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

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

[1]Key Laboratory of Core Tech on NWP-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 observation error distribution is the fundamental principle of some data assimilation and machine learning algorithms, the error structure of radar reflectivity becomes 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 give rise to the non-Gaussian error distribution by 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 unveil the error structure of CVMRs as a function of symmetric rain rates, which is the average of observed and simulated rain rates. Unlike satellite radiance, the error structure of CVMRs shows a sharper slope in light precipitations than moderate precipitations. 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 in comparison with the probability density function normalized by the whole samples. Moreover, the possibility of using third-party predictor to construct the symmetric error model are also discussed in this study. The Gaussianity of OmBs can be further improved by using a 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 reflectivity factor (unit: $mm^6\ m^{-3}$), is proportional to the sixth power of the hydrometeor diameter according to the Rayleigh scattering. Thanks to the high accuracy and spatiotemporal resolution, the reflectivity factor can provide quantitative precipitation estimation (QPE) over a larger area in comparison with rain gauges (Chang et al., 2021; Yo et al., 2021). On the other hand, the decibels, called equivalent reflectivity (unit: dBZ) which is a logarithmic transformation of reflectivity factor, have been used in either data assimilation (DA) or machine learning (ML) algorithms to improve the forecast and nowcast of convective systems in 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 the Gaussian error distribution of observations in order to guarantee statistically optimal estimations, meanwhile some classical ML

30   algorithms employ Gaussian distribution to solve a convex optimization problem. However, few studies have investigated whether the error distribution of radar reflectivity is Gaussian.

    To attack the non-Gaussian error distribution, some 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 ~~idea~~ ideal model. The Quadratic Programming Ensemble Kalman Filter (QPEns), incorporating nonnegativity

35 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 computation, above DA algorithms toward non-Gaussian distribution are hardly employed by current operational systems. To further explore potentials of high resolution reflectivity data in currently operational DA algorithms, the aim of this study is to improve the Gaussianity of reflectivity error.

40   The error statistics associated with radar reflectivity, consisting of both the instrument error and representation error (Janjić et al. 2018), 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 to a Gaussian distribution with a constant value. Thousands of reflectivity data were discarded in the thinning process. Recently, with the popularity of the Desroziers method (Desroziers et al., 2005),

45 the spatial error correlations of radar reflectivity were investigated in 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 equivalent reflectivity and attempt to understand what give rise to the non-Gaussian error distribution.

    Similar to the satellite radiance in all-sky reported by Geer and Bauer (2011), we can summarize that the radar reflectivity

50 error also exhibits substantial non-Gaussian behaviour because:

    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. The similar boundedness issue leads to the non-Gaussian error distribution in satellite radiance assimilation. The second boundedness indicates that the radar reflectivity could decrease fast to zero outside the rainy areas, because the distribution of hydrometeors is limited by geophysical boundaries, such as precipitation and non-

55 precipitation areas. Different 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 ~~equivalent~~ reflectivity, defined by Observations minus Backgrounds (hereafter shorted by OmBs), can change with the convective strength. ~~as a function of precipitation.~~ It is clear in reflectivity assimilation, where the representation error including mismatch between scales and observational operator error increases with

60 the intensification of convection. ~~where errors including representation errors and operator errors increase with the precipitation amount.~~ The mismatch between scales becomes worse when the convection intensifies rapidly, ~~representation errors, described by the departures between observations and simulations, usually called Observations minus Backgrounds (hereafter shorted by OmBs), increase rapidly with model errors for intense conventions,~~ 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 of the warm microphysics in a weak convection. Some assumptions about the shapes and sizes of ice-phased hydrometeors could bring additional uncertainty in the observational operator of reflectivity. It is also leads to large OmBs in the melting layer or upper level of a strong convection. Thus, the heteroscedasticity of reflectivity OmBs can be described by the convective strength. ~~the errors of observation operator in reflectivity assimilation also become large when the convective systems intensify. For instance, the simplified reflectivity operator is insufficient to describe the shapes and sizes of ice-phased hydrometeors in strong conventions (Jung et al., 2008).~~

In an idealized system, Bishop (2019) demonstrated that the state-dependent observation error variance should be anticipated and estimated whenever the observation is of a bounded variable, whose error variance tends to zero as the observation approaches the bound. Xue et al. (2007) also pointed to the importance of properly modelling reflectivity errors when the observation operator is nonlinear. The radar reflectivity is distinctly a bounded measurement and has complicated nonlinear observation operator. As inspired by these previous studies, the error of radar reflectivity 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 equivalent reflectivity by following the successful construction of symmetric error model in all-sky satellite radiance assimilation (Geer and Bauer, 2011; Migliorini and Candy, 2019; Zhu et al., 2019; Shahabadi and Buehner, 2021; Johnson et al., 2022).

