Improving the Gaussianity of Radar Reflectivity Departures between Observations and Simulations Using the Symmetric Rain Rates

Yudong Gao[1], Lidou Huyan[1], Zheng Wu[1], Bojun Liu[2]

[1]Key Laboratory of Core Tech on Numerical Model-AI Integrated Forecast for Hazardous Precipitation, Chongqing Institute of Meteorological Sciences, Chongqing 401147, China.
[2]Chongqing Meteorological Observatory, Chongqing 401147, China.

Correspondence to: Yudong Gao (stephencool@163.com)

Abstract. Given that the Gaussianity of the observation error distribution is the fundamental principle of some data assimilation and machine learning algorithms, the error structure of radar reflectivity has become increasingly important with the development of high resolution forecasts and nowcasts of convective systems. This study examines the error distribution of radar reflectivity and discusses what causes the non-Gaussian error distribution using 6 month observations minus backgrounds (OmBs) of composites of vertical maximum reflectivity (CVMRs) in mountainous and hilly areas. By following the symmetric error model in all-sky satellite radiance assimilation, we reveal the error structure of CVMRs as a function of symmetric rain rates, which is the average of the observed and simulated rain rates. Unlike satellite radiance, the error structure of CVMRs shows a sharper slope for light precipitation than for moderate precipitation. Thus, a three-piecewise fitting function is more suitable for CVMRs. The probability density functions of OmBs normalized by symmetric rain rates become more Gaussian than the probability density functions normalized by all samples. Moreover, the possibility of using third-party predictor to construct the symmetric error model is also discussed in this study. The result shows the Gaussian distribution of OmBs can be further improved via more accurate precipitation observations. According to the Jensen-Shannon divergence, a more linear predictor, the logarithmic transformation of rain rate, can provide the most Gaussian error distribution in comparison with other predictors.

1 Introduction

The radar echo signal, called the equivalent reflectivity factor (unit: $mm^6\ m^{-3}$), is proportional to the sixth power of the hydrometeor diameter according to Rayleigh scattering. Thanks to its high accuracy and spatiotemporal resolution, the equivalent reflectivity factor can provide quantitative precipitation estimation (QPE) over a larger area in comparison with rain gauge data (Chang et al., 2021; Yo et al., 2021). the decibel relative to the equivalent reflectivity factor (hereafter shorted by reflectivity, unit: dB) has been commonly used in either data assimilation (DA) or machine learning (ML) algorithms. Applying DA or ML to reflectivity has enhanced forecasts and nowcasts of convective systems in the last ten years (Stensrud et al., 2013; Sun et al., 2014; Gustafsson et al., 2018; Ayzel et al., 2020; Cuomo and Chandrasekar, 2021; Baron et al., 2023). Most current DA algorithms assume a Gaussian error distribution of observations to guarantee statistically optimal estimations,

while some classical ML algorithms employ a Gaussian distribution to solve the convex optimization problem. However, few studies have investigated whether the error distribution of reflectivity is Gaussian.

To address non-Gaussian error distributions, several ensemble DA algorithms have been designed. For instance, the Gamma, Inverse-Gamma and Gaussian (GIGG) algorithm, proposed by Bishop (2016), can handle a highly skewed uncertainty distribution in an ideal model. The Quadratic Programming Ensemble Kalman Filter (QPEns), incorporating nonnegativity constraints such as mass, energy and enstrophy conservations into the classical Kalman Filter, has been recognized as another effective approach (Janjić et al., 2014; Gleiter et al., 2022). Because of the complex and expensive computations, the above DA algorithms toward non-Gaussian distributions are rarely employed by current operational systems. To further explore the potential of high resolution reflectivity data in current operational DA algorithms, the aim of this study is to improve the Gaussianity of the reflectivity error.

The error statistics associated with radar reflectivity, consisting of both instrument error and representation error (Janjić et al. 2018), have become increasingly important in DA. In earlier studies, defining super observation over a large area satisfied the assumption of uncorrelated errors (Sun and Crook, 1997; Snyder and Zhang, 2003; Tong and Xue, 2005). The error of these "superobbed" reflectivity data could approximate a Gaussian distribution with a constant value. Thousands of reflectivity data were discarded during the thinning process. Recently, with the popularity of the Desroziers method (Desroziers et al., 2005), the spatial error correlations of radar reflectivity were investigated by the Met Office (Waller et al., 2017) and the Deutscher Wetterdienst (Zeng et al., 2021), but the non-Gaussian error distribution is still a challenge in radar reflectivity assimilation. In this study, we critically examine the non-Gaussian error structure of the reflectivity and attempt to understand what causes the non-Gaussian error distribution.

Similar to the satellite radiance in all-sky reported by Geer and Bauer (2011), the radar reflectivity error also exhibits substantial non-Gaussian behaviour for several reasons:

1. Boundedness. There are two kinds of boundednesses for radar reflectivity. First, radar reflectivity itself is a bounded variable since the hydrometeors cannot be less than zero. A similar boundedness issue leads to a non-Gaussian error distribution in satellite radiance assimilation. The second boundedness indicates that the radar reflectivity could decrease rapidly to zero outside rainy areas because the distribution of hydrometeors is limited by geophysical boundaries, such as precipitation and non-precipitation areas. In contrast to satellite radiance assimilation, the discontinuity of hydrometeors in the background prevents non-precipitation area from assimilating reflectivity. It is called the "zero gradient" effect (Bannister et al., 2020).

2. Heteroscedasticity. The representation error of the reflectivity, defined by Observations minus Backgrounds (hereafter shorted by OmBs), can change with the convective strength. In reflectivity assimilation, the representation error, including mismatch between scales and observational operator error, increases with the intensification of convection. The mismatch between scales becomes worse when the convection intensifies rapidly, which often exhibit low predictability (Sun and Zhang, 2020), leading to large reflectivity OmBs. Moreover, the cold microphysics in a strong convection, including ice-phased and mix-phased hydrometeors, complicates the transformation from model variables to reflectivity (Jung et al., 2008) in comparison with the warm microphysics in a weak convection. Some assumptions about the shapes and sizes of ice-phased

65   hydrometeors could bring additional uncertainty to the observational operator of reflectivity. This also leads to large OmBs in the melting layer or upper levels of strong convection. Thus, the heteroscedasticity of reflectivity OmBs can be described by the convective strength.

In an idealized system, Bishop (2019) demonstrated that the state-dependent observation error variance should be anticipated and estimated whenever the observation is a bounded variable, whose error variance tends to zero as the observation approaches

70   the bound. Xue et al. (2007) also noted the importance of properly modelling reflectivity errors when the observation operator is nonlinear. The radar reflectivity is a distinct bounded measurement and has a complicated nonlinear observation operator. Inspired by these previous studies, the radar reflectivity error should be a state-dependent function instead of a constant value. In this study, we present the first in-depth study to unveil the error structure of reflectivity by following the successful construction of a symmetric error model in all-sky satellite radiance assimilation (Geer and Bauer, 2011; Migliorini and Candy,

75   2019; Zhu et al., 2019; Shahabadi and Buehner, 2021; Johnson et al., 2022).

