Posselt Derek (Orcid ID: 0000-0002-5670-5822)

Nonlinear Data Assimilation for Clouds and Precipitation using a Gamma-Inverse Gamma Ensemble Filter

Derek J. Posselt[1] and Craig H. Bishop[2]

[1]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA*
[2]*Naval Research Laboratory, Monterey, CA, USA*

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.3374

Corresponding author address: Derek J. Posselt, Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109. E-mail: Derek.Posselt@jpl.nasa.gov

Abstract

Where clouds occur, their water content is always positive definite, and may be near zero. In addition, it is common for errors in remote sensing observations of clouds and rainfall to be represented as a fraction of the measurement. Furthermore, there is nonlinearity in the relationships among cloud environment, cloud microphysical processes, and the amount and distribution of cloud and precipitation. For these reasons, data assimilation algorithms that rely on linearity and assumptions of Gaussian probability distributions may have difficulty in assimilating observations in cloudy regions, as well as producing an analysis that realistically represents the actual distribution of clouds and precipitation. A recently developed ensemble filter algorithm, the Gamma, Inverse-Gamma, and Gaussian Ensemble Kalman Filter (GIGG-EnKF), allows for fractional observation errors and positive definite quantities. As such, it has promise for producing more effective and accurate data assimilation and retrievals of clouds and precipitation.

This study evaluates the effectiveness of the GIGG-EnKF by using it to estimate the values of warm rain cloud microphysical parameters in a cloud model using observations of precipitation. GIGG-EnKF estimates of cloud parameters and rainfall are compared with a Bayesian reference solution returned by a Markov chain Monte Carlo algorithm, and with a perturbed-observations Ensemble transform Kalman filter (ETKF). The ETKF produces an ensemble with a substantial number of negative (non-physical) precipitation rates and parameter values. In contrast, the GIG-EnKF precipitation analysis is positive definite, and is able to adapt

to changes in the observation value (and observation uncertainty). Non-linearity in the parameter-precipitation relationship is handled by implementing an iterative outer loop regression from the observation space to state space. The positive definite constraint and natural accommodation of fractional error, along with the iterative outer loop, produces an accurate reproduction of the Bayesian analysis of the precipitation rate and cloud microphysical parameter values.

Key words: Data Assimilation, Cloud Microphysics, MCMC, Ensemble Kalman Filter, Gamma, Inverse-Gamma, Non-Gaussian, Nonlinear, Convection

1. Introduction

Estimation of cloud properties and precipitation rates from remote sensing observations, and their use in modern forecast and data assimilation (DA) systems, can be challenging. In some cases, there exist relatively simple relationships between cloud properties and remote sensing measurements. For example, cloud optical depth and cloud droplet effective radius at cloud top are known to have a robust and simple relationship to visible and near infrared reflectances (Nakajima and King 1990; Nakajima et al. 1991; King et al. 1992; Platnick et al. 2001). Over the global oceans, the vertically integrated amount of liquid water (liquid water path) is closely related to low-frequency microwave brightness temperatures (Greenwald et al. 1993; Weng and Grody 1994). However, characterization of cloud vertical structure and surface precipitation rate is more difficult, especially over land, and for clouds containing a mixture of liquid and ice (Delanoe et al. 2010; Posselt and Mace 2014; Wood et al. 2014). The use of DA to estimate cloud and precipitation properties has proven even more challenging, as model representations of cloud processes are approximate, and the atmospheric state – cloud process relationships are nonlinear. The models that relate cloud and precipitation properties to remote sensing observations are also nonlinear, which adds to the complexity of the problem.

Recent work has demonstrated progress in the use of cloud-affected observations in DA systems (Vukicevic et al. 2004; Otkin 2010; Polkinghorne and Vukicevic 2011; Geer et al. 2018), and various other studies have shown promise for the assimilation of cloud and precipitation properties on small scales (e.g., Jung et al. 2008; Li and Mecikalski 2010;2012;

Putman et al. 2014; Li et al. 2017). However, fundamental problems remain, and they stem from the theory that underlies modern forecast and DA systems. In nearly every modern DA algorithm, information from measurements, along with a prior estimate of the state, is quantified by assigning it a probability distribution. The analysis is then computed as the (often approximate) solution to Bayes' relationship for conditional probabilities; including an estimate of the centre of mass (e.g., the mean, median, or mode) of the distribution, along with a measure of its dispersion (e.g., variance, inter-quartile range, etc.). The large dimensionality of most geophysical systems makes direct computation of the posterior distribution in Bayes' relationship intractable (e.g., Snyder et al. 2008; van Leeuwen 2009). For this reason, most modern DA algorithms make approximations that simplify the form of the probability distributions. Of these, ensemble based Kalman Filter (EnKF; Houtekamer and Mitchell 1998; Anderson 2001; Bishop et al. 2001; Tippett et al. 2003; Evensen, 2009) methods have proven to be particularly useful, as they are able to accommodate a range of different state variables, allow the use of nonlinear models, and return the error variance minimizing estimate (as long as the model is linear and the probability distributions are additive and Gaussian).

In our previous work (Vukicevic and Posselt 2008; Posselt and Bishop 2012 (PB12); Posselt et al. 2014 (PHB14)), we showed that the Gaussian assumption and linear update in EnKF framework returns a flawed representation of the Bayesian posterior distribution when the control variables are positive definite quantities, and when the state – observation space relationship is nonlinear. Adaptations to EnKF, in the form of quadratic nonlinear regression

(Hodyss 2011;2012), provide an improved approximation of the analysis mean and variance (PHB14). The nonlinear ensemble transform Kalman filter (Todter and Ahrens, 2015; Todter et al., 2016) and particle filter (PF) algorithms make fewer assumptions about the form of the probability distributions, and, as such, are capable (in theory) of accommodating highly nonlinear models (van Leeuwen 2009; Ades et al. 2015; Poterjoy 2016). However, in these, and in nearly all applications of ensemble based DA methods, observation errors are assumed to be fixed, and do not scale with the measurement value. In the case of clouds and precipitation, the observation uncertainties are more appropriately expressed as a fraction of the observation itself (Huffman 1997; Deng and Mace 2006; Posselt and Mace 2014; Tang et al. 2014; Wood et al. 2014). In this case, it is more correct to specify a constant *relative observation error* variance. The inverse of the relative variance is equivalent to the scale parameter of the gamma distribution (Bishop and Satterfield, 2013), and it follows that an algorithm that allows a gamma or inverse gamma observation likelihood may be more effective for assimilation (or retrieval) of observations with fractional errors.

Bishop (2016) introduced a variation on the perturbed observations form of the EnKF that enables it to utilize Gamma, Inverse-Gamma as well as Gaussian uncertainties. This Gamma, Inverse-Gamma, and Gaussian (GIGG)-EnKF filter accommodates positive-definite variables, as well as observation error standard deviations that scale with the true value of the observed variable. The GIGG-EnKF is a special case of Anderson's (2003, Section 2c) two stage multi-variate ensemble filter. The first stage of Anderson's filter produces an ensemble-based estimate of the posterior distribution of a single observed variable, given a prior ensemble distribution and an observation with known error characteristics. The second stage uses linear regression to estimate the posterior distribution of the extended model-plus-observation state vector given the new posterior distribution of the observed variable whose observation was just assimilated. The two stages are repeated until all observations have been assimilated.

In Anderson's filter, the first stage assumes Gaussian uncertainty distributions for both the prior forecast and the observation. This approach is inaccurate when the prior and observational uncertainty are better described in terms of skewed positive definite distributions such as gamma, inverse-gamma and log-normal pdfs such as those depicted in Figure 1 of Bishop (2016). In contrast, the GIGG-EnKF provides very accurate solutions to Bayes' theorem when the prior forecast pdf and the *likelihood* pdf of observations given truth come in one of the following pairs:

(i) The GIG case, in which the prior is a gamma pdf and the observation likelihood is an inverse-gamma.

(ii) The IGG case, in which the prior pdf is inverse-gamma and the observation likelihood is gamma.

(iii) The G or Gaussian case, in which both the prior and observation-likelihood pdfs are Gaussian.

For a detailed description of the GIGG-EnKF, we refer the reader to Bishop (2016), who compared the performance of the GIGG-EnKF with that of traditional EnKF using an idealized dust uptake model. Here, we investigate the potential utility of the GIGG-EnKF for cloud and precipitation applications by using it to estimate positive definite cloud microphysical parameters in a 1D cloud model using observations of precipitation rate. The performance of the GIGG-EnKF is compared with both the EnKF and an accurate solution to Bayes' theorem found using a Markov Chain Monte Carlo (MCMC; Posselt 2016) approach. This example allows us to test the GIGG-EnKF in a low-dimensional system that has direct application to both precipitation retrievals and ensemble DA for cloud variables. Specific aims of the study include:

1. Evaluate the potential utility of the GIGG-EnKF for cloud and precipitation assimilation and remote sensing retrievals.
2. Explore the manner in which the observation-space and state-space solution changes with observation value to illustrate the properties of the GIGG-EnKF.
3. Determine whether the GIGG-EnKF solution, which is analytically robust in observation space, returns a robust estimate of the state-space ensemble. Demonstrate how an outer loop formulation can be used to ensure consistency between state and observation space analysis distributions.

The remainder of this paper is structured as follows. Section 2 contains a description of the cloud model and the MCMC algorithm that is used to find the reference posterior distribution. Results from the MCMC, EnKF, and GIGG-EnKF analyses are presented in section 3 and a discussion of these results is contained in section 4. A summary of the study and the major conclusions are presented in section 5, and two appendices detail the changes made to MCMC to accommodate gamma and inverse gamma distributions (Appendix A), as well as a description of the GIGG-EnKF algorithm (Appendix B).

2. Methodology

2a. Model and observations

In this study, we aim to test the efficacy of the GIGG-EnKF for cloud and precipitation estimation for a system that is simple (low-dimensional), but sufficiently sophisticated so as to represent the complexity in real cloud models. Posselt and Vukicevic (2010; PV10) described a single column version of the NASA Goddard Cumulus Ensemble (GCE) model (Tao et al. 2003, Tao et al., 2014) that emulates the changes experienced by an atmospheric column during the evolution of a deep convective cloud system. This simplified model was designed to explore uncertainty in cloud microphysical parameters in isolation from any feedback to the cloud-scale dynamics. The model includes a single-moment bulk cloud microphysical scheme, with two liquid and three ice hydrometeor species (Lin et al. 1983, Rutledge and Hobbs 1983, 1984, Tao et al. 2003, Lang et al. 2007). Clouds are generated in the column model by specifying time-varying vertical profiles of vertical motion and water vapour tendency, and advection only operates on vertical distributions of cloud liquid and ice condensate. Clouds interact fully with both longwave and shortwave radiative transfer. By specifying appropriate base state potential temperature and water vapour, along with time-varying vertical profiles of vertical motion and water vapour tendency, PV10 forced the model to emulate flow through a deep convective squall line. The model is attractive in that it has the full realism of a three-dimensional cloud-scale model, while being computationally efficient enough to allow large ensembles of realizations (van Lier-Walqui et al. 2012; 2014).

In PV10, PB12, and PHB14, 10 cloud microphysical parameters were used as control variables in the ensemble analysis. In this study, we focus on warm clouds, and use the rain particle size distribution slope intercept (N_{0r}) and the threshold mass mixing ratio for conversion of cloud to rain (q_{c0}) as our control variables. Both were shown in our previous work to exert a strong influence on convective precipitation. Increasing the autoconversion threshold delays the onset of rain, while increasing the slope intercept of the rain PSD leads to the production of more small rain droplets. As such, increases in both parameters lead to decreases in rain rate, and vice versa. The ranges of variability for each model parameter are taken from observations of rain particle size distributions in the tropics (Tokay and Short, 1996; Roy et al., 2005). In these studies, evidence can be seen for a distribution of rain slope intercept that is approximately gamma in nature. We specified gamma prior distributions for N_{0r} and q_{c0} accordingly (Table 1).

Our sole observation consists of precipitation rate (PCP) at a single time, as would be the case in assimilation of precipitation from remote sensing measurements over observation-sparse regions. To focus on warm rain processes, we select a time 40 minutes into the simulation, during which only liquid cloud was present. We generate three simulated measurements of precipitation rate by selecting values for each warm rain parameter so that the model produces a low (< 1 mm hour^{-1}), moderate (5 mm hour^{-1}) or relatively high (10 mm hour^{-1}) precipitation rate

40 minutes into the simulation (Table 2). An ensemble of 1000 simulations, generated from 1000 draws from the gamma distribution of N_{0r} and q_{c0} (Table 2), yields a positively skewed distribution of precipitation rates for which there is a finite probability of zero precipitation (Fig. 1, grey filled boxes). Consistent with observed precipitation rates, the prior distribution of precipitation in our ensemble has more qualitative similarities with a gamma or inverse-gamma pdf than a Gaussian pdf. In our conceptual framework, we know in advance that precipitation has occurred. As such, we include in the prior only those ensemble members with rain rates exceeding 0.001 mm hour^{-1}. The resulting 818 member ensemble was used as the prior in both the EnKF and GIGG-EnKF experiments. The mean and variance of the prior ensemble of precipitation rates are listed in Table 3 (column 1).

2b. Fractional Observation Error and Relative Observation Error Variance

For many state variables (e.g., vector wind components, temperature, geopotential height) it is reasonable to assume that the observation error variance is independent of the value of the observation. However, in cloud remote sensing and DA, evidence suggests that errors in the observations are not, in fact, constant, and instead vary with the quantity observed. For example, errors in retrieved cloud properties obtained from radar are best represented as a fractional uncertainty (Wood et al., 2014), and uncertainties in visible and near infrared optical depth and effective radius are commonly presented in terms of fractional error (Platnick et al. 2001). Precipitation uncertainties have been shown to be proportional to the rain rate in global estimates of rainfall (Huffman 1997), with fractional uncertainties ranging from approximately 50% to 300% for satellite derived rain rate estimates (Tang et al. 2014). Geer and Bauer (2011) and Harnisch et al. (2016) showed that errors scaled with the cloud amount in microwave imager and infrared brightness temperatures, respectively. In variational (Optimal Estimation; Rodgers 2000) cloud retrievals and DA algorithms, the observation error variance is commonly assumed to be constant, which likely leads to mis-characterization of the uncertainties in the estimates. In ensemble DA, EnKFs typically assume that the observation error variance is independent of the true value of the variable. In contrast, the GIGG-EnKF algorithm allows for the use of a constant *fractional uncertainty*, or relative observation error variance.

In this study, the distribution of observed precipitation rates given a true precipitation rate is assumed to be inverse-gamma for the GIGG-EnKF algorithm. The mean is set equal to the true value, and the relative/fractional error standard deviation is assumed to be 1.0 (100% uncertainty), consistent with estimates of uncertainty in satellite based rain rate retrievals (Tang et al. 2014). Since it is the relative/fractional observational error variance rather than the absolute observation error variance that is constant, the three precipitation values used in this study (low, medium, and high) correspond to three different observation error variances, with the largest observation error variances corresponding to the largest precipitation rates. To be consistent with the assumption of fixed observation error variance in the EnKF, we compute an average observation error variance by: (i) multiplying each precipitation rate in the prior by its relative error standard deviation (1.0), (ii) squaring each resulting observation error standard deviation to obtain an observation error variance, and (iii) computing the average over all of the observation error variances in the prior ensemble forecast of precipitation rates.