To construct 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 ~~hydrometeors~~ convective strength and can be predicted by 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~~, demonstrating that the radar reflectivity is highly related to precipitation in convective systems~~. Meanwhile, the rain rate is also indicative of the convective strength that correlates the reflectivity and rain rate in physics. Thus, this study uses the rain rate as the predictor of the symmetric error model of radar reflectivity to describe the heteroscedasticity of reflectivity ~~error~~OmBs.

It naturally steps forward to examine the effects of some properties of rain rate on the symmetric error model of radar reflectivity. The accuracy of rain rates is the most uncertain property. It could vary from one data set to another. In this study, we first focus on the effects of observation accuracy on the symmetric error model. As reported by reflectivity and precipitation assimilation (Liu et al., 2020; Lopez, 2011), the logarithmic transform on 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. Properties of various predictors are discussed in section 3. The error structure of radar reflectivity constructed by symmetric rain rates is presented in section 4. This section also shows the effects of the accuracy and linearization of 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 covers the center and east of the Sichuan Basin. The two black rectangles A and B limit the research areas in order to exclude those model results out 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) is 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 in 2021, are examined before matching with the rain rate data. Both of 1 km CAPPIs and CVMRs have 1 km resolution.
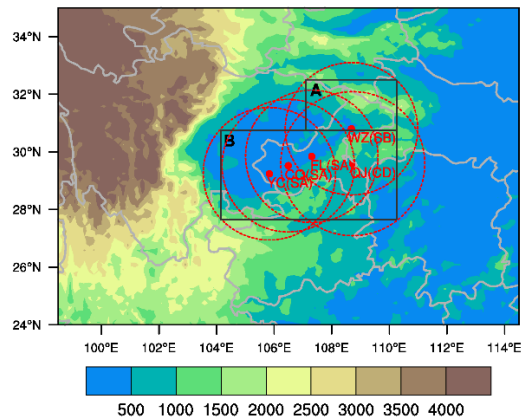
Figure 1: The inner domain and its topography (shaded; units: m) of WRF model. The red dots and red dash circles denote radar stations and the coverage of radar network respectively. The research areas are limited by the black rectangles A and B to exclude areas that are not covered by radar network.

The 1 km CAPPIs and CVMRs are interpolated linearly to 5 km resolution in Fig. 2 in order to match with the resolution of rain rate data. The linear interpolation uses Euclidean distances as weights, without the effects of terrain or earth sphere. Figure 2a shows a southwest–northeast convective system was captured by 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 owing to the terrain blockage. Although both 1 km CAPPIs and CVMRs indicate clear geophysical boundaries between precipitation and non-precipitation areas, the CVMRs could present better representations in mountainous areas. It is worth noting that the zero gradient of hydrometeors caused by geophysical boundaries could create difficulties in applications of some DA and ML algorithms.
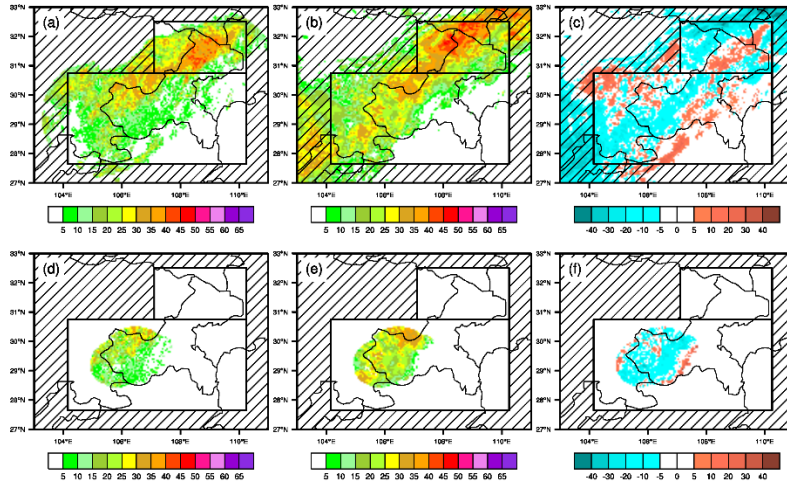
Figure 2: Distributions of CVMRs (the first row, unit: dBZ) and 1 km CAPPIs (the second row, unit: dBZ) observed by radars (a and d), simulated by model (b and e) and their OmBs (e and f) at 1800 UTC on August 28th, 2021. The black rectangles indicate the research areas, same as Fig. 1.