To construct a symmetric error model, we need a symmetric predictor, which is the average of simulations and observations. For radar reflectivity, this predictor should be an estimation of convective strength and can be predicted by a numerical weather model. Similar to the liquid water path derived from satellite radiance observations, the rain rate can be estimated by the radar reflectivity in terms of the Z-I relationship and its variations. Meanwhile, the rain rate is also indicative of the convective

80   strength that correlates the reflectivity and rain rate in physics. Thus, this study uses the rain rate as a predictor of the symmetric error model of radar reflectivity to describe the heteroscedasticity of reflectivity OmBs.

It naturally steps forward to examine the effects of certain properties of the rain rate on the symmetric error model of radar reflectivity. The accuracy of rain rate data is the most uncertain property. It could vary from one dataset to another. In this study, we first focus on the effects of observation accuracy on the symmetric error model. As reported for reflectivity and

85   precipitation assimilation (Liu et al., 2020; Lopez, 2011), the logarithmic transform of hydrometeor control variables or observations can alleviate the nonlinear issue in reflectivity assimilation. Here the linearization, the logarithmic transform of rain rates, is the second property we attempt to investigate.

The rest of this study is organized as follows. In Section 2, observations, model equivalents and their OmBs are introduced. The properties of various predictors are discussed in Section 3. The error structure of radar reflectivity constructed by

90   symmetric rain rates is presented in Section 4. This section also shows the effects of the accuracy and linearization of the predictor on the symmetric error model of radar reflectivity. Finally, conclusions are given in Section 5.

2 Observations, model equivalents and their OmBs

2.1 Composite reflectivity observations

The weather radar network in Chongqing Municipality, denoted by red circles and dots in Fig. 1, consists of 5 radars and

95   covers the central and eastern Sichuan Basin. The two black rectangles, A and B, limit the research areas to exclude the model results outside of the radar coverage because the truth outside the radar network is unknown. While the Constant Altitude Plan

Position Indicators at 1 km altitude (hereafter shorted by 1 km CAPPIs) are more consistent with precipitation observations, the composites of vertical maximum reflectivity (hereafter shorted by CVMRs) can provide more samples in mountainous and hilly areas. Thus, the features of 1 km CAPPIs and CVMRs, from April to September 2021, are examined before matching

100   with the rain rate data. Both the 1 km CAPPIs and CVMRs have a 1 km resolution.

Figure 1: The inner domain and its topography (shaded; units: m) of the WRF model. The red dots and red dashed circles denote the radar stations and the coverage of the radar network, respectively. The research areas are limited by black rectangles A and B
105   **to exclude areas that are not covered by the radar network.**

The 1 km CAPPIs and CVMRs are interpolated linearly to a 5 km resolution in Fig. 2 to match the resolution of the rain rate data. Linear interpolation uses Euclidean distances as weights without the effects of terrain or the Earth's sphere. Figure 2a shows a southwest–northeast convective system was captured by the CVMRs at 1800 UTC on August 28th. Area A contains more convective cells than area B. In contrast, the 1 km CAPPIs, as shown in Fig. 2d, miss the convective cells in area A

110   owing to terrain blockage. Although both the 1 km CAPPIs and CVMRs indicate clear geophysical boundaries between precipitation and non-precipitation areas, the CVMRs could present better representations in mountainous areas. Notably, the zero gradient of hydrometeors caused by geophysical boundaries could create difficulties in the application of some DA and ML algorithms.

Figure 2: Distributions of CVMRs (first row, unit: dB) and 1 km CAPPIs (second row, unit: dB) observed by radars (a and d), simulated by model (b and e) and their OmBs (e and f) at 1800 UTC on August 28th, 2021. The black rectangles indicate the research areas, as in Fig. 1.

2.2 Model equivalents

The 6 month model equivalents of 1 km CAPPIs and CVMRs are simulated by the Weather Research and Forecasting (WRF; Skamarock et al., 2019) model Version 4.1. The Lambert projection, whose standard latitudes are 20° N and 30° N and standard longitude is 106.5° E, is used. Same physics packages, including the new Kain-Fritsch scheme (Kain, 2004), the Yonsei University planetary scheme (YSU, Hong et al., 2006), the Thompson scheme (Thompson et al., 2008) and Unified Noah Land Surface Model (Ek et al., 2003), are employed in the 6 month simulations. The WRF model has been nested in one-way with a coarse resolution of 9 km and a fine resolution of 3 km. Figure 1 gives the topography in the inner domain of the WRF model, whose central location is at (29.8° N, 106.58° E) and whose horizontal grids are 480×360. In the outer domain, the central location is at (30° N, 104.5° E) and the horizontal grids are 600×480. Both domains have 51 vertical layers.

The initial and lateral boundary conditions of the WRF model are 0.5°×0.5° Global Forecast System (GFS) data produced by the National Centers for Environmental Prediction. More information about GFS data sets is available at https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forecast-system-gfs. The GFS analyses at 0000 UTC and 1200 UTC from April to September are used to drive the WRF model. The model equivalents are computed using 6 hour simulations, because a shorter simulation time causes spin-up issues and a longer simulation time brings large model errors. The overall growth of model errors can be described by the 6 hour integration of the WRF model since various observations are assimilated by the GFS. No reflectivity assimilation has been performed here since we investigated the impacts of the symmetric error model on the climatology of representation error. The model equivalents have 12 hour time intervals (i.e., 0600 UTC and 1800 UTC) in this study.

The diagnostic algorithm of three-dimensional reflectivity, consisting of rain drops, snow particles and graupel particles, can be briefly described as follows:

$$Z = 10 \log_{10}(Z_{er} + Z_{es} + Z_{eg}) \qquad\qquad (1)$$

where Z_{er}, Z_{es} and Z_{eg} are the equivalent reflectivity factors for rain, snow and graupel droplets, respectively. This diagnostic algorithm (Stoelinga 2005) employs 8×10^6, 2×10^7 and 4×10^6 m^{-4} as intercept parameters for rain, snow and graupel droplets, respectively. The densities of rain, snow and graupel droplets are 1000, 100 and 400 kg m^{-3}, respectively. The Unified Post Processor (UPP) package (https://dtcenter.org/community-code/unified-post-processor-upp) interpolates diagnostic reflectivities from the coordinates of the WRF model to altitude levels and then generates the model equivalents of 1 km CAPPIs and CVMRs. Despite some empirical assumptions, this diagnostic algorithm can transform model variables, such as rain, snow and graupel mixing ratios, to reflectivity. Liu et al. (2022) used a similar diagnostic algorithm based on double-moment Thompson microphysics as the forward operator in reflectivity assimilation.