It may be tempting to try to provide the EnKF with an observation error variance that is consistent with the underlying unknown true value by, for example, using an available proxy for the true value to turn the known relative error variance into an estimate of the error variance. Possible proxies for the true value include the observed value, or some weighted average of the observed value and the prior ensemble mean. Once the proxy for the true value is available, one can find the observation error variance associated with it, by multiplying the type 1 relative observation error variance by the square of the value of the proxy. However, Bishop (2018) shows that (i) giving an EnKF the exact observation error variance associated with the unknown truth results in far larger analysis errors than using the three step method outlined above, and (ii) the derivation for the (Kalman) gain matrix that minimizes analysis error variance actually explicitly requires that the observation error covariance matrix **R** used in the EnKF be the covariance of the prior distribution of observation errors associated with the prior distribution of truth. Thus, the three step procedure outlined above gives the EnKF a close approximation to the observation error covariance matrix that would minimize analysis error variance – particularly when the ensemble is large. In effect, use of a scaled observation error variance in the EnKF makes the algorithm more susceptible to noisy observations.

2c. Analysis framework

As was mentioned in the introduction, we employ a MCMC algorithm that produces a sample from the true Bayesian posterior distribution, then compare this reference solution with the analyses obtained from the GIGG-EnKF and the EnKF. Below, we provide a brief description of each algorithm and refer the reader to more complete theoretical treatments in the references cited.

1) BAYESIAN MCMC ALGORITHM

DA is fundamentally concerned with information. Specifically, the goal is to update prior information about the state of a system with new information from observations that are obtained from the true state. Bayes' theorem compactly summarizes the combination of various pieces of information via the relationship between conditional probabilities

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{L(\mathbf{y} \mid \mathbf{x}) P(\mathbf{x})}{P(\mathbf{y})} \tag{1}$$

where $\mathbf{x}$ is the set of microphysical parameters and $\mathbf{y}$ are the specific values of the observations. $P(\mathbf{x})$ is a pdf that represents prior knowledge of the control parameters $\mathbf{x}$, $L(\mathbf{y}|\mathbf{x})$ is the likelihood function that describes the distribution of error prone observed values for a given true state $\mathbf{x}$, and $P(\mathbf{y})$ is a normalizing factor that consists of the integral of the joint PDF of $\mathbf{x}$ and $\mathbf{y}$ $P(\mathbf{x},\mathbf{y})$ over all possible $\mathbf{x}$, holding $\mathbf{y}$ constant. Computation of the Bayesian posterior distribution requires specification of the form of the prior and likelihood distributions. Once these distributions are assigned, most approximate solutions to Eq. (1) also assume the posterior distribution has a specific distribution (e.g., Gaussian; Vukicevic and Posselt 2008). Markov chain Monte Carlo (MCMC) algorithms are completely general, and require no particular assumption as to the form of any of the distributions in Eq. (1). Rather, given a prior and likelihood, MCMC methods produce a sample of the Bayesian posterior probability distribution via a random walk (a Markov chain) that revisits regions in the posterior parameter space with high probability, while avoiding regions with low probability (Tamminen and Kyrola 2001; Tarantola 2005; Posselt et al., 2008; PV10; PB12; Posselt 2013; PHB14, Posselt 2016). In each step of the chain, candidate values of all parameters are randomly generated from a "proposal" distribution that depends on the current parameter estimate. These candidate parameter values are used to generate simulated measurements, which are in turn used in an evaluation of the likelihood distribution. Comparison of the prior and likelihood with the values generated at the previous step in the chain determines whether the candidate parameter values are stored as a sample of the posterior distribution.

In our experiments, we let the prior cloud microphysical parameters (N_{0r} and q_{c0}) be distributed gamma (Table 2) consistent with what we expect from theory (e.g., Tokay and Short 1996). Use of a gamma distribution allows for a probability density function with a hard bound at zero, and a long tail that accommodates a wide range of possibilities. For consistency with the assumptions employed in the GIGG-EnKF, we assume the observation likelihood is distributed inverse gamma with a 100% relative error standard deviation (Tang et al. 2014). As mentioned above, either the gamma or the inverse gamma distribution is a more accurate depiction of the uncertainties in remote sensing observations of clouds and precipitation, as both can be constructed to easily accommodate a fixed *relative* error variance (fractional uncertainty). An expanded description of the MCMC algorithm, along with modifications required to accommodate gamma and inverse gamma distributions, is contained in Appendix A.

2) ENSEMBLE KALMAN SMOOTHER

We use the perturbed observations version of the Ensemble Transform Kalman Smoother (ETKS) that is described in detail in PB12. In our experiments, the ensemble size K is 818, while the total number of observations p is 1. To be consistent with the assumptions used in GIGG-EnKF and MCMC, we multiplied each precipitation rate value in the prior sample by 1.0 to obtain the error standard deviation associated with the fixed relative error specified for GIG and MCMC. These values were then squared, and the mean over all values of error variance (32.6 mm^2 $hour^{-2}$) was used as the fixed observation error variance in the EnKF experiment for all values of precipitation rate. This average observation error variance is the constant that minimizes its mean square difference about the true observation error variance, and is broadly consistent with the rigorous quantification of retrieved rainfall uncertainty from the Tropical Rainfall Measuring Mission (TRMM, L'Ecuyer and Stephens 2002, 2003; Table 1), and from multi-sensor satellite based precipitation estimates (Tang et al. 2014). It is also the error variance that minimizes the posterior error variance in the EnKF framework (Bishop 2018).

3) THE GIGG-ENKF ANALYSIS ALGORITHM

Bishop (2016) introduced an ensemble DA algorithm that accurately estimates the observation-space solution to Bayes theorem for cases in which the prior and likelihood are distributed: gamma – inverse gamma, inverse gamma – gamma, or (the standard) Gaussian – Gaussian case. Here, we focus solely on the case of a Gamma prior and Inverse-Gamma observation likelihood (GIG). In Appendix B, we briefly summarize how the GIG produces a posterior analysis ensemble $y_i^a, i = 1, 2, \cdots, K$ of the observed variable y. In the GIGG-EnKF form, the information in the analysis space ensemble is passed from the observation-space analysis y_i^a to the m^{th} model variable x_m^f using linear regression. Specifically, if x_{mi}^f gives the value assigned to x_m^f by the i^{th} ensemble member then the GIGG-EnKF updates this value using

$$x_{mi}^a = x_{mi}^f + \frac{\operatorname{cov}\left(x_{mi}^f, y_i^f\right)}{\operatorname{var}\left(y_i^f\right)}\left(y_i^a - y_i^f\right), i = 1, 2, \cdots, K \tag{2}$$

where y_i^f is the result of mapping $\mathbf{x}_i^f$ into observation space. Hereafter, we will refer to this regression-based initial state space analysis as the GIG linear analysis to distinguish it from the outer-loop based nonlinear analysis we introduce later in section 4. The gamma (G, Eq. A.3) prior is defined by k and θ, the distribution shape and scale factors, respectively. The mean of a gamma distributed random variable x_g is $E[x_g] = k\theta$ and the variance is $Var[x_g] = k\theta^2$. Using the definition of the mean, we note that $Var[x_g] = E[x_g]^2 / k$; the variance is proportional to the square of the mean.

The inverse-gamma (IG, A.4) likelihood is defined by its shape α and scale β parameters. In the case of a variable y_{ig} distributed inverse gamma, the mean and variance are $E[y_{ig}] = \beta / (\alpha - 1)$ and $Var[y_{ig}] = \beta^2 / [(\alpha - 1)^2 (\alpha - 2)]$. Note that, as in the case of the gamma distribution, the inverse gamma variance is also proportional to the square of its mean $Var[y_{ig}] = \left(E[y_{ig}]\right)^2 / (\alpha - 2)$. If observations are assumed to be unbiased, so that the value of the observation is equal to the mean of the likelihood distribution, *then the observation error variance will scale with the value of the observation itself*. As in the gamma prior, the observation error variance scales with the square of the mean of the observation. As noted above, this is a desirable property for observations of clouds and precipitation, which are often better characterized by fractional errors. Bishop (2016) showed that, when the prior and likelihood distributions are paired G – IG, the Bayesian posterior distribution in observation space will be distributed gamma.

In our study, we use exactly the same *K*=818 prior ensemble members and *p*=1 observations as are employed in the EnKF, and use the perturbed observations GIG variant of the GIGG-EnKF to assimilate the single precipitation rate observation. Because the true observation likelihood pdf used by the MCMC is an inverse-gamma pdf, it follows that the GIG case is more appropriate than the IGG case. (We do not consider the Gaussian case of the GIGG-EnKF because it is identical to the regular EnKF). We refer the reader to Bishop (2016) for the complete derivation of the GIGG algorithm.

3. Results

3a. Precipitation Rate Analysis

Draws from the gamma prior distribution of model cloud microphysical parameters produce a skewed and bimodal ensemble distribution of precipitation rates, with one mode near zero, and another at approximately 7 mm hour^{-1} (Fig. 1, grey filled bars). The mean, variance, and skewness of the prior distribution of rain rates are listed in the first column of Table 3, and the maximum precipitation rate produced by the prior parameter distribution is 17.1 mm/hour. Recall that, although the true observation error standard deviation is 1.0 times the true value being observed, the average of the observational error variance over the prior is 32.6 mm^2 hour^{-2} for the EnKF. For the GIG filter, the true type 1 relative error variance of R^r =1.0 (100% fractional error) is used. Both the gamma and inverse gamma distributions become increasingly symmetric as the relative error variance decreases. With a type 1 relative error variance of 1.0, the inverse gamma likelihood distribution will have moderate positive skewness. Likelihood distributions are most consistent between MCMC, GIG, and EnKF when the observation values are moderate to large, and as such we show the results for the relatively high precipitation rate case first, moving then to the moderate and finally the low precipitation rate cases.

MCMC generates a weighted combination of prior and likelihood entirely consistent with Bayes' theorem (Eq. 1). When the prior is multiplied by the likelihood, it produces a solution with a posterior mode located between the observation and the higher valued prior mode (Fig. 1a, Table 3). The GIG (Fig. 1b) analysis has a shape that is similar to the reference MCMC posterior distribution; it has positive skewness, the mode is in approximately the same location, and there are similar numbers of members with precipitation rates larger than 5 mm hour^{-1}. The posterior mean of the GIG is similar to the MCMC posterior mean, but the ensemble is over-variant (Table 3), due to the presence in the GIG ensemble of a larger number of members with relatively low (< 5 mm hour^{-1}) precipitation rates. This can be directly attributed to the gamma fit to the observation-space prior, which has a mode at approximately 3 mm hour^{-1}. The EnKF posterior distribution is symmetric, with a smaller mean and larger posterior variance than the MCMC reference solution. There appear to be a similar number of ensemble members with low precipitation rates in the EnKF analysis as there are in the GIG; however, because there are fewer members with large (> 10 mm hour^{-1}) precipitation rates, the EnKF variance is smaller.

Use of a constant fractional error means that the observation error variance is smaller for the 5 mm hr^{-1} observation, relative to the 10 mm hr^{-1} observation, leading to smaller posterior variance (Table 3). The GIG algorithm (Fig. 1e) produces an observation space PDF that is closer to the MCMC solution at this time than the EnKF (Fig. 1f), as both the GIG and MCMC algorithms are able to account for the decrease in observation error magnitude associated with observation of a smaller precipitation rate. The GIG posterior distribution remains slightly over-variant, again as a consequence of fitting a gamma distribution with a single mode to the bi-modal prior. Comparison of the GIG posterior analysis with the EnKF analysis clearly shows the GIG algorithm is able to track the change in posterior variance caused by a change in the value of the observation. As previously noted, it is technically feasible to utilize a fractional observation error variance in the EnKF. It is possible that this may yield improved performance in select cases (e.g., for particular weather events, etc). However, when integrating over a great many experiments, the optimal observation error variance for the EnKF is independent of the value of the observation, and is given by the expected value of the observation error variance over the prior distribution of truth (Bishop 2018). As a consequence, the shape of the EnKF posterior distribution (Fig. 1f) is identical to that shown in Fig. 1c (note the change in axis range in Fig. 1d-f vs. Fig. 1a-c). Close inspection of the EnKF analysis ensemble indicates there are a few members with negative precipitation rates in this case.

At near-zero values of the observation, the observation error variance is quite small (0.64 mm^2 $hour^{-2}$ for a 0.8 mm $hour^{-1}$ observation). As such, the posterior variance in observation space is also quite small (Fig. 1g, Table 3). The GIG analysis distribution in observation space (Fig. 1h) is very similar to the MCMC solution in both mean and standard deviation (Table 3). In contrast to the EnKF, the posterior variance in the GIG scales with the MCMC's Bayesian posterior error variance, and *produces no negative precipitation rates in the posterior ensemble.* As noted above, the posterior variance for the EnKF (Fig. 1i) is typically assumed to be independent of the observation value, and the posterior mean is a weighted combination of the prior mean and the observations. Consequently, the posterior distribution in observation space is shifted in the direction of the observation. Because the variance of the posterior analysis does not change, a shift toward lower values results in a sizeable fraction (10%) of the ensemble members with (unphysical) negative values. In contrast, none of the GIG posterior ensemble members are negative.

3b. Cloud Microphysics (State Space) Analysis

The GIG algorithm operates in observation space, and its analysis consists of a posterior distribution of precipitation rates. As in many other DA schemes that do not account for non-linearity, such as EnKFs and variational schemes with no outer loop, multivariate *linear* regression (Eq. 2) is used to obtain the state space distribution of cloud microphysical parameters from the observation-space analysis. In our previous study, the response of precipitation to changes in warm rain parameters appeared to be approximately linear early in the simulation (c.f. Fig. 9 in PV10). As such, one might expect linear regression to return a reasonable state space ensemble.

Before examining the microphysical parameter (state space) distributions, it is worth remembering the physics of the warm rain process. As was shown in Table 1, and discussed in section 2, low values of N_{0r} correspond to relatively larger numbers of large (vs. small) rain drops, and produce large values of precipitation rate. Conversely, high values of N_{0r} correspond to large numbers of small rain drops (and few large drops), and produce small values of precipitation rate. The same is not necessarily true of the autoconversion threshold. Its value sets the threshold at which cloud mass mixing ratio water is reclassified as rain, but the actual rain rate also depends on the rain particle size distribution. This has interesting consequences for the state space analysis, as we shall see in a moment.