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 WRF model, whose central location is at (29.8° N, 106.58° E) and horizontal grids are 480×360. In the ~~outmost~~ outer domain, the central location is at (30° N, 104.5° E) and the horizontal grids are 600×480. Both two domains have 51 vertical layers.

The initial and lateral boundary conditions of the WRF model are 0.5°×0.5° Global Forecast ~~system~~ System (GFS) data sets 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 ~~in the 6 months~~ 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 WRF model since various observations are assimilated by GFS. No reflectivity assimilation has been performed here since we investigate the impacts of symmetric error model on the climatology of representation error. The model equivalents have 12 hour time interval (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:

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

where Z_{er}, Z_{es} and Z_{eg} are reflectivity factor for rain, snow and graupel droplets, respectively. This diagnostic algorithm (Stoelinga 2005) employs $8{\times}10^6$, $2{\times}10^7$ and $4{\times}10^6$ m^{-4} as intercept parameters of rain, snow and graupel droplets, respectively.

145 The densities of rain, snow and graupel droplets are 1000, 100 and 400 kg m^{-3}, respectively. More details of this diagnostic algorithm, including the densities and intercept parameters, can be found in Stoelinga (2005). The Unified Post Processor (UPP) package (https://dtcenter.org/community-code/unified-post-processor-upp) interpolates diagnostic reflectivities from the coordinates of 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 water, snow water

150 and graupel water 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 CVMRs capture the southwest–northeast rain belt with strong convective cells in area A, illustrating that WRF model is capable to simulate this convective system. The CVMRs and their model equivalents still presents discrepancy in the comparison of Fig. 2a and Fig. 2b. As shown in Fig. 2c, the OmBs can vary widely from place to

155 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 CVMRs. Thus, regardless of 1 km CAPPIs or CVMRs, the model equivalents are misplaced, ill-shaped, or have erroneous intensities when compared to observations point by point. Followed by Geer and Bauer (2011), we also refer all these errors to 'mislocation' error. The mislocation errors of 1 km CAPPIs and CVMRs can result in the non-Gaussian error distribution that violates the Gaussian

160 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 dBZ 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

165 numbers along the abscissa imply the large mislocation error of CVMRs resulting from considerable missed simulations. By comparing with the satellite radiance departures (Fig. 5 in Migliorini and Candy, 2019), these considerable missed simulations are associated with the worse spatial discontinuity in OmBs of CVMRs. For convenience we refer to the discontinuous scenario as 'any-reflectivity'.

To examine effects of the large mislocation error on the error structure of CVMRs, we removed all missed simulations and

170 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). It could be interpreted as the comparison of Fig. 3a and Fig. 3b showing that the reflectivity in 'any-reflectivity' has a more complicated error structure than 'both-reflectivity', illustrating that non-Gaussian error distribution in radar reflectivity assimilation is likely to be stronger than that in satellite radiance assimilation.
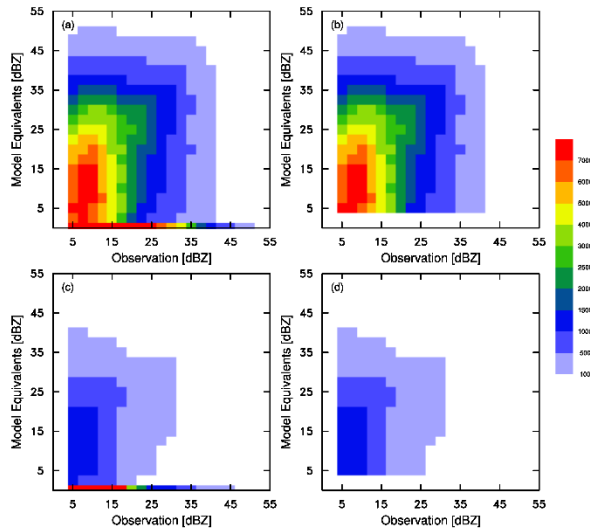
Figure 3: Histograms of observed (a and b) CVMRs and (c and d) 1 km CAPPIs (abscissa, unit: dBZ) against their model equivalents (ordinate, unit: dBZ) in 'any-reflectivity' (the first column) and 'both-reflectivity' (the second column) scenarios.