In Fig. 2b, the model equivalents of the CVMRs capture the southwest–northeast rain belt with strong convective cells in area A, illustrating that the WRF model is capable of simulating this convective system. The CVMRs and their model equivalents still present discrepancies in the comparison of Fig. 2a and Fig. 2b. As shown in Fig. 2c, the OmBs can vary widely from place to place, implying that a constant standard deviation may be insufficient to describe the error structure of CVMRs. For the 1 km CAPPIs, the model equivalents (Fig. 2e) and their OmBs (Fig. 2f) present similar features to those of the CVMRs. Thus, regardless of 1 km CAPPIs or CVMRs, the model equivalents are misplaced, ill-shaped, or have erroneous intensities compared to observations point by point. Following Geer and Bauer (2011), we refer to all these errors as 'mislocation' errors. The mislocation errors of 1 km CAPPIs and CVMRs can result in a non-Gaussian error distribution that violates the Gaussian assumptions underlying some DA and ML algorithms.

2.3 Observations minus Backgrounds

To represent the rainy echoes, the 1 km CAPPIs and CVMRs less than 5 dB are removed in this study. Thus, the samples in Fig. 3 do not contain false simulations (i.e., simulated, but not observed). Figure 3a shows a histogram of all CVMRs against their model equivalents based on 1165529 samples, including missed simulations (i.e., observed, but not simulated). The high numbers along the abscissa imply the large mislocation error of CVMRs resulting from considerable missed simulations. Compared with the satellite radiance departures (Fig. 5 in Migliorini and Candy, 2019), these considerable missed simulations are associated with the worse spatial discontinuity in the CVMR OmBs. For convenience, we refer to the discontinuous scenario as 'any-reflectivity'.

To examine the effects of the large mislocation error on the CVMR error structure, we removed all missed simulations and obtained 504123 samples (Fig. 3b). We refer to this scenario as 'both-reflectivity', whose histogram is similar to the nonprecipitating cloud affected satellite radiance observed by the AMSR-E channel 37v (Geer and Bauer, 2011). A comparison of Fig. 3a and Fig. 3b shows that 'any-reflectivity' scenario has a more complicated error structure than 'both-reflectivity' scenario, illustrating that the non-Gaussian error distribution in radar reflectivity assimilation is likely to be stronger than that in satellite radiance assimilation.
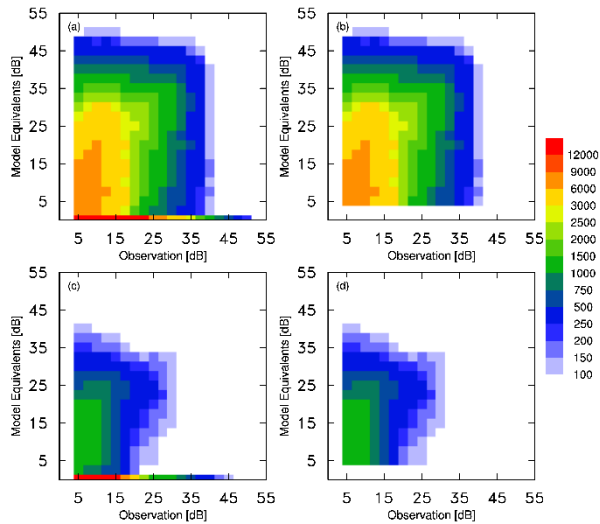
Figure 3: Histograms of the observed (a and b) CVMRs and (c and d) 1 km CAPPIs (abscissa, unit: dB) against their model equivalents (ordinate, unit: dB) in 'any-reflectivity' (the first column) and 'both-reflectivity' (the second column) scenarios.

The sample numbers of the 1 km CAPPIs decreased to 232681 and 71516 for 'any-reflectivity' and 'both-reflectivity', respectively. In the comparison of Fig. 3c and Fig. 3d, the 1 km CAPPIs also contain considerable missed simulations in terms of the high numbers along the abscissa. The error structure of the 1 km CAPPIs estimated by OmBs is similar to that of the CVMRs.

It is critical to understand the statistical features of several OmBs by examining their probability density functions (PDFs) before building a symmetric error model. Compared with the normal Gaussian distributions in Fig. 4, the PDF of CVMR OmBs (red solid line) in 'any-reflectivity' presents a positive skewness. Instead, the PDF for 'both-reflectivity' (blue solid line) is closer to a Gaussian distribution. The comparison illustrates that the numerous missed simulations along the abscissa in Fig. 3 have an undesirable effect on some DA and ML algorithms. In practice, the mismatches between observations and simulations provide valuable information related to convective systems. This non-Gaussian distribution cannot be ignored in radar reflectivity applications.

Similarly, the PDF of the 1 km CAPPI OmBs also approximates the Gaussian distribution after removing the missed simulations in Fig. 4. The means and standard deviations of the 1 km CAPPI and CVMR OmBs, denoted by μ and σ in Fig. 4, respectively, are similar as well. According to above comparisons, the statistical features of the 1 km CAPPI and CVMR OmBs are comparable in this study. Thus, the CVMR data in the 'any-reflectivity' scenario are used to match the rain rate data in the following sections.
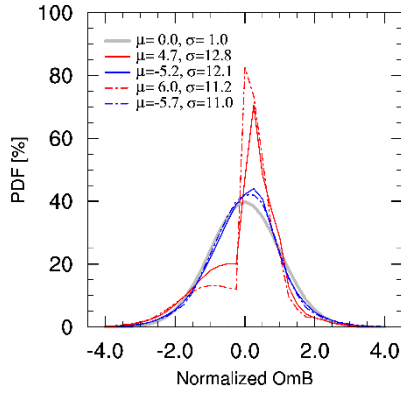
Figure 4: Probability density functions of CVMR (solid lines) and 1 km CAPPI (dashed lines) OmBs in the 'any-reflectivity' (red lines) and 'both-reflectivity' (blue lines) scenarios, normalized by the mean and standard deviation of all samples. The grey line represents a normal Gaussian distribution. The μ and σ denote the mean and standard deviation of OmBs, respectively.

3 Predictors of the symmetric error model

3.1 Predictor derived from reflectivity

The predictors of previous symmetric error models for satellite radiance assimilation were derived from the satellite radiance observations. Similarly, the rain rate can be derived from the echo signal in terms of the Z-I relationship, which is an empirical formula for estimating the rain rate I (unit: mm h^{-1}) from the equivalent reflectivity factor Z_e (unit: mm^6 m^{-3}):

$$Z_e = aI^b \qquad (2)$$

Here, the equivalent reflectivity factor at 3 km altitude and typical coefficients a=300 and b=1.4 are employed. Therefore, the 'symmetric' rain rate, rr_{sym}, which is used as the symmetric predictor in this study, is the average of the derived rain rate, rr_{obs}, and simulated rain rate, rr_{model}:

$$rr_{sym} = 0.5 \times (rr_{obs} + rr_{model}) \qquad (3)$$

In this study, the rr_{model} is the average of two consecutive hourly precipitations simulated by the WRF model, not derived by the reflectivity simulation.