MCMC (Figs. 2a – 2c) produces an ensemble of cloud microphysical parameters whose mode is close, but not equal, to the true parameter value for both q_{c0} and N_{0r}. Since the prior distribution is non-uniform (e.g., it has a mode), the posterior mode cannot be equal to the true value unless the prior mode is also located at the true value. The Gamma prior distribution for each parameter peaks at a value that is generally smaller than the pre-set values. It is clear that the covariance between q_{c0} and N_{0r} changes with the value of the observation, with little co-variability at high precipitation rates and strong (negative) co-variability at low precipitation rates. The increased observational constraint provided by the 0.8 mm hour^{-1} case can also be seen in the fact that the posterior mode is nearly coincident with the true values. The shapes of the posterior distributions are consistent with the physics of the warm rain process: relatively large precipitation can be produced for a large range of autoconversion threshold values, as long as there are sufficient numbers of large rain drops (N_{0r} is relatively low). In contrast, if the autoconversion threshold is high, low precipitation rates can only be produced for small values of N_{0r} . In effect, there is a tighter joint constraint on both parameters at low precipitation rates. Note that the univariate (marginal) posterior statistics (Table 4) mask this constraint. In our previous work, we demonstrated the potential pitfalls one may encounter when examining the statistics of marginal distributions (Posselt 2016). In this case, marginal posterior variance for N_{0r} and q_{c0} is larger for low precipitation rates than for high precipitation rates in MCMC, a reflection of the larger *range* of values for both parameters (Figs. 2a – 2c).

The linear regression from precipitation (observation) to parameter (state) space in the GIG (Figs. 2d – 2f) algorithm does not produce distributions of cloud microphysical parameters similar to the reference MCMC posterior distributions. However, there is evidence of increasing constraint on parameter values with decreasing values of observed precipitation. Specifically, the covariance between N_{0r} and q_{c0} increases in magnitude with decrease in precipitation, and the posterior density increases around the mode. While the summary statistics (Table 4) do not reflect this increase in density, plots of the joint distributions indicate that the GIG algorithm is sensitive to changes in the parameter-parameter relationship with changes in the precipitation rate.

The cloud microphysical parameter PDF returned by EnKF (Figs. 2g – 2i), appears to be similar to what one might obtain from an average over all of the MCMC posterior distributions (Figs. 2a – 2c). PHB14 averaged 100 centred MCMC posterior distributions computed for different observations drawn from the same prior, and showed that the EnKF does in fact return an approximation to the mean over all possible posterior analysis distributions. Consistent with its construction, the variance of the EnKF analysis distribution is independent of the value of the observation; the measurement serves only to change the position of the mean.

The mapping from observation to state space is clearly imperfect for the GIG; the regression produces a distribution of cloud microphysical parameters with negative (non-physical) values, especially for moderate and large precipitation rates. Problems with the linear regression from observation space to state space can be illustrated by running the model with every positive (physical) set of parameter values, and comparing the resulting distribution of precipitation rates with the observation-space distributions shown in Fig. 1. Since the model cannot be run with non-physical parameter values, we neglect all negative values in the state space analysis. We refer to the distribution of precipitation rates produced by the posterior distribution of microphysics parameters as the *forward analysis*. The resulting forward precipitation rate ensembles are shown for GIG and EnKF in Fig. 3 for all three observation values. The MCMC results are shown for reference, but note that MCMC produces a state space analysis that is perfectly consistent with the observation space analysis.

When high and moderate precipitation rates are observed (Figs. 3a - 3c and 3d – 3f, respectively), the ensemble of precipitation rates produced by the state-space GIG and EnKF microphysical parameter analyses (in red) are similar to the original precipitation rate distributions (in black). When the relatively large fraction of non-physical (negative) N_{0r} and q_{c0} values (Fig. 2) are removed from the GIG and EnKF posterior ensembles, the remaining members form a state space ensemble that is not dissimilar to the Bayesian solution (Figs. 2a and 2b). As such, the observation space values are similar to the observation space ensemble produced by each of the two algorithms (Fig. 1). The exception is the larger number of ensemble members with precipitation rates less than 2 mm hour^{-1}. This is consistent with the fact that large N_{0r} values (greater numbers of small droplets) produce smaller precipitation rates, and there are more members with large N_{0r} values in the GIG and EnKF ensembles than in the MCMC analysis.

With a decrease in observed precipitation rate to low values (Figs. 3g - 3i), the forward precipitation rate ensemble produced by the analysed microphysics parameters the GIG (Fig. 3h) and EnKF (Fig. 3i) increasingly differ from the observation-space (precipitation rate) solution. While the GIG returns a forward analysis distribution that is reasonably similar to the observation-space analysis for high and moderate precipitation rates, the regression from observation to state space fails for the low precipitation rate observation (Fig. 2f). The increasing number of ensemble members that produce a zero precipitation rate in the GIG and EnKF is due to the fact that high values of N_{0r} and q_{c0} generate small precipitation rates, and vice versa. The ensemble members with large precipitation rates in the GIG forward ensemble come from state space members with small q_{c0} and N_{0r}. The large number of near-zero precipitation rates are due to larger q_{c0}, N_{0r}, or both.

4. Improved State Space Analysis Using an Ensemble Outer Loop

4a. Functional Form of the Parameter – Precipitation Relationship

The linear regression from observation to state space produces a state space analysis that is inconsistent with the observation space solution. If the relationship between N_{0r}, q_{c0} and precipitation rate were linear, then the regression should yield a set of parameter values that produce a statistically similar observation-space ensemble. The parameter-precipitation relationship can be examined by plotting the response of the simulated precipitation to changes in each parameter, while the other is held fixed. The response functions are shown in Fig. 4 for

N_{0r} (Fig. 4a) and q_{c0} (Fig. 4b). N_{0r} bears a qualitatively similar relationship to precipitation for multiple different values of q_{c0}. Precipitation rate decreases with increasing N_{0r}, independent of q_{c0}. However, the linearity in the relationship is a function of the value of q_{c0}. As q_{c0} increases, N_{0r} is increasingly nonlinearly related to precipitation rate. In addition, as q_{c0} increases, sensitivity to changes in N_{0r} increases at small N_{0r} values, and decreases at large N_{0r} values. This is the reason for the high values of posterior marginal variance for N_{0r} (Table 4). At relatively large N_{0r} values, there is decreased sensitivity in precipitation to changes in N_{0r}, and what sensitivity there is may be offset by a change in q_{c0}.

In contrast to N_{0r}, there is a fundamental change in the q_{c0} - precipitation response function as N_{0r} decreases (Fig. 4b). At relatively high values of N_{0r}, the response is non-monotonic, with precipitation first increasing with q_{c0}, then decreasing. At relatively low values of N_{0r}, the response is still nonlinear, but is monotonic and positive. Clearly, linear regression will not adequately capture the range of responses. This is reflected in the fact that there is little sensitivity in the q_{c0} posterior distributions to changes in observation value (Table 4). The regression relationship in the prior appears to favour the inverse N_{0r} – precipitation rate response.

Changes in the response function for different values of q_{c0} and N_{0r} are reflected in the Bayesian posterior joint state space PDF in Figs. 2a-2c. These plots demonstrate that, as the true precipitation rate decreases in magnitude, the structure of the posterior PDF changes as well. As the precipitation rate decreases, the covariance between the state space variables increases. At the same time, the range of values of the two state space variables that produce near identical precipitation rates (as reflected in the size of the region of high posterior probability) grows. At low values of precipitation, there is a narrow range of q_{c0} values that will produce reasonable rain rates for any given value of N_{0r}; however, there is a very large range of *paired* values that will produce low precipitation rates. The GIG algorithm is able to represent some form of this change in posterior covariance between state space variables, but the regression from precipitation rate to N_{0r} and q_{c0} is clearly imperfect.

4b. Ensemble Outer Loop: Enforcing Consistency Between Observation and State Space in the Presence of Non-Linearity

The inaccuracy of the linear regression from observation to state space has implications for the production of a forecast ensemble from the state space analysis. If the state space analysis is incorrect, the forecast ensemble will also be incorrect, even if the full nonlinear model is used to generate the forecast ensemble. Clearly, when the state and observation space are nonlinearly related, linear regression will lead to inconsistencies. When the response changes sign (non-monotonic nonlinearity), the problems are particularly acute. What is needed is a way to (1) enforce consistency between the observation and state space analysis in a way that preserves nonlinearity in the relationships while respecting the merits of the GIG observation and model space state estimates, and (2) a method for dealing with non-physical state-space parameter values (which cannot be used to run the model). There are a variety of ways one might envisage doing this, and the authors believe that there is a strong need for more research in this area aimed at identifying approaches that result in posterior distributions that are as close as possible to the

Bayesian posterior. In what follows, we describe an *ad-hoc* approach for addressing the above concerns that yields promising results. We offer it in the hope that it will inspire research into more rigorous approaches. We wish to point out at the outset that the methodology described below is, in theory, applicable to any ensemble DA algorithm that involves nonlinear mapping from observation to state space.

One way of enforcing observation – state space consistency, while respecting the merits of the GIG observation and model space state estimates, is to apply the observation to state space regression in stages. The goal is to retain the GIG observation-space analysis, while iteratively adjusting the GIG model space ensemble members obtained from linear regression. Ultimately, this iterative procedure produces a *non-linear* GIG model space analyses that precisely maps onto the GIG observation space analyses. Let the vector $\mathbf{x}_i^n$ contain the model state variables N_{0r} and q_{c0} at the nth iteration for the ith ensemble member, let $H\left(\mathbf{x}_i^n\right)$ denote the forecast precipitation given by $\mathbf{x}_i^n$, $H\left(\mathbf{x}_j^n\right)$ denote the forecast precipitation given by $\mathbf{x}_j^n$, and let y_i^a denote the GIG analysis of the observed precipitation rate. Now consider the following equation that generates a new *guess* $\mathbf{x}_i^{n+1}$ of the model state/parameter variables from the old guess $\mathbf{x}_i^n$:

for $n = 1:N$

for $i = 1:M$

$$\mathbf{x}_i^{n+1} = \mathbf{x}_i^n + \left(\mathbf{PH}^T\right)_i^n \left[\left(\mathbf{HPH}^T\right)_i^n\right]^{-1} \left[y_i^a - H\left(\mathbf{x}_i^n\right)\right]$$

end (3)

end

where $\left(\mathbf{PH}^T\right)_i^n = \sum_{j=1}^{M} w_{ij}^n \left(\mathbf{x}_j^n - \mathbf{x}_i^n\right)\left[H\left(\mathbf{x}_j^n\right) - H\left(\mathbf{x}_i^n\right)\right]^T$,

$$\left(\mathbf{HPH}^T\right)_i^n = \sum_{j=1}^{M} w_{ij}^n \left[H\left(\mathbf{x}_j^n\right) - H\left(\mathbf{x}_i^n\right)\right]\left[H\left(\mathbf{x}_j^n\right) - H\left(\mathbf{x}_i^n\right)\right]^T.$$

Here, M denotes the total number of ensemble members, while N is the total number of outer loop iterations. After N iterations, the state space GIG analysis ensemble $\mathbf{x}^a$ is set equal to $\mathbf{x}^N$. Note the similarity of the first equation to the equation for the update of the ensemble mean in an ensemble Kalman filter. Note also y_i^a does not change from one iteration to the next and that the typical EnKF covariances used to compute the covariance matrices $\left(\mathbf{PH}^T\right)^n$ and $\left(\mathbf{HPH}^T\right)^n$ have been replaced by *weighted* covariances about the current guess field $\mathbf{x}_i^n$. The weights w_{ij} are based on the distance of each of the ensemble members from $\mathbf{x}_i^n$ using a formula inspired by the typical formula for the weights given to ensemble members in a particle filter (e.g., Eq. 14 in van Leeuwen 2009) of the form

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\mathbf{x}_j^n - \mathbf{x}_i^n\right)^T \mathbf{D}_w^{-1}\left(\mathbf{x}_j^n - \mathbf{x}_i^n\right)\right]\left\{\sum_{k=1}^{M} \exp\left[-\frac{1}{2}\left(\mathbf{x}_j^n - \mathbf{x}_i^n\right)^T \mathbf{D}_w^{-1}\left(\mathbf{x}_j^n - \mathbf{x}_i^n\right)\right]\right\}^{-1}, \quad (4)$$

where $\mathbf{D}_w = \begin{pmatrix} d_{N_{0r}}^2 & 0 \\ 0 & d_{q_{c0}}^2 \end{pmatrix}$

Here, $d_{N_{0r}}$ and $d_{q_{c0}}$ are tuneable positive definite distance parameters that control the number of members that contribute to the local regression in state space. Note that $\sum_{j=1}^{M} w_{ij} = 1$; i.e. the weights to sum to unity. Also note that in the limit as $d_{N_{0r}}, d_{q_{c0}} -> \infty$, $w_{ij} = (1/M)$ for $j = 1, 2, ..., M$; i.e. all ensemble members would receive equal weight in the covariance calculation. In contrast, in the limit as $d_{N_{0r}}, d_{q_{c0}} -> 0$, the ensemble member that is closest to member i gets a w_{ij} weight of unity, whereas all other members have a weight of zero. As a reasonable measure of the number of ensemble members that play a significant role in the computation of the covariances, the effective ensemble size can be computed as $K_{eff}^{i} = 1 \Big/ \sum_{j=1}^{K} w_{ij}^{2}$. Note that $K_{eff}^{i} = M$ in the limit as $d_{N_{0r}}, d_{q_{c0}} -> \infty$ and $w_{ij} = (1/M)$ for $j = 1, 2, ..., M$ whereas $K_{eff}^{i} = 1$ in the limit as $d_{N_{0r}}, d_{q_{c0}} -> 0$.

In the first iteration (n=1), $\mathbf{x}_i^1$ is set equal to the model space values obtained using the standard EAKF linear regression (Eq. 2), provided this linear regression has yielded physically plausible values for the parameter values. If, for some ensemble member, the linear regression yields a physically implausible value, the kernel method described below is used to find a nearby physically plausible value. Note that the use of EAKF linear regression for the first guess in the iteration ensures that the approach delivers the GIG *linear* analysis given by Eq. (2) in the case that the mapping from model space to observation space was linear. For a 2-parameter state space, in theory, just two parameters would be required to define the local regression from state to observation space. We tested a range of effective ensemble sizes, and obtained the best performance for a relatively low K_{eff}^{i} value between 4 and 7. In practice, we start by setting $d_{N_{0r}}, d_{q_{c0}}$ equal to 0.1 times the variance of the prior guess analysis ensemble in state space $\mathbf{x}^n$, then adjust them until the effective ensemble size is less than or equal to 7. It was found that convergence, defined as $H\left(\mathbf{x}_i^n\right) \cong y_i^a$, was obtained within 2 to 5 iterations, and further iterations had no substantial effect on the model state estimate. Furthermore, at convergence, the model space analysis for the *i*th ensemble member is precisely consistent with the GIG analysis y_i^a for the *i*th ensemble member in observation space. Since the convergence rate differed slightly among ensemble members, we fixed the total number of iterations at N=10 – even though we did not see cases where more than 5 iterations were required for convergence.

In addition to ensuring consistency between the observation space and state space analyses, it is necessary to deal with negative (unphysical) state space (model parameter) values produced by the regression. This happens primarily in the initial regression from observation space to state space, which uses the entire ensemble (all members weighted equally), but also may occur for members with parameter values near zero. To maintain a robust sample, it is desirable to replace non-physical members with physically reasonable parameter values. We chose to employ the following particle filter resampling methodology for each unphysical state space member. First, draw a new state space member from the previous analysis $\mathbf{x}^n$ (in the first iteration, this is simply the GIG linear analysis). Second, perturb each parameter using a

Gaussian kernel with mean equal to the unphysical values associated with this member, and standard deviation set using the so-called rule-of-thumb bandwidth selector (Wand and Jones, 1995). This kernel standard deviation comes from kernel density estimation (KDE) theory, and is the bandwidth that minimizes the mean squared error of the KDE, if the underlying distribution is Gaussian. The equation that defines the rule of thumb bandwidth h_x for state variable x is

$$h_x = \left(\frac{4\hat{\sigma}_x^{\ 5}}{3K} \right)^{\frac{1}{5}}, \tag{5}$$

where $\hat{\sigma}_x$ is the standard deviation of the ensemble for parameter x, and K is the number of ensemble members. We can see from Eq. (5) that the bandwidth h_x is proportional to the sample standard deviation $\hat{\sigma}_x$ divided by $K^{0.20}$. In our case, then, $h_x \approx \hat{\sigma}_x/4$. Note that, if the perturbation drawn from the kernel results in a non-physical parameter value, then we simply draw another perturbation.