The sample numbers of 1 km CAPPIs decrease to 232681 and 71516 in '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 1 km CAPPIs estimated by OmBs is similar to CVMRs.

It is critical to understand statistical features of several OmBs by examining their probability density functions (PDFs) before building the symmetric error model. By comparing 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 in 'both-reflectivity' (blue solid line) is apparently closer to the Gaussian distribution. The comparison illustrates that the numerous missed simulations along the abscissa in Fig. 3 give an undesirable effect on some DA and ML algorithms. In practice, the mismatches between observations and simulations provide valuable information related to convective systems. The non-Gaussian distribution cannot be ignored in applications of radar reflectivity.

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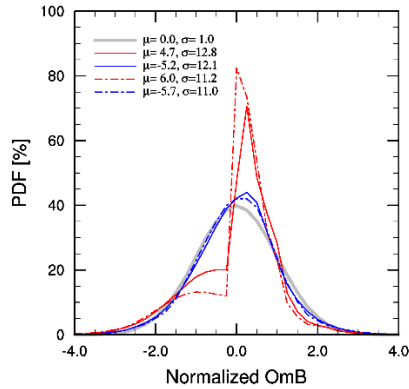
Figure 4: Probability density functions of CVMR (solid lines) and 1 km CAPPI (dash lines) OmBs in 'any-reflectivity' (red lines) and 'both-reflectivity' (blue lines) scenarios, normalized by the mean and standard deviation of the whole sample. The gray line represents the normal Gaussian distribution.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

3 Predictors of symmetric error model

3.1 Predictor derived from reflectivity

The predictors of previous symmetric error models in 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 estimating rain rate I (unit: mm h^{-1}) from reflectivity factor Z_e (unit: mm^6 m^{-3}):

$$Z_e = aI^b \tag{2}$$

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

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

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

Figure 5 shows the distributions of rain rate data derived from observations and simulated by WRF model. Despite some disagreements when CVMRs less than 15 dBZ in area A, the rain belt derived from reflectivity factors at 3 km altitude presents a similar southwest–northeast distribution to 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 CVMRs in Fig. 2b. Consequently, the OmBs of rain rates in Fig. 5c agree with the OmBs of CVMRs in Fig. 2c, illustrating that the error structure of CVMRs can be described by the rain rates regardless of the discrepancy between CVMRs and rain rates.
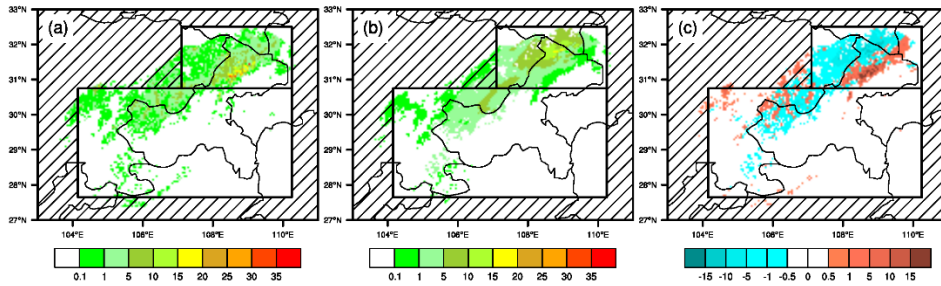
Figure 5: Distributions of rain rates (unit: mm h⁻¹) (a) derived from reflectivity factors (unit: mm⁶ m⁻³) at 3 km altitude, (b) simulated by WRF model and (c) their OmBs at 1800 UTC on August 28th, 2021. The black rectangles indicate the research areas, same as Fig. 1

3.2 Predictors from third-party observations

Derivation from reflectivity factor is not the only way to obtain the 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 National Meteorological Information Center of the China Meteorological Administration (NMIC/CMA). The hourly CMPAS data with 0.05° resolution, merging precipitation observations from rain gauge, radar QPE and satellite QPE, capture a number of details of hourly precipitations and are more accurate than other single source precipitation observations (Pan et al., 2018; Li et al., 2022).