Figure 5 shows the distributions of the rain rate data derived from the observations and simulated by the WRF model. Despite some disagreements for CVMRs below 15 dB in area A, the derived rain belt has a southwest–northeast distribution similar to that of the actual CVMRs. Moreover, the large rainy centers in Fig. 5a are associated with the strong convective cells in Fig. 2a. The simulated rain belt in Fig. 5b also presents similarities to the model equivalents of the CVMRs in Fig. 2b. Consequently, the rain rate OmBs in Fig. 5c agree with the CVMR OmBs in Fig. 2c, illustrating that the CVMR error structure can be described by the rain rates regardless of the discrepancy between the CVMRs and rain rates.
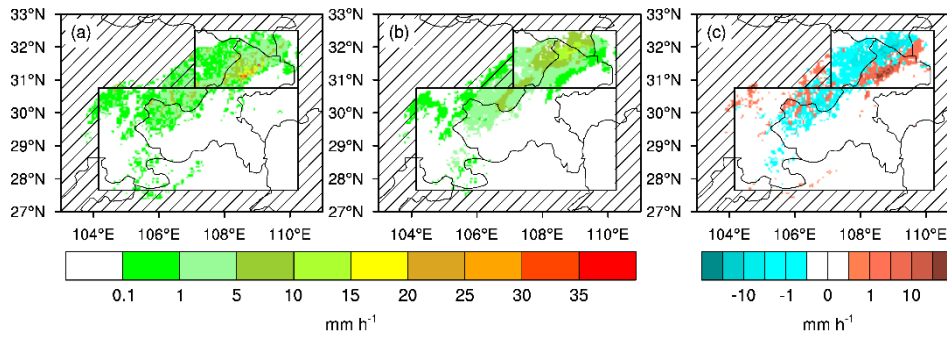
Figure 5: Distributions of rain rates (unit: mm h⁻¹) (a) derived from equivalent reflectivity factors at 3 km altitude, (b) simulated by the WRF model and (c) their OmBs at 1800 UTC on August 28th, 2021. The black rectangles indicate the research areas, as in Fig. 1.

3.2 Predictors from third-party observations

Derivation from the equivalent reflectivity factor is not the only way to obtain rain rate data. Other hourly precipitation observations can be used to produce rain rate data. Thus, it is of interest to discuss how the accuracy of rain rate affects the symmetric error model.

In this study, the derived rain rates are replaced by the CMA Multisource Precipitation Analysis System (CMPAS) data produced by the National Meteorological Information Center of the China Meteorological Administration (NMIC/CMA). Hourly CMPAS data with a 0.05° resolution, merging precipitation observations from rain gauges, radar QPEs and satellite QPEs, capture a number of hourly precipitation details and are more accurate than other single source precipitation observations (Pan et al., 2018; Li et al., 2022).

Comparing Fig. 5a and Fig. 6a, the CMPAS rain rates are comparable to the derived rain rates, especially for heavy precipitation in area A. Because radar observations have been used to generate the CMPAS data. The CMPAS rain rates present a smoother southwest–northeast rain belt and a more evident precipitation center in the mountainous area. A few small and moderate precipitations in area B are captured by the CMPAS rain rates, leading to a wider distribution of OmBs, as shown in Fig. 6b. Thus, a more accurate precipitation data can provide more reliable samples for the construction of symmetric error model.
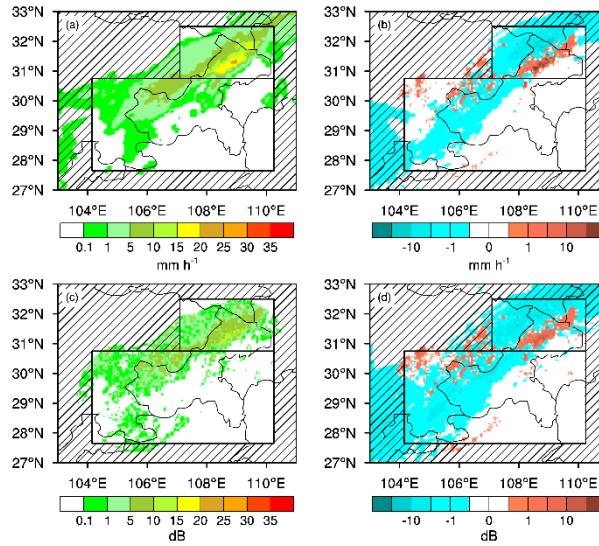
Figure 6: Distributions of (a) CMPAS rain rates (unit: mm h^{-1}) and (c) logarithmic rain rates (unit: dB) at 1800 UTC on August 28th, 2021. (b) and (d) are OmBs of the CMPAS rain rates (unit: mm h^{-1}) and logarithmic rain rates (unit: dB), respectively. The black rectangles indicate the research areas, as in Fig. 1

3.3 The linearization of predictor

The Z-I relationship exists between the rain rate I and the equivalent reflectivity factor Z_e (unit: mm^6 m^{-3}), not the reflectivity Z (unit: dB). A natural step forward is imposing a logarithmic transformation on Eq. 2 to obtain a more linear relationship between Z and I:

$$Z = 10 \log_{10} Z_e = 10 \log_{10} a + 10b \log_{10} I \qquad (4)$$

where a and b are the coefficients of the Z-I relationship. In this study, Eq. 4 is not a formula for accurately obtaining the quantitative reflectivity. It merely transforms the relationship between the CVMRs and symmetric rain rates to a more linear relationship, which allows us to discuss the effects of the linearization of predictor on the symmetric error model. Thus, this subsection uses $10 \log_{10}(I + 1.0)$, hereafter referred to as the logarithmic rain rate (unit: dB), as a linear predictor. Adding 1.0 to the rain rate ensures that the base of the logarithm is greater than zero, which is the same as for precipitation assimilation (Lopez, 2011).

The logarithmic rain rates also present the southwest–northeast rain belt in Fig. 6c. However, the precipitation center in area A is smoothed out by the logarithm. The OmBs of the logarithmic rain rates in Fig. 6d present similar negative and positive distributions in comparison with the derived rain rates in Fig. 5c. Notably, a number of precipitations below 0.1 mm h^{-1} are amplified by the above logarithmic transform, resulting in more OmBs of logarithmic rain rates. The logarithmic rain rates allow us to obtain more small precipitation samples.