We implement the weighted regression and particle filter resampling, and cycle the weighted outer loop procedure until the state space ensemble produces an ensemble in observation space that converges to the GIG posterior observation-space analysis. Results of the iterative regression are shown for the observation space ensemble in Figs. 5d – 5f, and for the state space ensemble in Figs. 6d – 6f. For comparison, we reproduce the observation space result from the GIG that results from running the state space analysis through the model in Figs. 5a – 5c, and the MCMC solution in Figs. 5g – 5i and Figs. 6g – 6i. There is a visible improvement in the consistency between the state space ensemble (red) and the observation space analysis (black). After application of the outer loop, the state-space analysis now produces an observation-space ensemble that is nearly identical to the ensemble produced by the GIG analysis.

The improvement may be quantified by evaluating the difference between the MCMC reference distribution and the EnKF or GIG analysis. Of the many available metrics, we have chosen to use the integrated absolute difference (IAD; Posselt et al. 2012). For two frequency distributions $f_1(x)$ and $f_2(x)$ defined on discrete bins (i = 1, N), the IAD may be computed as:

$$IAD = \frac{1}{2}\sum_{i=1}^{N}\left|f_1(x)_i - f_2(x)_i\right|, \tag{6}$$

where $f_1(x)_i$ is the histogram value in the ith bin, and it is assumed that (1) both histograms are defined on equivalent ranges and with equivalent bin sizes, and (2) each histogram has been normalized so that it sums to 1. The IAD is similar to an L1 norm, and, as such, is less sensitive to outliers than a root mean squared difference (RMSD, L2 norm). When multiplied by a factor of 1/2, as it is in Eq. (6), the IAD has the added benefit of an intuitive interpretation, in that an IAD of 1.0 indicates a 100% difference between distributions (zero shared probability mass), an IAD of 0.5 indicates 50% overlap, etc. In all of the results that follow, IADs are multiplied by 100 so that they represent the percent difference between histograms. IAD values for the observation space analysis are plotted in Table 5, with the differences all computed with respect to the MCMC posterior distribution. All histograms had 51 bins spanning a range from [-5,15] mm/hour for the low and mid precipitation case, and from [-5,30] mm/hour for the high precipitation case. Tests were run in which the number of bins ranged from 40 to 70, and the IAD in each case changed by no more than 0.03 (3%) across all cases.

The prior (left-most column in Table 5) serves as the reference, and places a theoretical upper bound on the IAD for each analysis. The EnKF analysis constitutes an improvement over the prior for high and mid precipitation cases, with the low precipitation case a net degradation, as the posterior IAD is ~10% higher than the prior. The GIG returns a lower IAD than the prior for all cases, and a lower IAD than EnKF for every case but the high precip case. In this case, the two IAD values are within 2% of each other. As this is a smaller difference than that obtained by changing histogram bin numbers, we deem it not to be significant. In most cases, the observation-space distribution that results from running the state space analysis through the model (columns containing $F(\mathbf{x})$ in Table 5) is worse (has a higher IAD) than the observation-space analysis. The GIG outer loop (right-most column, Table 5) produces a state space analysis that maps nearly exactly into the observation space, as demonstrated by the fact that the IAD for the observation space analysis (column 4, Table 5) and the outer loop (column 6, Table 5) differ by less than 3%.

The state space linear regression GIG analysis, the analysis produced by iterative (outer loop) regression, and the Bayesian posterior analysis produced by MCMC, are shown in Fig. 6. The state space analysis produced by the outer loop (Figs. 6d – 6f) appears closer in form to the Bayesian posterior distribution (Figs. 6g – 6i) than does the analysis produced by the initial linear regression (Figs. 6a – 6c). There is little change in the posterior mean for the GIG outer loop, as compared with the linear regression; however, the posterior variance is smaller for every case (Table 4). The variance of the outer loop GIG is closer to the MCMC solution than the GIG regression analysis for moderate and high precipitation values, and under-variant in the low precipitation rate case. Examination of the IAD for the state space solutions (Table 6) reveals the EnKF posterior analysis (Figs. 2d - 2f) provides a modest improvement over the prior. The GIG analysis distribution that results from a single regression from observation to state space represents an improvement over both the prior and EnKF, but differences between EnKF and GIG are relatively small for all but the low precipitation case. Application of the outer loop (Figs. 6d – 6f, Table 6) results in an analysis that is quantitatively closer to the MCMC solution than either the first stage GIG regression or the EnKF in all cases. The results indicate that application of an iterative weighted regression makes it is possible to recover a close approximation to both the observation and state space analysis using the GIG filter.

5. Summary and Conclusions

DA for clouds and precipitation is challenging for several reasons: observations are positive definite, precipitation processes are nonlinear, and measurements do not have fixed uncertainties. In fact, many observations of cloud and precipitation related variables have errors that scale with the value of the observation itself: the relative observation error variance is fixed, rather than the observation error variance itself. Ensemble-based Kalman filter algorithms utilize the nonlinear model, but have been shown to have problems constraining positive definite variables that may be close to zero (Posselt et al. 2014). In contrast, a recently developed ensemble analysis technique (the GIG filter introduced by Bishop (2016)) is based on the assumption that the prior and likelihood are distributed gamma and inverse gamma, respectively. The GIG analysis naturally accommodates positive definite observations and state variables, as well as observations with fractional uncertainties. To test the efficacy of the GIG for cloud process applications, we conducted experiments in which two cloud microphysical parameters (the intercept of the rain particle size distribution and the cloud – rain autoconversion threshold) were constrained using an observation of precipitation rate. The reference Bayesian posterior analysis generated by a MCMC algorithm was compared with analyses produced by a perturbed-obs EnKF algorithm, as well as the GIG filter. The primary conclusions of our study include the following:

1. For cases in which the posterior distribution is skewed and bounded at zero, the GIG filter produces an ensemble of precipitation rates that is closer to the Bayesian posterior than the EnKF.
2. The EnKF analysis ensemble is insensitive to the value of the observation, and the EnKF produced an observation-space solution that included negative precipitation rates, especially for low observed precipitation values. In contrast, the GIG filter did not produce any negative precipitation rates in the observation space posterior distribution, and provided a closer match to the Bayesian solution than the EnKF.
3. The GIG produces its analysis ensemble in observation space, and requires a transformation to obtain the state space estimate. In contrast to the EnKF, the GIG was

able to approximate the changes in the state space posterior exhibited by the MCMC solution for varying measurement values. However, because the functional relationship between state and observation space changes (and changes sign) for different cloud microphysical parameter values, a regression-based transformation produced a state space parameter ensemble that did not match the MCMC posterior, and was inconsistent with the observation-space analysis.

4. An iterative weighted regression from observation to state space, similar in concept to a variational outer loop, produced a state-space solution for the GIG that was entirely consistent with the observation space solution and quite similar to the Bayesian posterior distribution.

The results illustrate the benefit of the GIG algorithm for computing the Bayesian analysis for observations with a skewed distribution bounded at zero. It returns an observation space estimate that is positive definite, and it has the added benefit of naturally accommodating observations with uncertainties that are proportional to the true measurement value. For these reasons alone, the GIG has particular promise for DA and retrievals of clouds and precipitation, as well as concentrations of pollutants, aerosol, and trace gasses. Implementation of an iterative weighted outer loop regression enforced dynamical consistency between observation and state space, and produced a state space posterior distribution that closely approximated the Bayesian MCMC state space analysis. It is possible that such a weighted regression may be useful in any ensemble-based DA algorithm used for nonlinear observation-state relationships. Because the outer loop requires only the repeated evaluation of the observation operator for each new estimate of the state space ensemble, the computational cost is expected to be reasonable. We wish to emphasize that, while the ensemble outer loop procedure works well for this specific case, we do not prove that the theory holds generally. We leave this proof for future work.

Acknowledgements

DJP was supported by U.S. Office of Naval Research Grant N00173-14-1-G907, while CHB was supported by NRL Base Program, PE0601153N. A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Appendix A: MCMC Algorithm, GIG Implementation

As was documented in section 2b, MCMC algorithms evaluate Bayes relationship between conditional probabilities. This amounts to computation of the right hand side of Eq. (1), which is the product of the prior and likelihood distributions, normalized by the integral over the union of prior and likelihood distributions. Since it is straightforward to numerically normalize a sample of the posterior distribution, it is common to evaluate the right hand side of Eq. (1) up to a constant so that Bayes theorem can be written

$$P\left(\mathbf{x} \mid \mathbf{y}^{o}\right) \propto L\left(\mathbf{y}^{o} \mid \mathbf{x}\right) P\left(\mathbf{x}\right) \tag{A.1}$$

where we have used the superscript *o* on the observation-space variables to indicate the use of observations. Markov chain Monte Carlo (MCMC) algorithms sample the non-normalized posterior probability distribution $P(\mathbf{x} \mid \mathbf{y}^o)$ by constructing a random walk (the Markov chain) in which proposed new posterior points are either accepted or rejected. The accept/reject criterion is chosen to ensure that the density of accepted points will converge to the true posterior density. Let the current set of state space variables (location) in the Markov chain be denoted $\mathbf{x}_l$. Each new set $\hat{\mathbf{x}}$ of $\mathbf{x}$ is made to depend on the previous set (Markovian) by drawing candidate values from a proposal distribution $q(\hat{\mathbf{x}} \mid \mathbf{x}_l)$ centred on the current set of values $\mathbf{x}_l$. For each new set, the (non-normalized) posterior density at $\hat{\mathbf{x}}$, as given by Eq. (A.1), is compared with the posterior density at the current location of the Markov chain $\mathbf{x}_l$. Depending on the ratio of these two densities, the proposal $\hat{\mathbf{x}}$ is either accepted or rejected as a sample from the posterior distribution. The dispersion (or variance) of the proposal distribution determines how large, on average, perturbations to $\mathbf{x}_l$ are in the Markov chain. The model $\hat{\mathbf{y}} = H\left[M\left(\hat{\mathbf{x}}\right)\right]$ is used to generate candidate observation space proposals from the candidate parameter values $\hat{\mathbf{x}}$ for use in evaluating the observation likelihood density $L(\mathbf{y}^o \mid \hat{\mathbf{x}})$ associated with the proposal $\hat{\mathbf{x}}$. It is important to keep in mind the fact that the candidate values $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}} = H\left[M\left(\hat{\mathbf{x}}\right)\right]$ are *not* draws from the prior; rather, they are test values only. The direction of the random walk (and, as such, the generation of the posterior sample) depends on whether proposed posterior particle $\hat{\mathbf{x}}$ is accepted or rejected. . This decision is made by comparing the value of Eq. (A.1) for $\hat{\mathbf{x}}$ with that of $\mathbf{x}_l$. As shown by (Tamminen and Kyrölä, 2001; Posselt, 2013), the MCMC Markov chain will succeed in constructing the posterior pdf if the probability *P* of accepting $\hat{\mathbf{x}}$ into the archive of posterior particles is made to be equal to (Tamminen and Kyrölä, 2001; Posselt, 2013):

$$P\left(\hat{\mathbf{x}} \mid \mathbf{x}_l\right) = \frac{L(\mathbf{y}^o \mid \hat{\mathbf{x}})P(\hat{\mathbf{x}})q(\hat{\mathbf{x}} \mid \mathbf{x}_l)}{L(\mathbf{y}^o \mid \mathbf{x}_l)P(\mathbf{x}_l)q(\mathbf{x}_l \mid \hat{\mathbf{x}})} \ . \tag{A.2}$$

If $P\left(\hat{\mathbf{x}} \mid \mathbf{x}_l\right) \geq 1$ the proposed particle is accepted into the archive of posterior perturbations. However, if $P\left(\hat{\mathbf{x}} \mid \mathbf{x}_l\right) < 1$, the proposed particle is accepted if and only if $u < P\left(\hat{\mathbf{x}} \mid \mathbf{x}_l\right)$ where *u* is a realization from a uniform distribution between zero and unity. This *Bernoulli trial* is what causes MCMC to pause at posterior points for numbers of step attempts that are, on average, proportional to the posterior density of the point in question and thus *sample* the posterior distribution. Note that the proposal distribution *q* appears in both the numerator and denominator to allow the possibility of a non-symmetric proposal (e.g., one that favours lower parameter values over higher). If the proposal distribution is symmetric, then $q(\mathbf{x}_l \mid \hat{\mathbf{x}}) = q(\hat{\mathbf{x}} \mid \mathbf{x}_l)$, and Eq. (A.2) depends only on the prior and likelihood distributions. This is the case in the current work, as the proposal distribution is assumed to be Gaussian, with fixed variance that is tuned so that the acceptance rate (the ratio of accepted parameter sets to the total candidate draws) is approximately 25% (Posselt 2013).

In previous work, we had assumed that the prior was Uniform, thus reducing Eq. (A.2) to a ratio of observation likelihoods and we also assumed the likelihood distribution was Gaussian.

In this paper, we use a more sophisticated MCMC algorithm that employs a gamma prior for the model state parameters and an inverse gamma observation likelihood pdf to represent observational uncertainty. The gamma prior probability density function for each parameter x_m can be written

$$P(x_m) = \frac{1}{\Gamma(k)\theta^k} x_m^{k-1} e^{-\left(\frac{x_m}{\theta}\right)}, \tag{A.3}$$

where “ is the gamma function, and *k* and ¸ are the shape and scale parameters, respectively. In our experiments, we assume parameters are mutually uncorrelated *a priori*, and have chosen specific shape and scale parameters for each state space variable that produce a prior distribution consistent with what we expect from theory and observations. Note, however, that if the prior is assumed to be distributed gamma, and the mean and standard deviation are known, then the shape and scale factors can be computed as: $k = \frac{\mu^2}{\sigma^2}$ and $\theta = \frac{\sigma^2}{\mu}$. Specific values of *k* and θ used to define the prior for N_{0r} and q_{c0} can be found in Table 2.