In the comparison of Fig. 5a and Fig. 6a, the CMPAS rain rates are comparable to the derived rain rates, especially for heavy precipitations in area A. Because the 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 mountainous area. A number of small and moderate precipitations in area B are captured by CMPAS rain rates, leading to a wider distribution of OmBs in Fig. 6b. Thus, a more accurate precipitation data can provide more reliable samples in construction of symmetric error model.
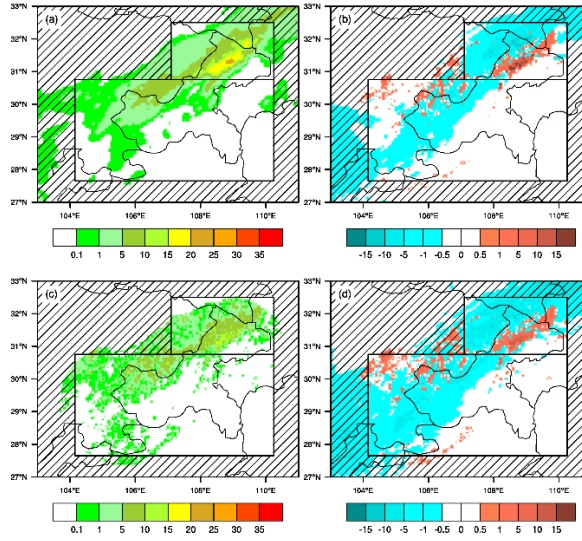
Figure 6: Distributions of (a) CMPAS rain rates (unit: mm h^{-1}) and (c) logarithmic rain rates at 1800 UTC on August 28th, 2021. The (b) and (d) are OmBs of CMPAS rain rates (unit: mm h^{-1}) and logarithmic rain rates, respectively. The black rectangles indicate the research areas, same as Fig. 1

3.3 The linearization of predictor

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

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

where a and b are the coefficients of Z-I relationship. In this study, the Eq. 4 is not a formula to obtain the quantitative equivalent reflectivity accurately. It merely transforms the relationship between 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 shorted by the logarithmic rain rate, as a linear predictor. Adding 1.0 on rain rate ensures that the base of logarithm is greater than zero, same as the 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 logarithmic rain rates in Fig. 6d present similar negative and positive distribution in comparison with derived rain rates in Fig. 5c. It is worth noting that a number of precipitations smaller than 0.1 mm h^{-1} are amplified by above logarithmic transform, resulting in more OmBs of logarithmic rain rates. The logarithmic rain rates allow us to obtain more small precipitation samples.

In order to examine the relationship between CVMR OmBs and symmetric rain rates, it is advisable to count the numbers of CVMR OmBs over the 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 derived rain rates (Fig. 7a) and CMPAS rain rates (Fig. 7b) locate from -20 dBZ to

30 dBZ when the symmetric rain rates 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}. Two peaks are at about 30 dBZ and -10 dBZ respectively.
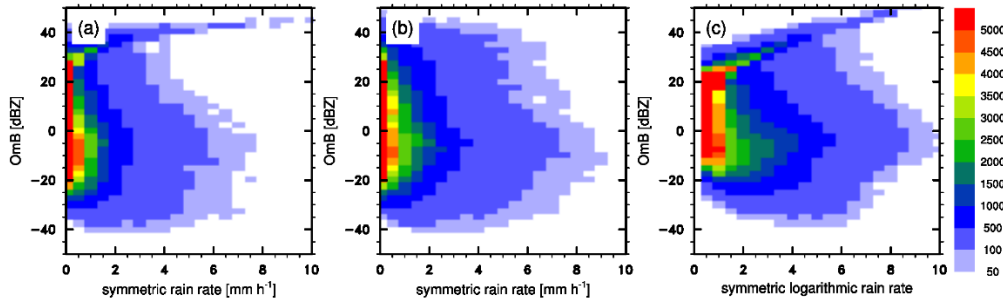
Figure 7: Histograms of CVMR OmBs (ordinate, unit: dBZ) against symmetric rain rates (abscissa, unit: mm h^{-1}), which are (a) derived by reflectivity factors and (b) computed by CMPAS data, and (c) the symmetric logarithmic rain rate.