To examine the relationship between CVMR OmBs and symmetric rain rates, it is advisable to count the number of CVMR OmBs over discrete intervals of symmetric rain rates, chosen here to be 0.5 mm h^{-1}. Owing to the numerous missed simulations

in Fig. 3a, most OmBs of the derived rain rates (Fig. 7a) and CMPAS rain rates (Fig. 7b) range from -20 to 30 dB when the symmetric rain rates are less than 0.5 mm h⁻¹. As shown in Fig. 7a, the major OmBs against derived rain rates, chosen to be larger than 500 samples, become bimodal as the symmetric rain rates increase roughly from 0.5 to 2 mm h⁻¹. The two peaks are at about 30 dB and -10 dB.
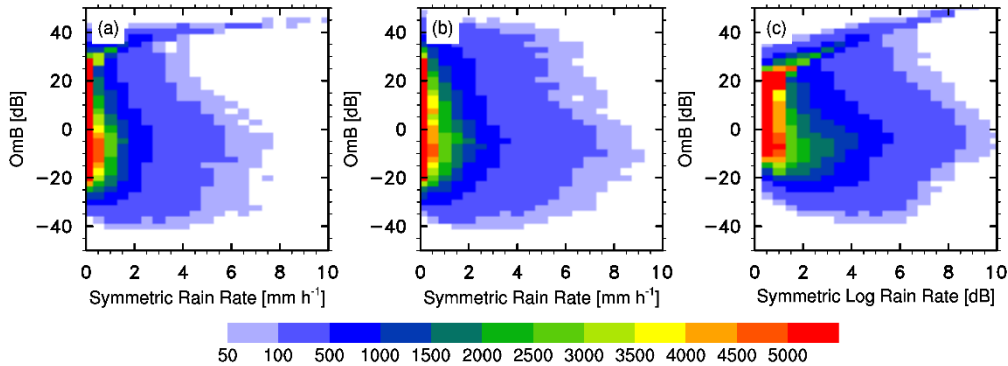
Figure 7: Histograms of CVMR OmBs (ordinate, unit: dB) against different symmetric predictors (abscissa), which are rain rates (unit: mm h⁻¹) (a) derived by equivalent reflectivity factors, (b) computed by CMPAS data, and (c) the symmetric logarithmic rain rate (unit: dB).

In contrast, the major OmBs against CMPAS rain rates in Fig. 7b exhibit a unimodal distribution peaking at about -10 dB. Although this unimodal distribution is not symmetric when the OmB equals zero, it is closer to a Gaussian distribution, confirming that more accurate CMPAS data can offer superior representation. When comparing the derived rain rates (Fig. 7a) with the logarithmic rain rates (Fig. 7c), the major OmBs exhibit a bimodal distribution, but become gentle along the abscissa. As a result, the logarithmic transformation reduces the rain rate gradient without altering the structure of the CVMR OmBs.

4 Errors as a function of symmetric rain rates

4.1 The CVMR symmetric error model

Similar to satellite radiances, it is possible to investigate the CVMR error structure over discrete rain rate bins, chosen to be 0.5 mm h⁻¹ in this study. As shown in Fig. 8a, the standard deviations of the CVMR OmBs vary from about 10 to 33 dB. A constant value is insufficient to describe the CVMR error structure. The difference between the first two bins is much greater than that between the other bins. To illustrate this, we may argue that light precipitation is closer to the geophysical boundary than moderate precipitation, resulting in a greater difference between the first two bins. From the second bin, the standard deviations of the CVMR OmBs increase with symmetric derived rain rates before peaking at 8.0 mm h⁻¹. The standard deviations that alternately increase and decrease after 8.0 mm h⁻¹ could be caused by poor initial conditions of the WRF model, small sample numbers or inaccurate diagnostic reflectivity.
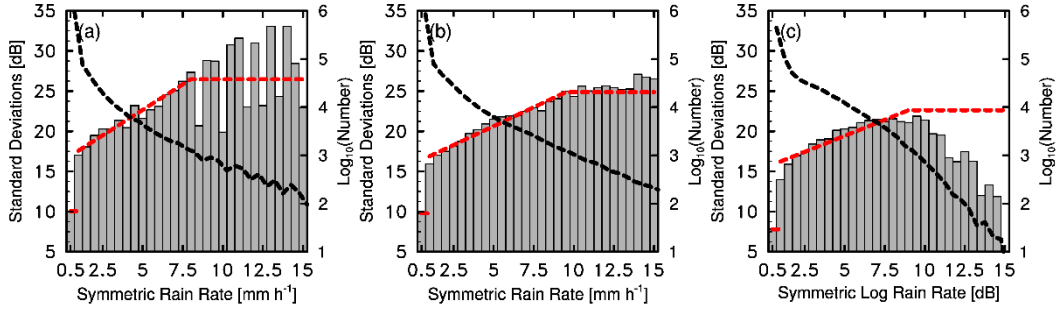
Figure 8: Standard deviations of CVMR OmBs for symmetric (a) derived rain rates (unit: mm h⁻¹), (b) CMPAS rain rates (unit: mm h⁻¹) and (c) logarithmic rain rates (unit: dB). The red dashed lines show the three-piecewise fitting functions (listed in Table 1). The black dashed lines show the logarithm of sample numbers over symmetric rain rate bins.

To simplify the complex CVMR error structure, a three-piecewise function (red dashed line) is fitted by using linear regression. The first bin must be isolated from the linear regression to pass the 95% confidence level for the F-test. A straight line rather than a linear regression is used to describe the reflectivity error for large symmetric derived rain rates. This is a cautious approach to fit a rational linear regression based on a large sample size (black dashed line), chosen to be larger than 10^3 samples. Table 1 lists the key parameters of the piecewise functions.

Table 1: the key parameters of the three-piecewise fitting functions.

Predictor	Function	Rain rate range	R^2
Derived rain rates	y=10.04	0.0<x≤0.5	0.94
	y=16.31+1.27x	0.5<x≤8.0	
	y=26.47	8.0<x	
CMPAS rain rates	y=9.78	0.0<x≤0.5	0.96
	y=15.94+0.94x	0.5<x≤9.5	
	y=24.87	9.5<x	
Logarithmic rain rates	y=7.8	0.0<x≤0.5	0.83
	y=15.43+0.80x	0.5<x≤9.0	
	y=21.64	9.0<x	

As shown in Fig. 8b, similar characteristics, such as the distinct difference between the first two bins and the increase in symmetric derived rain rates, are captured by the symmetric CMPAS rain rates as well. The standard deviations vary from about 10 to 25 dB when the symmetric CMPAS rain rate increases from 1 to 9.5 mm h⁻¹. The small variation in the standard deviations after 10 mm h⁻¹ results from the superior representation of the CMPAS data. For the symmetric logarithmic rain rates (Fig. 8c), the standard deviations of the CVMR OmBs grow gradually from roughly 14 to 21 dB as the symmetric logarithmic rain rates increase from 1 to 10, even if they still increase quickly from about 8 to 14 in the first two bins. The

decreasing trend at the tail of the logarithmic rain rates (lager than 9.0) results from the rapid decrease in sample size. The straight line prevents the three-piecewise fitting function from being obtained from an irrational linear regression. According to Table 1, the logarithmic rain rates obtain the smallest slope of the fitting function among the three symmetric predictors, despite having the smallest R^2.