To represent observational uncertainty we assume that the distribution of imperfect observations y^o given the true value y of the observed variable is inverse gamma. Specifically,

$$L(y^o \mid \mathbf{x}) = \frac{\left(y\,\tilde{R}^r\right)^{-\left[\left(\tilde{R}^r\right)^{-1}+1\right]} \left(y^o\right)^{-\left[\left(\tilde{R}^r\right)^{-1}+2\right]}}{\Gamma\left[\left(\tilde{R}^r\right)^{-1}+1\right]} e^{-\left[\left(\frac{y}{y^o}\right)\left(\tilde{R}^r\right)^{-1}\right]} . \tag{A.4}$$

where, in our case, y is the true rainfall given by the model parameters $\mathbf{x}$ and y^o is the observed rainfall, $\tilde{R}^r$ is the constant type 2 relative observation error variance given by

$\tilde{R}^r = \dfrac{\mathrm{var}\left(y^o - y\right)}{y^2 + \mathrm{var}\left(y^o - y\right)} = \dfrac{R}{y^2 + R}$ where R is the observation error variance. Because $\tilde{R}^r$ is assumed to be independent of y, it follows that $R = \dfrac{y^2}{\left(\tilde{R}^r\right)^{-1} - 1}$ and hence the observation error standard deviation is proportional to the true value y. This corresponds to the assumption of a constant type 1 fractional observation error variance $R^r = \dfrac{R}{y^2} = \left[\left(\tilde{R}^r\right)^{-1} - 1\right]^{-1}$, as is commonly the case in remote sensing retrievals of cloud properties and precipitation (Posselt and Mace, 2014; Posselt et al. 2017). Given the Type 1 relative error variance R^r, the Type 2 relative error variance required for Eq. (A.4) is given by $\tilde{R}^r = \left[\left(R^r\right)^{-1} + 1\right]^{-1}$, and we may then use Eq. (A.4) to evaluate the likelihood for a particular candidate set of parameters $\hat{\mathbf{x}}$ and actual observation y^o.

Appendix B: The GIG Filter

As discussed in Bishop (2016), the GIG assumes that the prior distribution of forecasts y^f of the observed variable y in observation space can be approximated by a gamma distribution of the form

$$P(y^f) = \frac{1}{\Gamma\left[\left(P_y^r\right)^{-1}\right]} \left[\frac{\left(P_y^r\right)^{-1}}{\left\langle y^f \right\rangle}\right]^{\left(P_y^r\right)^{-1}} y^{f\left[\left(P_y^r\right)^{-1} - 1\right]} e^{-\left[\left(P_y^r\right)^{-1} \frac{y^f}{\left\langle y^f \right\rangle}\right]} . \tag{B.1}$$

Here, $P_y^r = \dfrac{P_y^f}{\left\langle y^f \right\rangle^2}$ is the type 1 relative variance of the prior and P_y^f is the standard variance of the prior prediction of the observed variable *y*. Note that the *j* subscript used in Bishop (2016) to indicate which of $j = 1, 2, \cdots, p$ observations is being assimilated has been dropped because in this paper we only assimilate a single observation. In practice, we approximate $< y^f >$ as the prior ensemble sample mean $\overline{y^f}$, and P_y^f as the prior ensemble sample variance.

Observation errors are assumed in the GIG filter to be distributed inverse gamma as in Eq. (A.4). Computation of the posterior distribution, given a gamma prior and inverse gamma likelihood, is obtained by multiplying Eq. (B.1) and Eq. (A.4). Bishop (2016) and others have shown that this results in a gamma posterior distribution of the form

$$P_{\text{post}}(y^a \mid y^o, y^f) = \frac{1}{\Gamma\left(\left(\Pi^r\right)^{-1}\right)} \left(\frac{\left(\Pi^r\right)^{-1}}{\left\langle y^a \right\rangle} \right)^{\left(\Pi^r\right)^{-1}} \left\{ \left(y^a\right)^{\left(\left(\Pi^r\right)^{-1} - 1\right)} \exp\left[-\left(\Pi^r\right)^{-1} \frac{y^a}{\left\langle y^a \right\rangle} \right] \right\}. \tag{B.2}$$

where

$$\begin{aligned} \Pi^r &= \left[\left(\tilde{R}^r\right)^{-1} + \left(\tilde{P}^r\right)^{-1} \right]^{-1} = \tilde{P}^r - \tilde{P}^r \left(\tilde{P}^r + \tilde{R}^r\right)^{-1} \tilde{P}^r, \text{ and} \\ \frac{1}{\left\langle y^a \right\rangle} &= \frac{1}{\left\langle y^f \right\rangle} + \frac{\tilde{P}^r}{\tilde{R}^r + \tilde{P}^r} \left[\frac{1}{y^o} - \left(\tilde{R}^r + 1\right) \frac{1}{\left\langle y^f \right\rangle} \right], \text{ where} \\ \tilde{P}^r &= \frac{\operatorname{var}\left(y^f\right)}{\left\langle y^f \right\rangle^2 + \operatorname{var}\left(y^f\right)} = \left[\left(P^r\right)^{-1} + 1 \right]^{-1}. \end{aligned} \tag{B.3}$$

Bishop (2016) demonstrated that a *K* member analysis ensemble that is consistent with Eq. (B.2) and Eq. (B.3) can be obtained using the equation

$$y_i^{a'}=\frac{y_i^a-\overline{y^a}}{\overline{y^a}}=\frac{\left(y_i^f-\overline{y^f}\right)}{\sqrt{\left(\overline{y^f}\right)^2+\operatorname{var}\left(y^f\right)}}$$

$$+\tilde{P}^r\left(\tilde{P}^r+\tilde{R}^r\right)^{-1}\left[\frac{\left(y_i^{gig}-\left\langle y^{gig}\right\rangle\right)}{\sqrt{\left\langle y^{gig}\right\rangle^2-2\operatorname{var}\left(y^{gig}\right)}}-\frac{\left(y_i^f-\overline{y^f}\right)}{\sqrt{\left(\overline{y^f}\right)^2+\operatorname{var}\left(y^f\right)}}\right], \quad (B.4)$$

For $i=1,2,\cdots,K$. , In Eq. (B.4), y_i^{gig} is the counterpart of the perturbed observations used in EnKFs but it is not a Gaussian variable. It is a random sample from a gamma distribution with mean $\left\langle y^{gig}\right\rangle$ and type 1 relative variance R_{gig}^r given by

$$\left\langle y^{gig}\right\rangle=\frac{\left[\left(\tilde{R}^r\right)^{-1}+2\right]}{\left(\tilde{R}^r\right)^{-1}}y^o \text{ and } \left(R_{gig}^r\right)^{-1}=\left[\left(\tilde{R}^r\right)^{-1}+2\right], \quad (B.5)$$

In practice, we determine the k and theta parameters as $k^{gig}=\left(R^{r^{gig}}\right)^{-1}=\left[\left(\tilde{R}^r\right)^{-1}+2\right]$ and $\theta^{gig}=\overline{y^{gig}}R^{r^{gig}}=y^o\tilde{R}^r$, respectively, then generate random samples of y_j^{gig} from the associated gamma distribution using a gamma random number generator (e.g., gamrnd in MatLAB). Eq. (B.4) is then used to obtain ensemble perturbations, and each ensemble member *i* for each observation is computed from

$$y_i^a=y_i^{a'}\overline{y^a}+\overline{y^a}. \quad (B.6)$$

References

Ades, M. and P. J. van Leeuwen, 2015: The equivalent-weights particle filter in a high-dimensional system. *Q. J. R. Meteorol. Soc.*, **141**, 484–503. doi:10.1002/qj.2370

Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Wea. Rev.*, **129**, 2884-2903.

Anderson J.L., 2003. A Local Least Squares Framework for Ensemble Filtering. *Mon. Wea. Rev.*, **131**, 634–642.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev.*, **129**, 420–436.

Bishop, C. and E. Satterfield, 2013: Hidden Error Variance Theory. Part I: Exposition and Analytic Model. *Mon. Wea. Rev.*, **141**, 1454–1468, doi: 10.1175/MWR-D-12-00118.1.

Bishop, C. H., 2016: The GIGG-EnKF: ensemble Kalman filtering for highly skewed non-negative uncertainty distributions. *Q. J. R. Meteorol. Soc.*, 142, 1395-1412.

Bishop, C. H., 2018: Data assimilation and the inevitable state dependence of the error variance of unbiased observations of bounded variables. *Q. J. R. Meteorol. Soc.*, In Press.

Cho, H.-K., K. P. Bowman, and G. R. North, 2004: A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *J. Appl. Meteor.*, **43**, 1586-1597.

Cohn, S. E., A. da Silva, J. Gua, M. Sienkiewicz, and D. Lamich, 1998: Assessing the Effects of Data Selection with the DAO Physical-Space Statistical Analysis System. *Mon. Wea. Rev.*, **126**, 2913-2926.

Delanoë, J., and R. J. Hogan, 2010: Combined CloudSat-CALIPSO-MODIS retrievals of the properties of ice clouds, *J. Geophys. Res.*, **115**, D00H29, doi:10.1029/2009JD012346.

Deng, M. and G. Mace, 2006: Cirrus microphysical properties and air motion statistics using cloud radar Doppler moments: Part I – Algorithm description. *J. Applied Meteorology and Climatology*, **45**, 1690-1709.

Evensen, G., 2009: Data Assimilation: The Ensemble Kalman Filter. Springer, 279 pp.

Geer, A. J., and P. Bauer, 2011: Observation errors in all-sky data assimilation. *Quart. J. Roy. Meteor. Soc.*, **137**, 2024-2037, doi:10.1002/qj.830.

Geer, A. J., K. Lonitz, P. Weston, M. Kazumori, K. Okamoto, Y. Zhu, E. H. Liu, A. Collard, W. Bell, S. Migliorini, P. Chambon, N. Fourrié, M.-J. Kim, C. Köpken-Watts, and C. Schraff, 2018: All-sky satellite data assimilation at operational weather forecasting centres. *Quart. J. Roy. Meteor. Soc.* Accepted. doi:10.1002/qj.3202

Greenwald, T. J., G. L. Stephens, T. H. Vonder Haar, and D. L. Jackson, 1993: A physical retrieval of cloud liquid water over the global oceans using special sensor microwave/imager (SSM/I) observations, *J. Geophys. Res.*, **98**(D10), 18471–18488, doi:10.1029/93JD00339.

Harnisch, F., M. Weissmann, and A. Perianez, 2016: Error model for the assimilation of cloud-affected infrared satellite observations in an ensemble data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **142**, 1797-1808, doi:10.1002/qj.2776.

Hodyss, D., 2011: Ensemble State Estimation for Nonlinear Systems Using Polynomial Expansions in the Innovation. *Mon. Wea. Rev.*, **139**, 3571–3588.

Hodyss, D., 2012: Accounting for skewness in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 2346-2358.

Hodyss, D., and W. F. Campbell, 2013: Square root and perturbed observation ensemble generation techniques in Kalman and quadratic ensemble filtering algorithms. *Mon. Wea. Rev.*, **141**, 2561-2573.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811.

Huffman, G. J., 1997: Estimates of root-mean-square random error for finite samples of estimated precipitation. *J. Appl. Meteor.*, **36**, 1191–1201.

Husak, G. J., J. Michaelsen, and C. Funk, 2007: Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *Int. J. Clim.*, **27**, 935-944.

Jung, Y., G. Zhang, and M. Xue, 2008: Assimilation of Simulated Polarimetric Radar Data for a Convective Storm Using the Ensemble Kalman Filter. Part I: Observation Operators for Reflectivity and Polarimetric Variables. *Mon. Wea. Rev.*, **136**, 2228–2245, https://doi.org/10.1175/2007MWR2083.1

King, M. D., Y. J. Kaufman, W. P. Menzel and D. Tanre, 1992: Remote sensing of cloud, aerosol, and water vapor properties from the moderate resolution imaging spectrometer (MODIS), *IEEE Trans. Geosci. Rem. Sens.*, **30**, 2-27.

Lang, S., W.-K. Tao, R. Cifelli, W. Olson, J. Halverson, S. Rutledge, and J. Simpson, 2007: Improving Simulations of Convective Systems from TRMM LBA: Easterly and Westerly Regimes. *J. Atmos. Sci*., **64**, 1141-1164.

L'Ecuyer, T. S., and G. L. Stephens, 2002: An estimation-based precipitation retrieval algorithm for attenuating radars. *J. Appl. Meteor*., **41**, 272-285.

L’Ecuyer, T.S., and G. L. Stephens, 2003: The tropical oceanic energy budget from the TRMM perspective. Part I: Algorithm and uncertainties. *J. Climate*, **16**, 1967-1985.

Li, X., and J. R. Mecikalski, 2010: Assimilation of the dual-polarization Doppler radar data for a convective storm with a warm-rain radar forward operator, *J. Geophys. Res.*, **115**, D16208, doi:10.1029/2009JD013666.

Li, X. and J.R. Mecikalski, 2012: Impact of the Dual-Polarization Doppler Radar Data on Two Convective Storms with a Warm-Rain Radar Forward Operator. *Mon. Wea. Rev.*, **140**, 2147–2167, https://doi.org/10.1175/MWR-D-11-00090.1

Li, X., J. R. Mecikalski, and D. J. Posselt, 2017: An Ice-Phase Microphysics Forward Model and Preliminary Results of Polarimetric Radar Data Assimilation. *Mon. Wea. Rev.*, **145**, 683-708, doi: 10.1175/MWR-D-16-0035.1.

Lin, Y.-L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Climate Appl. Meteor*., **22**, 1065-1092.

Nakajima, T. and M.D. King, 1990: Determination of the Optical Thickness and Effective Particle Radius of Clouds from Reflected Solar Radiation Measurements. Part I: Theory. *J.*

Atmos. Sci., **47**, 1878–1893, https://doi.org/10.1175/1520-0469(1990)047<1878:DOTOTA>2.0.CO;2

Nakajima, T., M.D. King, J.D. Spinhirne, and L.F. Radke, 1991: Determination of the Optical Thickness and Effective Particle Radius of Clouds from Reflected Solar Radiation Measurements. Part II: Marine Stratocumulus Observations. *J. Atmos. Sci.*, **48**, 728–751, https://doi.org/10.1175/1520-0469(1991)048<0728:DOTOTA>2.0.CO;2

Otkin, J. A. (2010), Clear and cloudy sky infrared brightness temperature assimilation using an ensemble Kalman filter, *J. Geophys. Res.*, **115**, D19207, doi:10.1029/2009JD013759.

Platnick, S., J. Y. Li, M. D. King, H. Gerber, and P. V. Hobbs, 2001: A solar reflectance method for retrieving the optical thickness and droplet size of liquid water clouds over snow and ice surfaces, J. Geophys. Res., **106**(D14), 15185–15199, doi:10.1029/2000JD900441.

Polkinghorne, R. and T. Vukicevic, 2011: Data Assimilation of Cloud-Affected Radiances in a Cloud-Resolving Model. *Mon. Wea. Rev.*, **139**, 755–773, https://doi.org/10.1175/2010MWR3360.1

Posselt, D. J., and T. Vukicevic, 2010: Robust Characterization of Model Physics Uncertainty for Simulations of Deep Moist Convection. *Mon. Wea. Rev.*, **138**, 1513–1535.

Posselt, D. J., and C. H. Bishop, 2012: Nonlinear parameter estimation: Comparison of an Ensemble Kalman Smoother with a Markov chain Monte Carlo algorithm. *Mon. Wea. Rev.*, **140**, 1957-1974.

Posselt, D. J., 2013: Markov chain Monte Carlo Methods: Theory and Applications. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, 2nd Ed. S. K. Park and L. Xu, Eds., Springer, pp 59–87.

Posselt, D. J., D. Hodyss, and C. H. Bishop, 2014: Errors in Ensemble Kalman Smoother Estimates of Cloud Microphysical Parameters, *Mon. Wea. Rev.*, **142**, 1631-1654.