In contrast, the major OmBs against CMPAS rain rates in Fig. 7b become a unimodal distribution peaking at about -10 dBZ. Although this unimodal distribution is not symmetric along OmB equals zero, it is closer to Gaussian distribution, confirming that the 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 very gentle along the abscissa. As a result, the logarithmic transformation just reduces the gradient of rain rates without altering the structure of CVMR OmBs.

4 Errors as a function of symmetric rain rates

4.1 The symmetric error model of CVMRs

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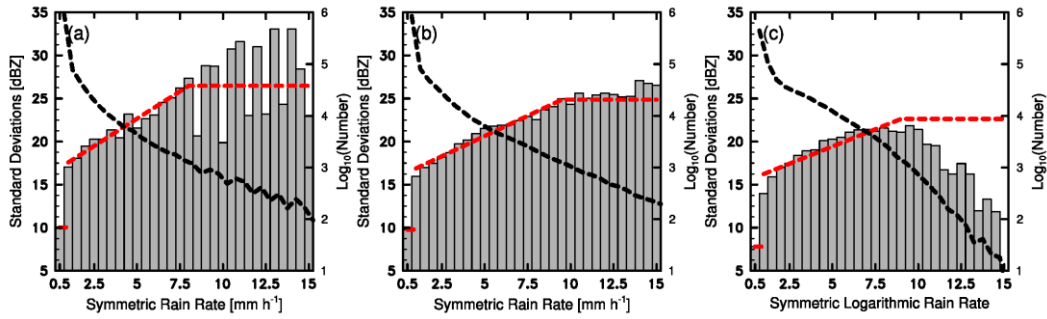
Figure 8: Standard deviations of CVMR OmBs over the symmetric (a) derived rain rates, (b) CMPAS rain rates and (c) logarithmic rain rates. The red dash lines show the three-piecewise fitting functions (listed in Table 1). The black dash lines show the logarithm of sample numbers over symmetric rain rate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

To simplify the complex error structure of CVMRs, a three-piecewise function (red dash line) is fitted by using linear regression. Because the first bin has to be isolated to pass the 95% confidence level for F-test. A straight line rather than the 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 dash line), chosen to be larger than 10^3 samples. Table 1 lists key parameters of piecewise functions.

Table 1: key parameters of 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 with symmetric derived rain rates, are captured by the symmetric CMPAS rain rates as well. The standard deviations vary from about 10 to 25 dBZ when the symmetric CMPAS rain rates increase from 1 to 9.5 mm h^{-1}. The small variation of standard deviations after 10 mm h^{-1} results from the superior representation of CMPAS data. For the symmetric logarithmic rain rates (Fig. 8c), the standard deviations of CVMR OmBs grow gradually from roughly 14 to 21 dBZ 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 decrease trend at the tail of the logarithmic rain rate (lager than 9.0) results from the rapid decline of sample size. The straight line prevents

295 the three-piecewise fitting function from an irrational linear regression. According to Table 1, the logarithmic rain rates obtain the smallest slope of fitting function among three symmetric predictors, despite of the smallest R^2.

4.2 Improvements on Gaussianity

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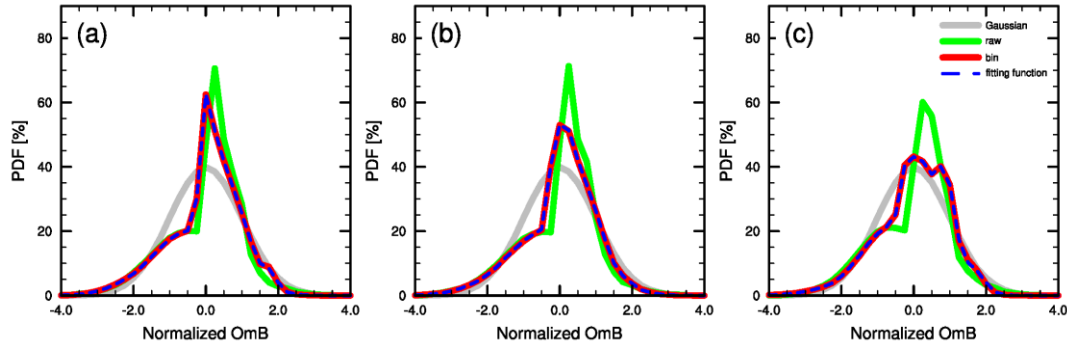
Figure 9: Probability density functions (PDFs) of CVMR OmBs normalized by symmetric (a) derived rain rates, (b) CMPAS rain rates and (c) logarithmic rain rates. The green, red, blue and gray lines represent the raw, binned, three-piecewise and normal
310 **Gaussian PDFs, respectively.**