4.2 Improvements in Gaussianity

To illustrate the potential benefits of symmetric error models to some DA and ML algorithms, the Gaussianity of PDF is examined in this subsection. Although the PDF of CVMR OmBs is not Gaussian, the CVMR OmBs can be divided into a number of subgroups with Gaussian PDFs according to the binned standard deviations or piecewise functions from the above subsection. Figure 9 shows the PDFs of the CVMR OmBs normalized by various symmetric rain rates, with the raw and normal Gaussian PDFs displayed for comparison. Compared with the raw PDF (green line), the PDFs normalized by the binned standard deviations (red line) become more Gaussian. The three-piecewise function, which simplifies the CVMR error structure, also corrects the positive skewness of the raw PDF. We argue that the three-piecewise function is sufficient in this study because it shows an identical PDF to the binned standard deviations.
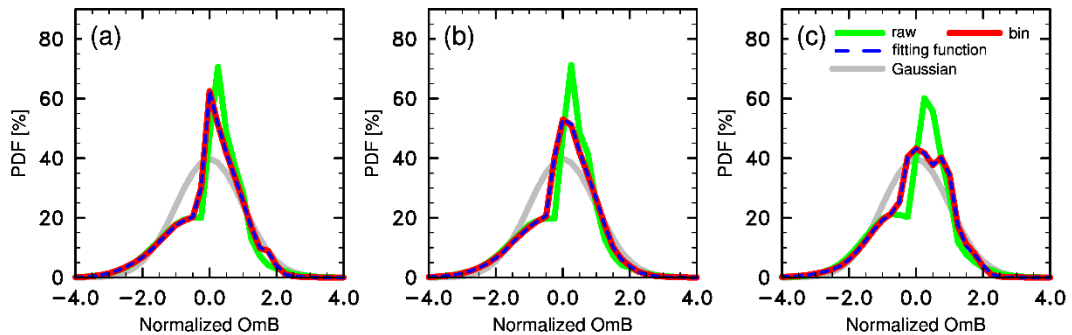
Figure 9: PDFs of CVMR OmBs normalized by symmetric (a) derived rain rates, (b) CMPAS rain rates and (c) logarithmic rain rates. The green, red, blue and grey lines represent the raw, binned, three-piecewise and normal Gaussian PDFs, respectively.

To quantify the similarity between the PDF normalized by the symmetric rain rates and a normal Gaussian PDF, Table 2 lists the Jensen-Shannon divergence (JSD):

$$JSD(P \parallel Q) = \frac{1}{2}\sum P(x) \log(\frac{2P(x)}{P(x)+Q(x)}) + \frac{1}{2}\sum Q(x) \log(\frac{2Q(x)}{P(x)+Q(x)}) \qquad (5)$$

where P is the PDF normalized by symmetric rain rates or raw standard deviations and Q represents a normal Gaussian PDF. When JSD is zero, the distributions P and Q are the same. For the derived rain rates, the JSDs of the PDFs normalized by the binned standard deviations and the three-piecewise function decrease from 0.010 to 0.006.

Table 2: Jensen-Shannon divergences of probability density functions normalized by various symmetric rain rates.

Predictor	Raw	Three-piecewise	Binned

Derived rain rates	0.010	0.006	0.006
CMPAS rain rates	0.010	0.005	0.005
Logarithmic rain rates	0.008	0.004	0.004

For the CMPAS rain rates in Fig. 9b, the PDFs normalized by the binned standard deviations and the three-piecewise function not only correct the positive skewness, but also reduce the overestimation in the central area. The CMPAS rain rates also have smaller JSDs than the derived rain rates, as listed in Table 2. This demonstrates that the accuracy of CMPAS rain rates can further improve the Gaussianity of the PDFs. For the logarithmic rain rates (Fig. 9c), the PDFs normalized by the binned standard deviations and three-piecewise function also approximate a normal Gaussian distribution according to comparison with the raw PDF. It is worth noting that the logarithmic rain rates obtain the smallest JSDs despite a few fluctuations in the PDFs normalized by the binned standard deviations and three-piecewise function.

5 Conclusions

In this study, the Gaussianity of two types of OmB data, i.e., the CVMRs and 1 km CAPPIs, is examined in Southwest China. Their features, such as horizontal distributions and PDFs, are similar regardless of the different definitions between the CVMRs and 1 km CAPPIs. Consequently, the 6 month CVMR OmBs, which exhibit superior representation to 1 km CAPPI OmBs in mountainous and hilly areas, are employed to discuss the handling of non-Gaussian PDFs.

In the comparison of the 'any-reflectivity' and 'both-reflectivity' scenarios, the Gaussianity of OmBs can be improved by removing the numerous mismatches between observations and simulations. These mismatches cannot be ignored in some DA or ML algorithms. They provide essential information related to convective systems. Moreover, the reflectivity OmBs often vary widely from place to place, demonstrating that a constant standard deviation is insufficient to describe the error structure of radar reflectivity in most studies and operations.

The symmetric error model, which has been broadly used in all-sky satellite radiance assimilation (Migliorini and Candy, 2019; Zhu et al., 2019; Shahabadi and Buehner, 2021), is built to improve the Gaussianity of CVMR OmBs. According to the symmetric derived rain rates, the standard deviations of CVMR OmBs can vary from about 10 to 33 dB. However, the instrument noise of radar is on the order 1 dB.

Similar to satellite radiance, the standard deviations of CVMR OmBs increase with the symmetric derived rain rates, illustrating that the largest component of the CVMR OmBs comes from the poor prediction associated with clouds and rain and the inaccurate diagnostic algorithm of radar reflectivity in some DA and ML applications. As discussed in Geer and Bauer (2011), using the symmetric error model in reflectivity assimilation may also compensate for the inadequate background error specification of hydrometeors, which will be investigated by DA experiments in our ongoing study. In contrast to satellite radiance, the symmetric error model of CVMR data shows that the difference between the first two bins is much greater than

that between the other bins, illustrating that a more complex structure, a three-piecewise function, should be formulated at the convective-allowing scale.

Compared with the raw PDF, the PDFs normalized by the binned standard deviations and the three-piecewise function become more Gaussian by reducing the positive skewness. Each subgroup of CVMR OmBs, separated by symmetric derived rain rates, approximates a Gaussian PDF despite the non-Gaussian PDF of all samples. Thus, this study demonstrates that the Gaussianity of CVMR OmBs can be improved by the symmetric error model based on the derived rain rates.

The effects of more accurate rain rate data on the symmetric error model of CVMRs are also examined in this study. Although the CMPAS rain rates build a three-piecewise function similar to that of the derived rain rates, the superior representation can further improve the Gaussianity of CVMR OmBs in terms of the JSDs in Table 2.