Posselt, D. J., and G. G. Mace, 2014: MCMC-Based Assessment of the Error Characteristics of a Surface-Based Combined Radar–Passive Microwave Cloud Property Retrieval. *J. Appl. Meteor. Clim.*, **53**, 2034-2057.

Posselt, D. J., 2016: A Bayesian Examination of Deep Convective Squall Line Sensitivity to Changes in Cloud Microphysical Parameters. *J. Atmos. Sci.*, **73**, 637–665.

Posselt, D. J., J. Kessler, and G. G. Mace, 2017: Bayesian retrievals of vertically resolved cloud particle size distribution properties. *J. Appl. Meteor. Clim.*, **56**, 745-765, doi: 10.1175/JAMC-D-16-0276.1.

Poterjoy, J., 2016: A Localized Particle Filter for High-Dimensional Nonlinear Systems. *Mon. Wea. Rev.*, **144**, 59–76, doi: 10.1175/MWR-D-15-0163.1.

Rodgers, C. D., 2000: Inverse Methods for Atmospheric Sounding, Theory and Practice, World Scientific, Singapore

Putnam, B.J., M. Xue, Y. Jung, N. Snook, and G. Zhang, 2014: The Analysis and Prediction of Microphysical States and Polarimetric Radar Variables in a Mesoscale Convective System Using

Double-Moment Microphysics, Multinetwork Radar Data, and the Ensemble Kalman Filter. *Mon. Wea. Rev.*, **142**, 141–162, https://doi.org/10.1175/MWR-D-13-00042.1

Rosmond, T., and L. Xu, 2006: Development of NAVDAS-AR: non-linear formulation and outer loop tests. *Tellus A*, **58**, 45-58.

Roy, S. S., R. K. Datta, R. C. Bhatia and A. K. Sharma, 2005: Drop size distributions of tropical rain over south India. *Geofizika*, **22**, 105-130.

Rutledge, S. A., and P. V. Hobbs, 1983: The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. VIII: A model for the ``seeder-feeder" process in warm-frontal rainbands. *J. Atmos. Sci.*, **40**, 1185-1206.

Rutledge, S. A., and P. V. Hobbs, 1984: The mesoscale and microscale structure and organization of clouds and precipitation in midlatitude cyclones. Part XII: A diagnostic modeling study of precipitation development in narrow cold frontal rainbands. *J. Atmos. Sci.*, **41**, 2949-2972.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to High-Dimensional Particle Filtering. *Mon. Wea. Rev.*, **136**, 4629–4640, https://doi.org/10.1175/2008MWR2529.1

Tamminen, J., and E. Kyrola, 2001: Bayesian solution for nonlinear and non-Gaussian inverse problems by Markov chain Monte Carlo method. *J. Geophys. Res.,* **106**, 14,377-14,390.

Tang, L., Y. Tian, and X. Lin (2014), Validation of precipitation retrievals over land from satellite-based passive microwave sensors. *J. Geophys. Res.*, **119**, 4546–4567, doi:10.1002/2013JD020933.

Tao, W.-K., J. Simpson, D. Baker, S. Braun, M.-D. Chou, B. Ferrier, D. Johnson, A. Khain, S. Lang, B. Lynn, C.-L. Shie, D. Starr, C.-H. Sui, Y. Wang, and P. Wetzel, 2003: Microphysics, radiation, and surface processes in the Goddard Cumulus Ensemble (GCE) model. *A Special Issue on Non-hydrostatic Mesoscale Modeling, Meteorology and Atmospheric Physics*, **82**, 97-137.

Tao, W.-K., S. Lang, X. Zeng, X. Li, T. Matsui, K. Mohr, D. J. Posselt, J. Chern, P. N. Norris, I.-S. Kang, I. Choi, and Y.-M. Yang, 2014: The Goddard Cumulus Ensemble (GCE) Model: Improvements and applications for Studying Precipitation Processes. *Atmos. Res.*, **143**, 392-424.

Tarantola, A., 2005: Inverse Problem Theory and methods for model parameter estimation. SIAM, Philadelphia, pp 342.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble Square Root Filters. *Mon. Wea. Rev.*, **131**, 1485–1490.

Tödter, J. and B. Ahrens, 2015: A Second-Order Exact Ensemble Square Root Filter for Nonlinear Data Assimilation. *Mon. Wea. Rev*., **143**, 1347–1367, doi: 10.1175/MWR-D-14-00108.1.

Tödter, J., P. Kirchgessner, L. Nerger, and B. Ahrens, 2016: Assessment of a Nonlinear Ensemble Transform Filter for High-Dimensional Data Assimilation. *Mon. Wea. Rev*., **144**, 409–427, doi: 10.1175/MWR-D-15-0073.1.

Tokay, A., and D. A. Short, 1996: Evidence from tropical raindrop spectra of the origin of rain from stratiform versus convective clouds. *J. Appl. Meteor.,* **35**, 355-371.

van Leeuwen, P., 2009: Particle Filtering in Geophysical Systems. *Mon. Wea. Rev.*, **137**, 4089–4114, doi: 10.1175/2009MWR2835.1.

van Lier-Walqui, M., T. Vukicevic, and D. J. Posselt, 2012: Quantification of Cloud Microphysical Parameterization Uncertainty using Radar Reflectivity, *Mon. Wea. Rev.*, **140**, 3442-3466.

van Lier-Walqui, M. A., T. Vukicevic, and D. J. Posselt, 2014: Linearization of microphysical parameterization uncertainty using multiplicative process perturbation parameters, Mon. Wea. Rev., 142, 401-413.

Vukicevic, T., and D. J. Posselt 2008: Analysis of the Impact of Model Nonlinearities in Inverse Problem Solving. *J. Atmos. Sci.*, **65**, 2803-2823.T. Greenwald, M. Zupanski, D. Zupanski, T. Vonder Haar, and A.S. Jones, 2004: Mesoscale Cloud State Estimation from Visible and Infrared Satellite Radiances. *Mon. Wea. Rev.*, **132**, 3066–3077, https://doi.org/10.1175/MWR2837.1

Wand, M. P. and M. C. Jones, 1995: Kernel Smoothing. Chapman and Hall, London.

Weng, F., and N. C. Grody, 1994: Retrieval of cloud liquid water using the special sensor microwave imager (SSM/I), *J. Geophys. Res.*, **99**(D12), 25535–25551, doi:10.1029/94JD02304.

Wood, N. B., T. S. L'Ecuyer, A. J. Heymsfield, G. L. Stephens, D. R. Hudak, and P. Rodriguez, 2014: Estimating snow microphysical properties using collocated multisensor observations. *J. Geophys. Res.*, **119**, 8941–8961, doi:10.1002/2013JD021303.

Tables

Table 1. List of parameter values and corresponding precipitation rates for each experiment.

Experiment	N_{0r} (cm^{-4})	q_{c0} (g kg^{-1})	PCP rate (mm hour^{-1})
Low Precipitation	2.714	2.186	0.8
Moderate Precipitation	2.201	0.486	5.0
High Precipitation	0.131	1.642	10.0

Table 2. Parameters of the gamma prior distribution used in the MCMC experiments. Here, k is the shape parameter, while ¸ is the scale parameter.

	μ	σ	k	¸
N_{0r} (cm^{-4})	2.5	2	1.563	1.6
q_{c0} (g kg^{-1})	2	1.5	1.778	1.125

Table 3. Observation-space results from each ensemble experiment for the case in which both q_{c0} and N_{0r} are used as control variables. Rows correspond to the three different precipitation rate experiments, while columns contain results for the prior ensemble (column 1) and each of the ensemble posterior analysis schemes (columns 2-4).

		Mean		
	Prior	MCMC	GIG	EnKF
Low Precip	5.23	1.51	1.77	4.00
Moderate Precip	5.23	7.09	6.34	5.17
High Precip	5.23	8.08	8.39	6.55
		Variance		
	Prior	MCMC	GIG	EnKF

Low Precip	12.52	0.68	0.60	8.82
Moderate Precip	12.52	5.14	7.74	8.82
High Precip	12.52	6.44	13.55	8.82

Table 4. State-space results from each ensemble experiment for the case in which both q_{c0} and N_{0r} are used as control variables. Rows correspond to the three different precipitation rate experiments, while columns contain results for the prior ensemble and each of the ensemble posterior analysis schemes for each parameter.

Mean

	Prior		MCMC		EnKF		GIG		GIG, Outer Loop	
	N_{0r}	q_{c0}	N_{0r}	q_{c0}	N_{0r}	q_{c0}	N_{0r}	q_{c0}	N_{0r}	q_{c0}
Low Precip	2.50	2.00	3.12	1.85	2.29	1.63	2.99	1.71	3.16	1.67
Moderate Precip	2.50	2.00	1.25	1.27	1.93	1.59	1.56	1.55	1.56	1.55
High Precip	2.50	2.00	1.00	1.44	1.49	1.54	0.92	1.47	1.13	1.70

Variance

	Prior		MCMC		EnKF		GIG		GIG, Outer Loop	
	N_{0r}	q_{c0}	N_{0r}	q_{c0}	N_{0r}	q_{c0}	N_{0r}	q_{c0}	N_{0r}	q_{c0}
Low Precip	4.00	2.25	2.26	1.07	1.98	0.99	1.17	0.98	1.11	0.46
Moderate Precip	4.00	2.25	0.65	0.60	1.98	0.99	1.90	0.98	0.96	0.57
High Precip	4.00	2.25	0.50	0.74	1.98	0.99	2.48	0.99	0.88	0.66

Table 5. Integrated absolute difference (%) between MCMC and: prior, EnKF, and GIG for the analyses in observation space, as well as for the state space analysis mapped into observation space for both the one-step regression (italics), and for the GIG outer loop (bold italics). See text for details.

	Prior	EnKF	EnKF $F(\mathbf{x})$	GIG	GIG $F(\mathbf{x})$	GIG $F(\mathbf{x})$, Outer Loop
Low Precip	58.7%	67.8%	*66.7%*	22.7%	*42.5%*	***20.1%***
Moderate Precip	37.0%	28.9%	*30.6%*	18.1%	*20.8%*	***17.9%***
High Precip	42.9%	24.1%	*28.6%*	25.7%	*27.9%*	***24.7%***

Table 6. Integrated absolute difference (%) between MCMC and prior, EnKF, and GIG for the state space analyses, as well as for the state space analysis for GIG that results from the outer loop (bold italics). See text for details.

	Prior	EnKF	GIG	GIG, Outer Loop
Low Precip	86.8%	70.5%	59.1%	***42.4%***
Moderate Precip	51.7%	43.1%	37.4%	***33.2%***
High Precip	55.9%	40.7%	39.3%	***33.6%***

List of Figures

Figure 1. Histograms depicting the prior ensemble (grey) plotted along-side the posterior analysis (open boxes) of precipitation rates (mm hour^{-1}) from (a,d,g) MCMC, (b,e,h) GIG, and (c,f,i) EnKF for observed precipitation rates of 10 mm hour^{-1} (a, b, c), 5 mm hour^{-1} (d, e, f), and 0.8 mm hour^{-1} (g, h, i). The true precipitation rate value in each case is denoted by the dashed vertical line. The heavy solid lines depict the analytical gamma prior used in the GIG experiments (b, e, h), and the Gaussian prior used in the EnKF experiments (c, f, i).

Figure 2. Two-dimensional joint posterior PDFs for the state space parameters for the (left) high, (centre) moderate, and (right) low precipitation cases, as computed by the (a – c) MCMC, (d – f) GIG, and (g – i) EnKF algorithms. Colour filled contours depict the normalized probability, while the unfilled white contours surround regions containing 37% (heavy solid), 68% (solid), 95% (dashed), and 99% (dot-dashed) of the total probability mass. The red lines depict the true parameter values used in each experiment.

Figure 3. As in Fig. 1, these plots contain histograms depicting the prior ensemble (grey) in observation space, plotted along-side the posterior analysis (black open boxes). This plot also includes the histogram produced when the state space analysed parameter values are used to generate precipitation rates from the model (red open boxes). Results are shown for (a, d, g) MCMC, (b, e, h) GIG, and (c, f, i) EnKF for high (a, b, c), moderate (d, e, f), and low (g, h, I) precipitation rates. As in earlier figures, the true observation value is denoted by the dashed vertical line in each plot.

Figure 4. Plots of precipitation rate generated by the model when run with variable (a) slope intercept of the rain particle size distribution, and (b) cloud to rain autoconversion parameter. In each plot, one variable is allowed to change while the other is fixed. Each gray line represents the response function for a particular value of the other parameter. Colored lines correspond to the values that produce low (blue), moderate (black), and high (red) precipitation rates. The fixed parameter values are provided in the legend of each plot for reference, and are denoted by plus signs on each plot.

Figure 5. As in Figs. 1 and 3, these plots contain histograms depicting the prior ensemble (grey) in observation space, plotted along-side the posterior analysis (black open boxes), and histogram produced when the state space analysed parameter values are used to generate precipitation rates from the model (red open boxes). In this case, Results are shown for (a, b, c) MCMC, (d, e, f) GIG, and (g, h, i) the GIG with observation space analysis generated by the state space analysis following 10 outer loop iterations.

Figure 6. As in Fig. 2, but depicting a comparison between the GIG analysis formed via linear regression from the observation space analysis (a, b, c), the GIG analysis that results from 10 iterations of a weighted outer loop regression (d, e, f), and the MCMC state-space posterior analysis (g, h, i).