To quantify the similarity between the PDFs normalized by the symmetric rain rates and 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 PDFs normalized by symmetric rain rates or raw standard deviations and Q represents the normal Gaussian
315 PDF. The JSD is zero means distributions P and Q are the same. For the derived rain rates, the JSDs of PDFs normalized by the binned standard deviations and the three-piecewise function can decrease from 0.010 to 0.006.

Table 2: the 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 at central area. The CMPAS rain rates also obtain smaller JSDs than the derived rain rates as listed in Table 2. It demonstrates that the accuracy of CMPAS rain rates can further improve the Gaussianity of PDFs. For the logarithmic rain rates (Fig. 9c), the PDFs normalized by the binned standard deviations and three-piecewise function also approximate to the normal Gaussian distribution by comparing with the raw PDF. It is worth noting that the logarithmic rain rates obtain the smallest JSDs in spite of a few fluctuations on the PDFs normalized by the binned standard deviations and three-piecewise function.

5 Conclusions

In this study, the Gaussianity of two OmB data, including the CVMRs and 1 km CAPPIs, are examined in the southwest of China. Their features, such as horizontal distributions and PDFs, are similar regardless of the different definitions between 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 how to attack the non-Gaussian PDF.

In the comparison of '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. Because 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 researches 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 dBZ. Yet the instrument noise of radar is of order 1 dBZ.

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 rains and the inaccurate diagnostic algorithm of radar reflectivity in some DA and ML applications. As the discussion in Geer and Bauer (2011), using the symmetric error model in reflectivity assimilation may also compensate for an inadequate specification of hydrometeors in background error, 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 the other bins, illustrating that a more complex structure, the three-piecewise function, should be built at convective-allowing scale.

By comparing 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. Because each subgroup of CVMR OmBs, separated by symmetric derived rain rates, approximates to Gaussian PDF in spite of the non-Gaussian PDF of the whole 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.

Effects of a 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 similar three-piecewise function to the derived rain rates, the superior representation can further improve the Gaussianity of CVMR OmBs in terms of the JSDs calculated by PDFs in Table 2.

The logarithmic rain rates give profound effects on the symmetric error model of CVMR OmBs. Not only the gradients of standard deviations of CVMR OmBs become gentle from the second bin, but the PDFs normalized by the binned standard deviations and the three-piecewise function also obtain the smallest JSDs by comparing with 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 to build the symmetric error model of CVMRs.

In theory, the symmetric error models of CVMRs built in this study are more consistent to the fundamental principle in 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 a number of experiments to discuss the effects of symmetric error models on some DA and ML algorithms is also encouraged. An immature usage of symmetric error model is briefed here:

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

where RR_{avg} means the symmetric rain rate, σ_l and σ_u are the lower and upper boundaries of reflectivity error, respectively. The 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 future, it is of interest to add the effects of ice-phased hydrometeors on the symmetric error model of CVMRs. The polarized measurements and their combinations that provide additional information about hydrometeors may be the solutions.

Code and data availability

The observations, simulations and derived rain rates are available at https://doi.org/10.6084/m9.figshare.25093508.v1. The graphics were generated using 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 System (UPP) for WRF is also available at GitHub
380 (https://github.com/NOAA-EMC/UPP).

Author contribution

YG conceptualized this study and plotted all figures. LH and YG computed the Observation minus Background data sets and built the error model 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 authors.

385 **Competing interests**

The authors declare that they have no conflict of interest.

Acknowledgements

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

390 **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.
395 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*
400 *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.

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.

405 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 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
410 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 Simulations by Using the Symmetric Rain Rate [Dataset]. figshare, https://doi.org/10.6084/m9.figshare.25093508.v1, 2024.

Geer, A. J., and Bauer, P.: Observation errors in all-sky data assimilation. *Q J R Meteorol Soc*, 137, 2024–2037,
415 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.

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
420 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.

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.

425 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 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,
430 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, 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,
435 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,

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. 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.

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.

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 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 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.

470   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.

475   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.

480

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.

485