The logarithmic rain rates have profound effects on the symmetric error model of CVMR OmBs. Not only do the gradients of the standard deviations of CVMR OmBs decrease from the second bin, but also the PDFs normalized by the binned standard deviations and the three-piecewise function obtain the smallest JSDs compared with those of the other rain rates. It is convenient to create configuration files for the logarithmic rain rates in the operational system. Moreover, the logarithmic transform has been used to assimilate precipitation observations directly in operational four-dimensional variation system at the European Centre for Medium-Range Weather Forecasts (Lopez, 2011). Thus, using a more linear predictor is recommended for building a symmetric error model of CVMRs.

In theory, the symmetric error models of CVMRs built in this study are more consistent with the fundamental principle of some DA and ML algorithms than a constant value. However, the symmetric error model, estimated by OmB data, highly relies on the numerical weather model, DA or ML strategy and forward observation operator. Consequently, this study encourages readers to build an effective symmetric error model based on their own assimilation and prediction systems.

Performing several experiments to discuss the effects of symmetric error models on several DA and ML algorithms is also encouraged. The immature use of the symmetric error model is briefly described here:

$$\sigma = \begin{cases} \sigma_l & RR_{avg} < RR_{avg1} \\ \sigma_l + \alpha\beta(RR_{avg} - RR_{avg1}) & RR_{avg1} \le RR_{avg} < RR_{avg1} \\ \sigma_u & RR_{avg} \le RR_{avg2} \end{cases} \tag{6}$$

where RR_{avg} is the symmetric rain rate, and σ_l and σ_u are the lower and upper boundaries of the reflectivity error, respectively. β is the slope of the three-piecewise function and α is a tuning parameter, as designed by Geer and Bauer (2011). By tuning the parameter α, the representative error can either be assigned completely by the symmetric error model (α=1) or ignored (α=0). In the future, the effects of ice-phased hydrometeors on the symmetric error model of CVMRs should be considered. Polarization measurements and their combinations may provide additional information about hydrometeors.

Code and data availability

The observations, simulations and derived rain rates are available at https://doi.org/10.6084/m9.figshare.25093508.v2. The graphics were generated using the NCAR Commend Language (https://www.ncl.ucar.edu/Download/). The Weather Research and Forecasting (WRF) model (V4.1) used in this study is available from the public WRF-Model Release page on GitHub (https://github.com/wrf-model). The Unified Post Processing (UPP) System for the WRF model is also available at GitHub (https://github.com/NOAA-EMC/UPP).

Author contribution

YG conceptualized this study and generated all the figures. LH and YG computed the Observation minus Background datasets and built the symmetric error models of radar observations. WZ performed the WRF simulations, and BL implemented quality control for radar reflectivity. YG prepared the paper and its revised versions with contributions from all the authors.

Competing interests

The authors declare that they have no conflicts of interest.

Acknowledgements

This study was jointly funded by the National Natural Science Foundation of China (42375161 and U2342220) and the Natural Science Foundation of Chongqing Municipality (cstc2021jcyj-msxmX0698).

References

Ayzel, G., T. Scheffer, and M. Heistermann: Rainnet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.*, 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, 2020.

Bannister, R. N., Chipilski, H. G., and Martinez-Alvarado, O.: Techniques and challenges in the assimilation of atmospheric water observations for numerical weather prediction towards convective scales. *Q J R Meteorol Soc*, 146, 1–48. https://doi.org/10.1002/qj.3652, 2020.

Baron, P., and Coauthors: Nowcasting Multiparameter Phased-Array Weather Radar (MP-PAWR) Echoes of Localized Heavy Precipitation Using a 3D Recurrent Neural Network Trained with an Adversarial Technique. *J. Atmos. Oceanic Technol.*, 40, 803–821, https://doi.org/10.1175/JTECH-D-22-0109.1, 2023.

Bishop, C. H.: The GIGG-EnKF: ensemble Kalman filtering for highly skewed non-negative uncertainty distributions. *Q J R Meteorol Soc*, 142: 1395–1412. https://doi.org/10.1002/qj.2742, 2016.

Bishop, C. H.: Data assimilation strategies for state-dependent observation error variances. *Q J R Meteorol Soc*, 145, 217–227, https://doi.org/10.1002/qj.3424, 2019.

410 Chang, P., and Coauthors: An Operational Multi-Radar Multi-Sensor QPE System in Taiwan. *Bull. Amer. Meteor. Soc.*, 102, E555–E577, https://doi.org/10.1175/BAMS-D-20-0043.1, 2021.

Cuomo, J., and V. Chandrasekar: Use of Deep Learning for Weather Radar Nowcasting. *J. Atmos. Oceanic Technol.*, 38, 1641–1656, https://doi.org/10.1175/JTECH-D-21-0012.1, 2021.

Desroziers, G., Berre, L., Chapnik, B. and Poli, P.: Diagnosis of observation, background and analysis-error statistics in
415 observation space. *Q J R Meteorol Soc*, 131, 3385–3396, https://doi.org/10.1256/qj.05.108, 2005.

Ek, M. B., Mitchell, K. E., Rogers, E., Lin, Y., Grunmann, P., Koren, V., et al.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational Mesoscale Eta Model. *Journal of Geophysical Research*, 108, 8851. doi:10.1029/2002JD003296, 2003.

Gao, Y.: Data used in the publication: Improving the Gaussianity of Radar Reflectivity Departures between Observations and
420 Simulations by Using the Symmetric Rain Rate [Dataset]. figshare, https://doi.org/10.6084/m9.figshare.25093508.v2, 2024.

Geer, A. J., and Bauer, P.: Observation errors in all-sky data assimilation. *Q J R Meteorol Soc*, 137, 2024–2037, https://doi.org/10.1002/qj.830, 2011.

Gleiter, T., Janjić, T. & Chen, N.: Ensemble Kalman filter based data assimilation for tropical waves in the MJO skeleton model. *Q J R Meteorol Soc*, 148, 1035–1056, https://doi.org/10.1002/qj.4245, 2022.

425 Gustafsson, N., Janjić, T., Schraff, C., Leuenberger, D., Weissmann, M., Reich, H., Brousseau, P., Montmerle, T., Wattrelot, E., Bučánek, A. and Mile, M.: Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Q J R Meteorol Soc*, 144, 1218–1256, https://doi.org/10.1002/qj.3179, 2018.

Hong, S.-Y., Noh, Y. and Dudhia, J.: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, 134, 2318–2341, http://doi.org/10.1175/MWR3199.1, 2006.

430 Janjić, T., McLaughlin, D., Cohn, S.E. and Verlaan, M.: Conservation of mass and preservation of positivity with ensemble-type Kalman filter algorithms. *Mon. Wea. Rev.*, 142, 755–773, https://doi.org/10.1175/MWR-D-13-00056.1, 2014.