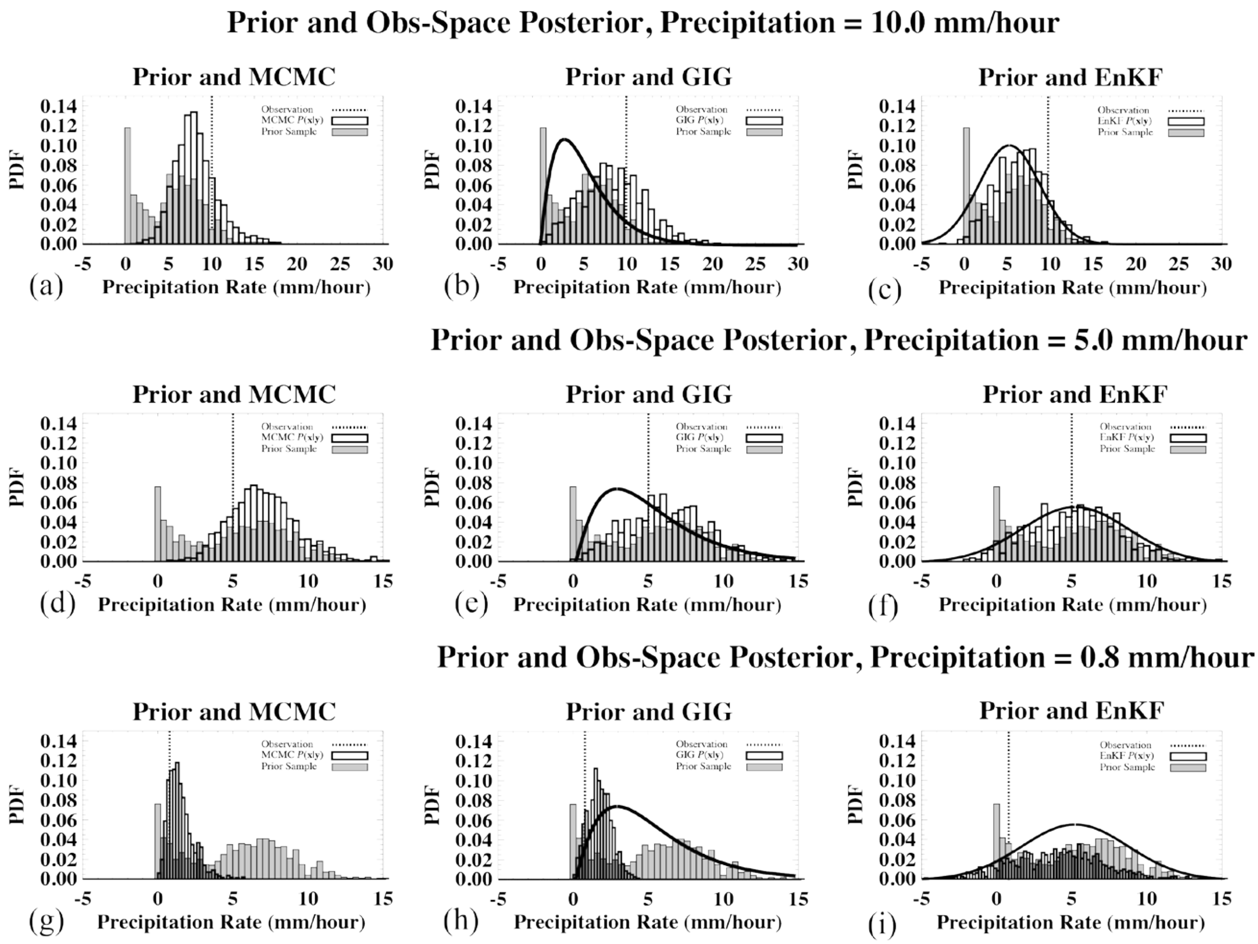


Figure 1. Histograms depicting the prior ensemble (grey) plotted along-side the posterior analysis (open boxes) of precipitation rates (mm hour^{-1}) from (a,d,g) MCMC, (b,e,h) GIG, and (c,f,i) EnKF for observed precipitation rates of 10 mm hour^{-1} (a, b, c), 5 mm hour^{-1} (d, e, f), and 0.8 mm hour^{-1} (g, h, i). The true precipitation rate value in each case is denoted by the dashed vertical line. The heavy solid lines depict the analytical gamma prior used in the GIG experiments (b, e, h), and the Gaussian prior used in the EnKF experiments (c, f, i).

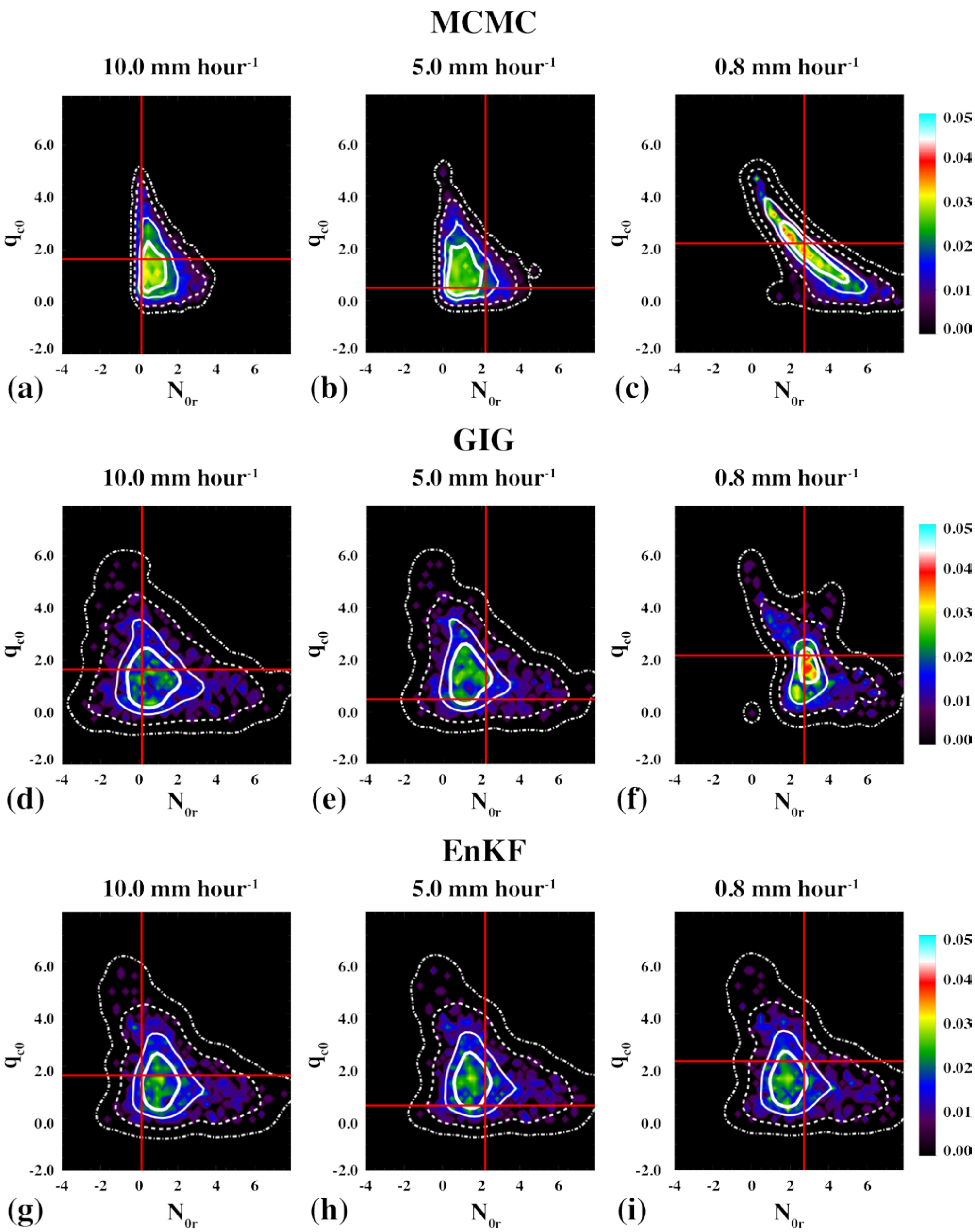


Figure 2. Two-dimensional joint posterior PDFs for the state space parameters for the (left) high, (centre) moderate, and (right) low precipitation cases, as computed by the (a – c) MCMC, (d – f) GIG, and (g – i) EnKF algorithms. Colour filled contours depict the normalized probability,

while the unfilled white contours surround regions containing 37% (heavy solid), 68% (solid), 95% (dashed), and 99% (dot-dashed) of the total probability mass. The red lines depict the true parameter values used in each experiment.

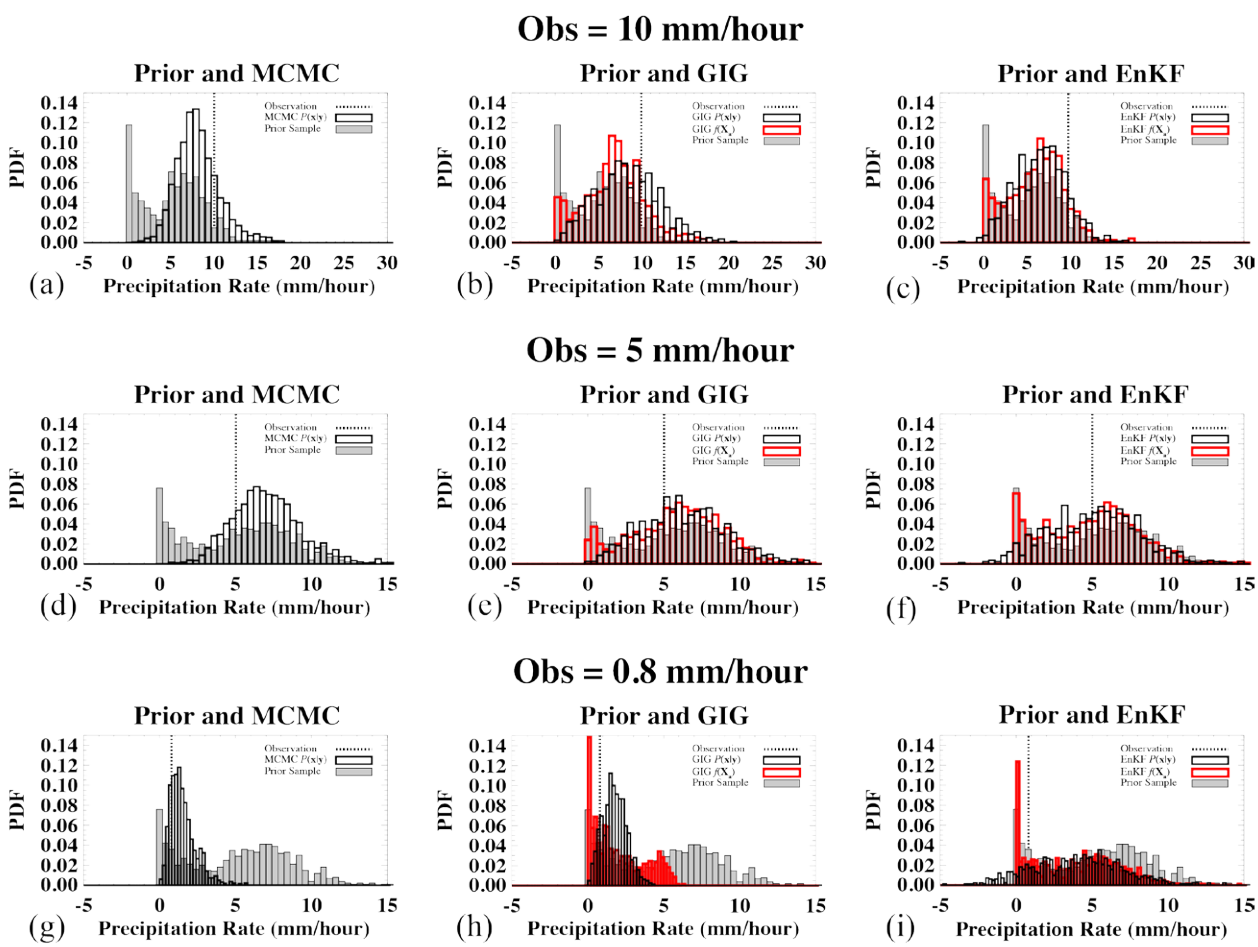


Figure 3. As in Fig. 1, these plots contain histograms depicting the prior ensemble (grey) in observation space, plotted along-side the posterior analysis (black open boxes). This plot also includes the histogram produced when the state space analysed parameter values are used to generate precipitation rates from the model (red open boxes). Results are shown for (a, d, g)

MCMC, (b, e, h) GIG, and (c, f, i) EnKF for high (a, b, c), moderate (d, e, f), and low (g, h, I) precipitation rates. As in earlier figures, the true observation value is denoted by the dashed vertical line in each plot.

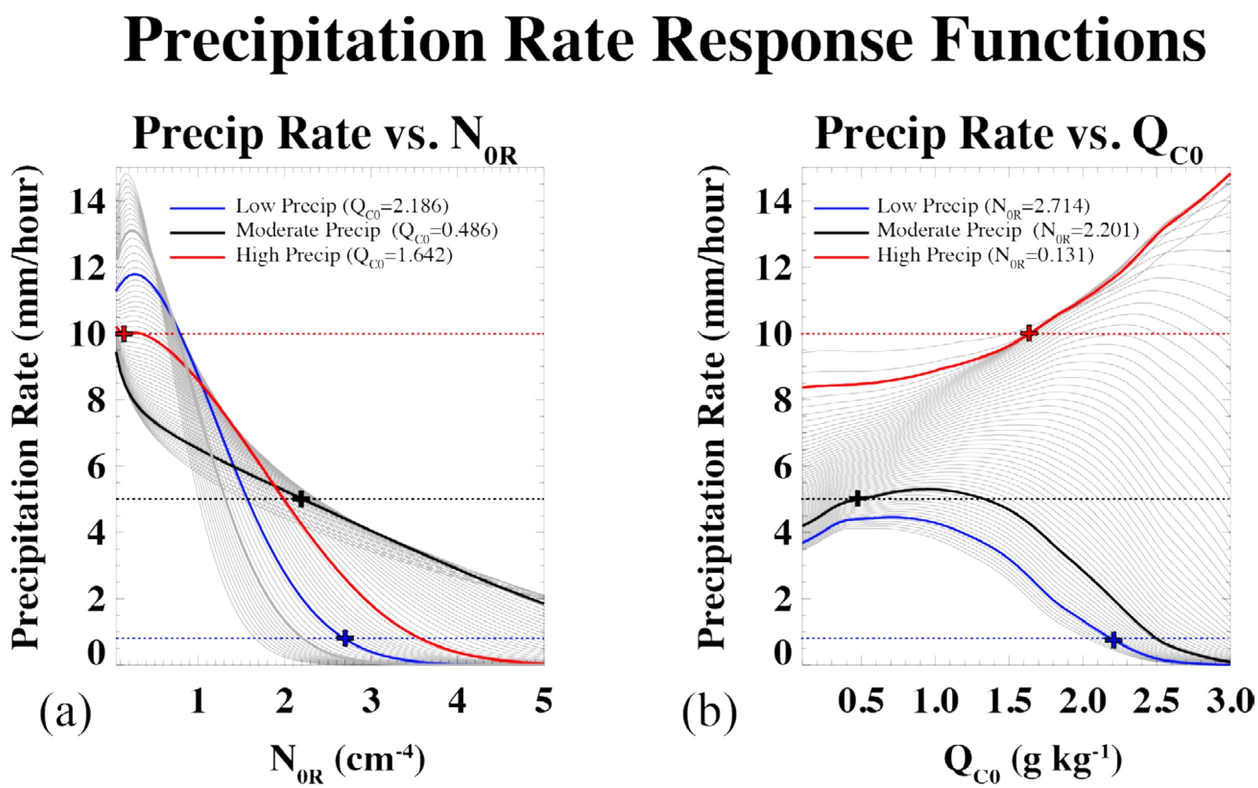


Figure 4. Plots of precipitation rate generated by the model when run with variable (a) slope intercept of the rain particle size distribution, and (b) cloud to rain autoconversion parameter. In each plot, one variable is allowed to change while the other is fixed. Each gray line represents the response function for a particular value of the other parameter. Colored lines correspond to the values that produce low (blue), moderate (black), and high (red) precipitation rates. The fixed parameter values are provided in the legend of each plot for reference, and are denoted by plus signs on each plot.