Janjić, T., and Coauthors: On the representation error in data assimilation. *Quart. J. Roy. Meteor. Soc.*, 144, 1257–1278, https://doi.org/10.1002/qj.3130, 2018.

Johnson, A., Wang, X., and Jones, T.: Impacts of assimilating GOES-16 ABI channels 9 and 10 clear air and cloudy radiance
435 observations with additive inflation and adaptive observation error in GSI-EnKF for a case of rapidly evolving severe supercells. *Journal of Geophysical Research: Atmospheres*, 127, e2021JD036157. https://doi.org/10.1029/2021JD036157, 2022.

Jung, Y., Xue, M., Zhang, G. F., and Straka, J. M.: Assimilation of simulated polarimetric radar data for a convective storm using the ensemble Kalman filter. Part II: Impact of polarimetric data on storm analysis. *Mon. Wea. Rev.*, 136, 2246–2260,
440 https://doi.org/10.1175/2007MWR2288.1, 2008.

Kain, J. S.: The Kain-Fritsch convective parameterization: An update. *Journal of Applied Meteorology*, 43, 170–181, https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2, 2004.

Li, S. Y., Huang, X. L., Wu, W., Du, B., and Jiang, Y. H.: Evaluation of CMPAS precipitation products over Sichuan, China. *Atmospheric and Oceanic Science Letters*, 15, https://doi.org/10.1016/j.aosl.2021.100129, 2022,

445   Liu, C. S., Xue, M., and Kong, R.: Direct Variational Assimilation of Radar Reflectivity and Radial Velocity Data: Issues with Nonlinear Reflectivity Operator and Solutions. *Mon. Wea. Rev.*, 148, 1483–1502, https://doi.org/10.1175/MWR-D-19-0149.1, 2020.

Liu, C., H. Li, M. Xue, Y. Jung, J. Park, L. Chen, R. Kong, and C. Tong: Use of a Reflectivity Operator Based on Double-Moment Thompson Microphysics for Direct Assimilation of Radar Reflectivity in GSI-Based Hybrid En3DVar. Mon. Wea.

450   Rev., 150, 907–926, https://doi.org/10.1175/MWR-D-21-0040.1, 2022.

Lopez, P.: Direct 4D-Var assimilation of NCEP stage IV radar and gauge precipitation data at ECMWF. *Mon. Wea. Rev.*, 139, 2098–2116, https://doi.org/10.1175/2010MWR3565.1, 2011.

Migliorini, S., and Candy, B.: All-sky satellite data assimilation of microwave temperature sounding channels at the Met Office. *Q J R Meteorol Soc*, 145, 867–883, https://doi.org/10.1002/qj.3470, 2019.

455   Pan, Y., Gu, J. X., Xu, B., Shen, Y., Han, S., and Shi, C. X.: Advances in multi-source precipitation merging research. *Advances in Meteorological Science and Technology*, 8, 143–152, doi: 10.3969/j.issn.2095-1973.2018.01.019, 2018. (in Chinese)

Shahabadi, M. B., and Buehner, M.: Toward All-Sky Assimilation of Microwave Temperature Sounding Channels in Environment Canada's Global Deterministic Weather Prediction System. *Mon. Wea. Rev.*, 149, 3725–3738, https://doi.org/10.1175/MWR-D-21-0044.1, 2021.

460   Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., et al.: A Description of the Advanced Research WRF Model Version 4. https://doi.org/10.6084/m9.figshare.7369994.v4, 2019.

Snyder, C., and Zhang, F. Q.: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, 131, 1663–1677, https://doi.org/10.1175//2555.1, 2003.

Stensrud, D. J., Wicker, L. J., Xue, M., Dawson, D. T., Yussouf, N., Wheatley, D. M., et al.: Progress and challenges with

465   warn-on-forecast. *Atmospheric Research*, 123, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004, 2013.

Stoelinga, M. T.: Simulated equivalent reflectivity factor as currently formulated in RIP: Description and possible improvements. University of Washington Tech. Rep., 5 pp. [Available online at: http://www.atmos.washington.edu/~stoeling/RIP_sim_ref.pdf], 2005.

Sun, J. Z., and Crook, N. A.: Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and

470   its adjoint. Part I: Model development and simulated data experiments. *Journal of the Atmospheric Sciences*, 54, 1642–1661, https://doi.org/10.1175/1520-0469(1997)054,1642:DAMRFD.2.0.CO;2, 1997.

Sun, J. Z., Xue, M., Wilson, J. M., Zawadzki, I., Ballard, S. P., Onvlee-Hooimeyer, J., et al.: Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95, 409–426, https://doi.org/10.1175/BAMS-D-11-00263.1, 2014.

Sun, Y. Q. and Zhang, F. Q.: A New Theoretical Framework for Understanding Multiscale Atmospheric Predictability. *Journal of the Atmospheric Sciences*, 77, 2297–2309, https://doi.org/10.1175/JAS-D-19-0271.1, 2020.

Thompson, G., Field, P. R., Rasmussen, M., and Hall, W. D.: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme, Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, 136, 5095–5115, https://doi.org/10.1175/2008MWR2387.1, 2008.

Tong, M. J., and Xue, M.: Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments. *Mon. Wea. Rev.*, 133, 1789–1807, https://doi.org/10.1175/MWR2898.1, 2005.

Waller, J. A., Dance, S. L. and Nichols, N. K.: On diagnosing observation-error statistics with local ensemble data assimilation. *Q J R Meteorol Soc*, 143, 2677–2686, https://doi.org/10.1002/qj.3117, 2017.

Xue, M., Jung, Y., and Zhang, G. F.: Error modeling of simulated reflectivity observations for ensemble Kalman filter assimilation of convective storms. *Geophyical Research Letters*, 34, L10802, doi: 10.1029/2007GL029945, 2007.

Yo, T.-S., Su, S.-H., Chu, J.-L., Chang, C.-W., and Kuo, H.-C.: A deep learning approach to radar-based QPE. *Earth and Space Science*, 8, e2020EA001340. https://doi.org/10.1029/2020EA001340, 2021.

Zeng, Y., Janjic, T., Feng, Y., Blahak, U., de Lozar, A., Bauernschubert, E., Stephan, K., and Min, J.: Interpreting estimated observation error statistics of weather radar measurements using the ICON-LAM-KENDA system, *Atmos. Meas. Tech.*, 14, 5735–5756, https://doi.org/10.5194/amt-14-5735-2021, 2021.

Zhu, Y. Q., Gayno, G., Purser, R. J., Su, X. J., and Yang, R. H.: Expansion of the All-Sky Radiance Assimilation to ATMS at NCEP. *Mon. Wea. Rev.*, 147, 2603–2620, https://doi.org/10.1175/MWR-D-18-0228.1, 2019.