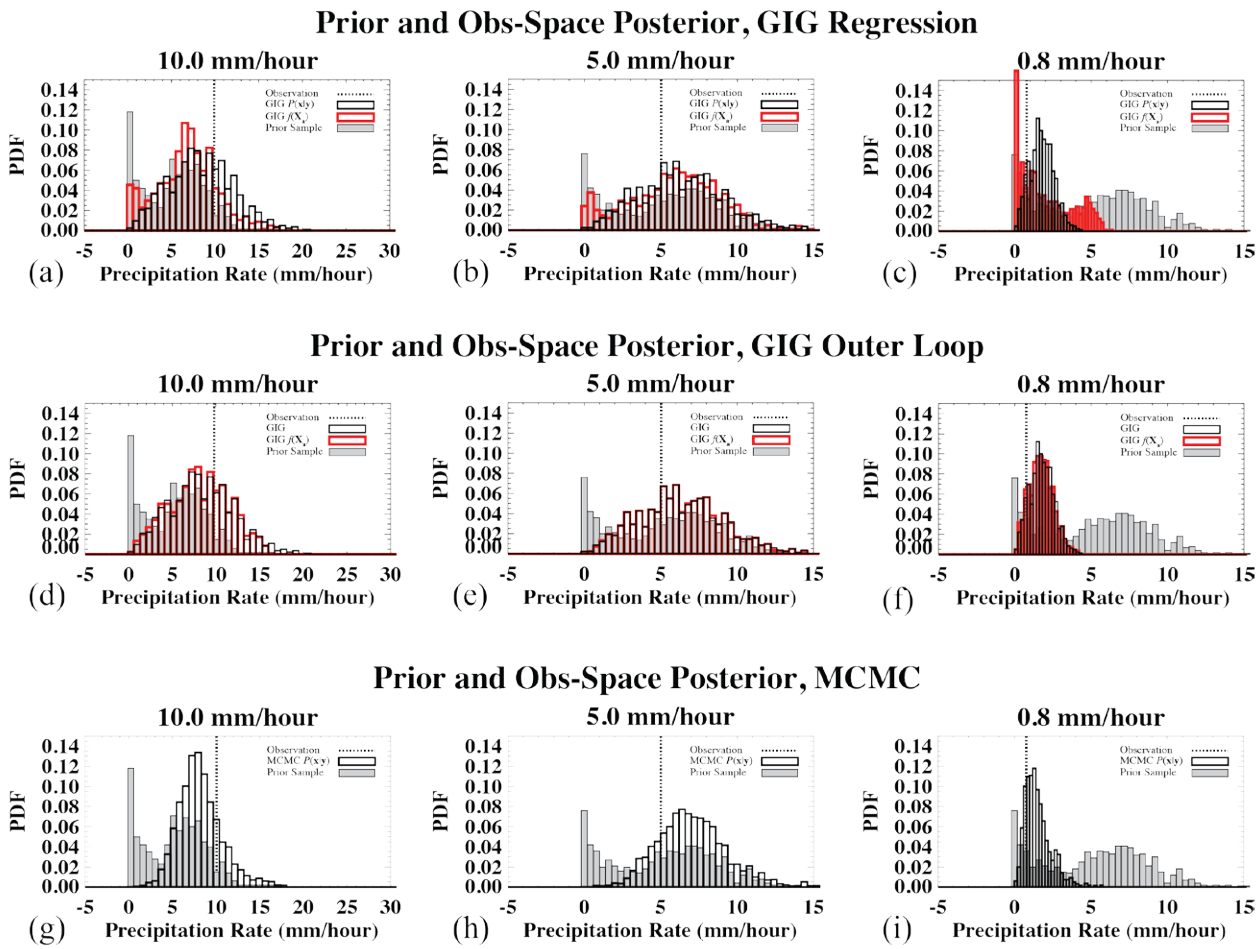


Figure 5. As in Figs. 1 and 3, these plots contain histograms depicting the prior ensemble (grey) in observation space, plotted along-side the posterior analysis (black open boxes), and histogram produced when the state space analysed parameter values are used to generate precipitation rates from the model (red open boxes). In this case, Results are shown for (a, b, c) GIG analysis formed via linear regression from the observation space analysis, (d, e, f) the GIG with

observation space analysis generated by the state space analysis following 10 outer loop iterations, and (g, h, i) MCMC.

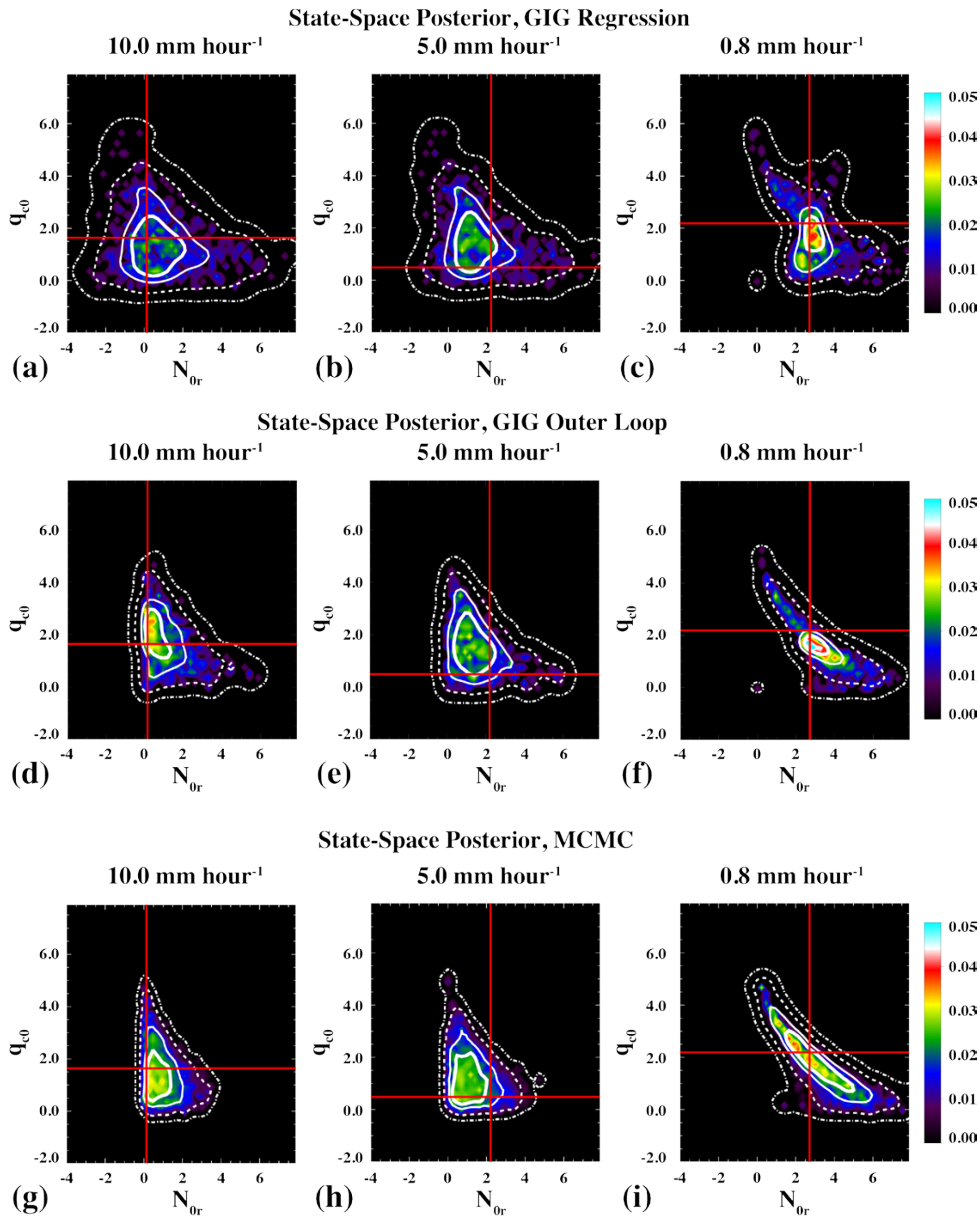
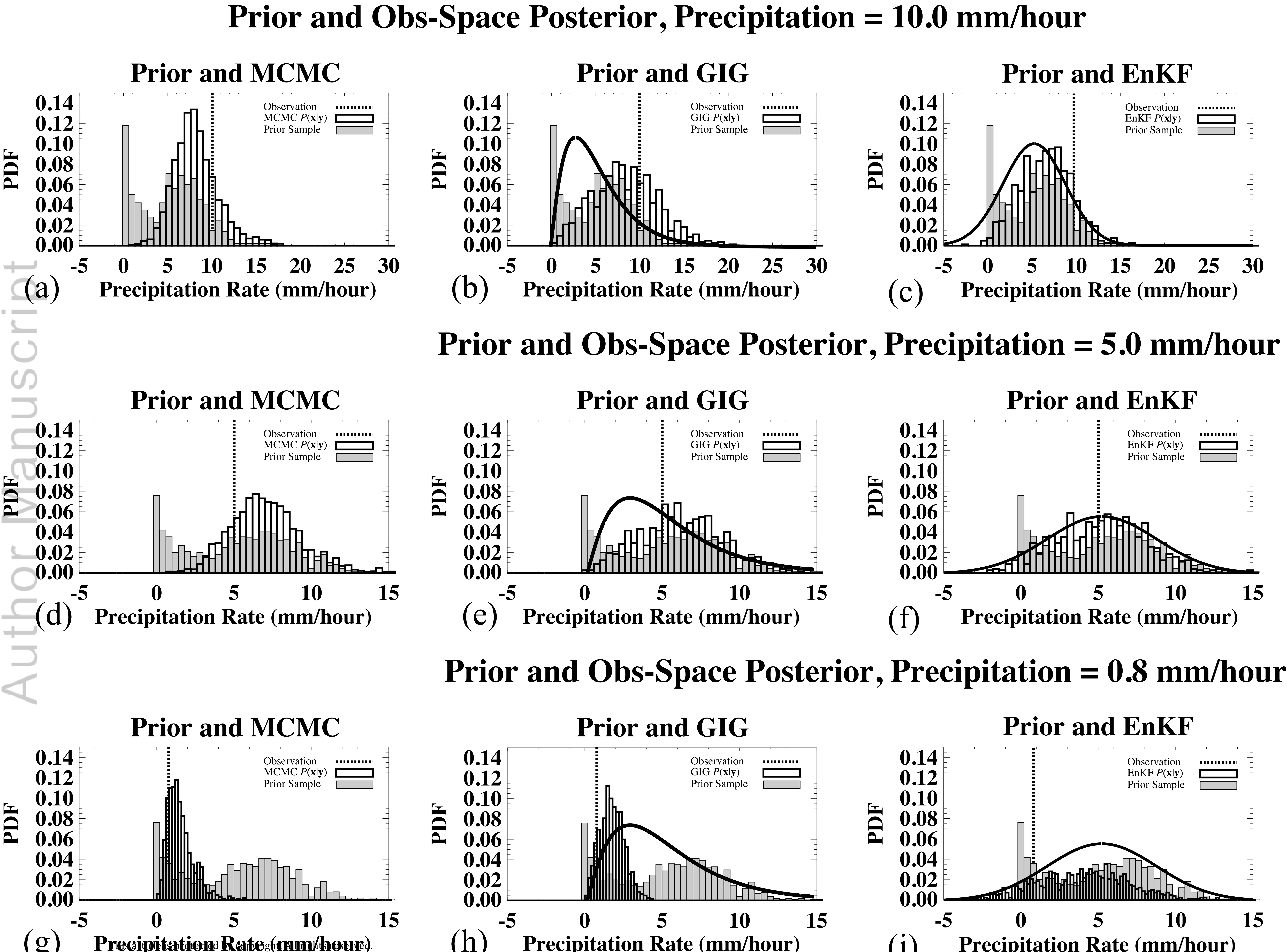
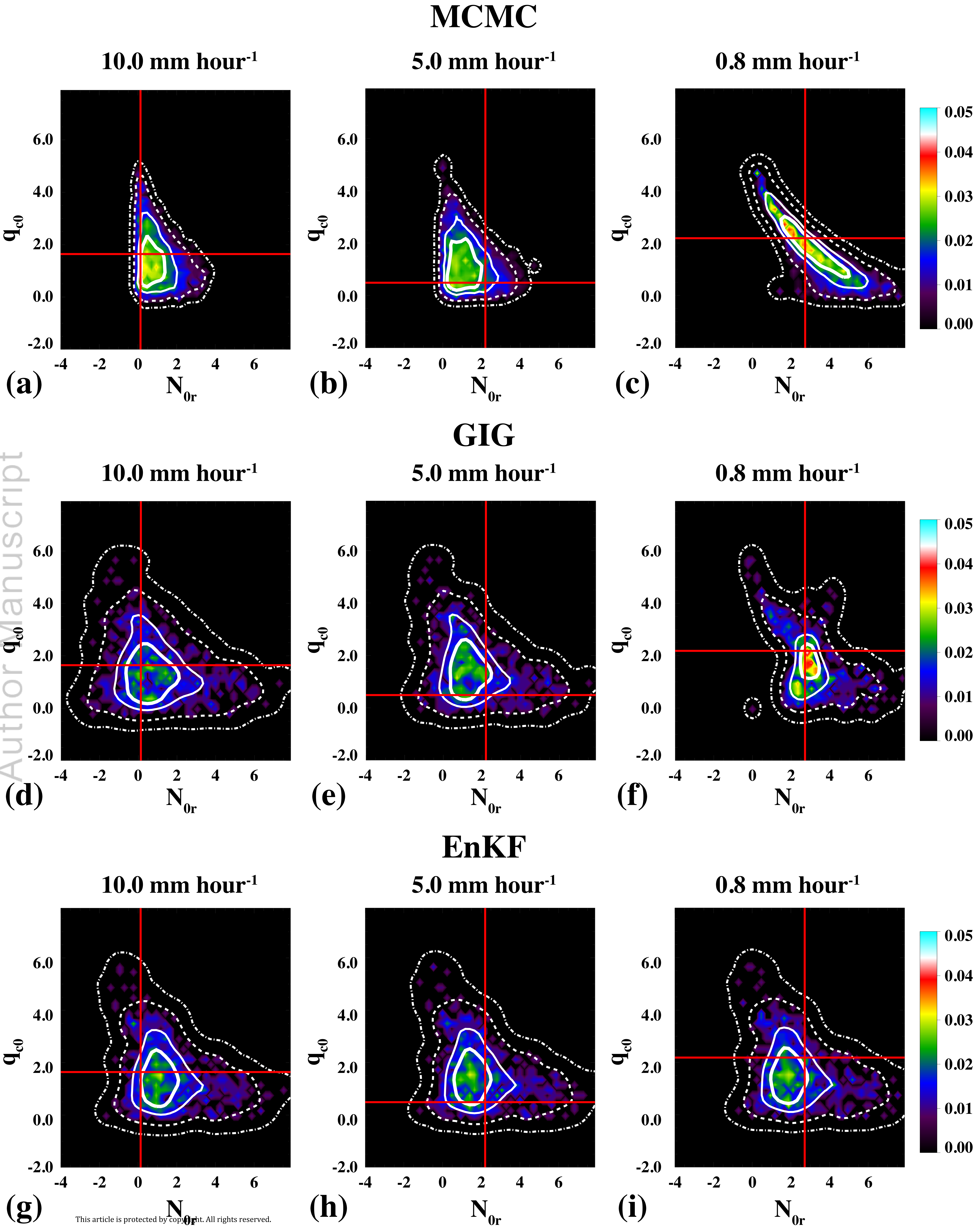
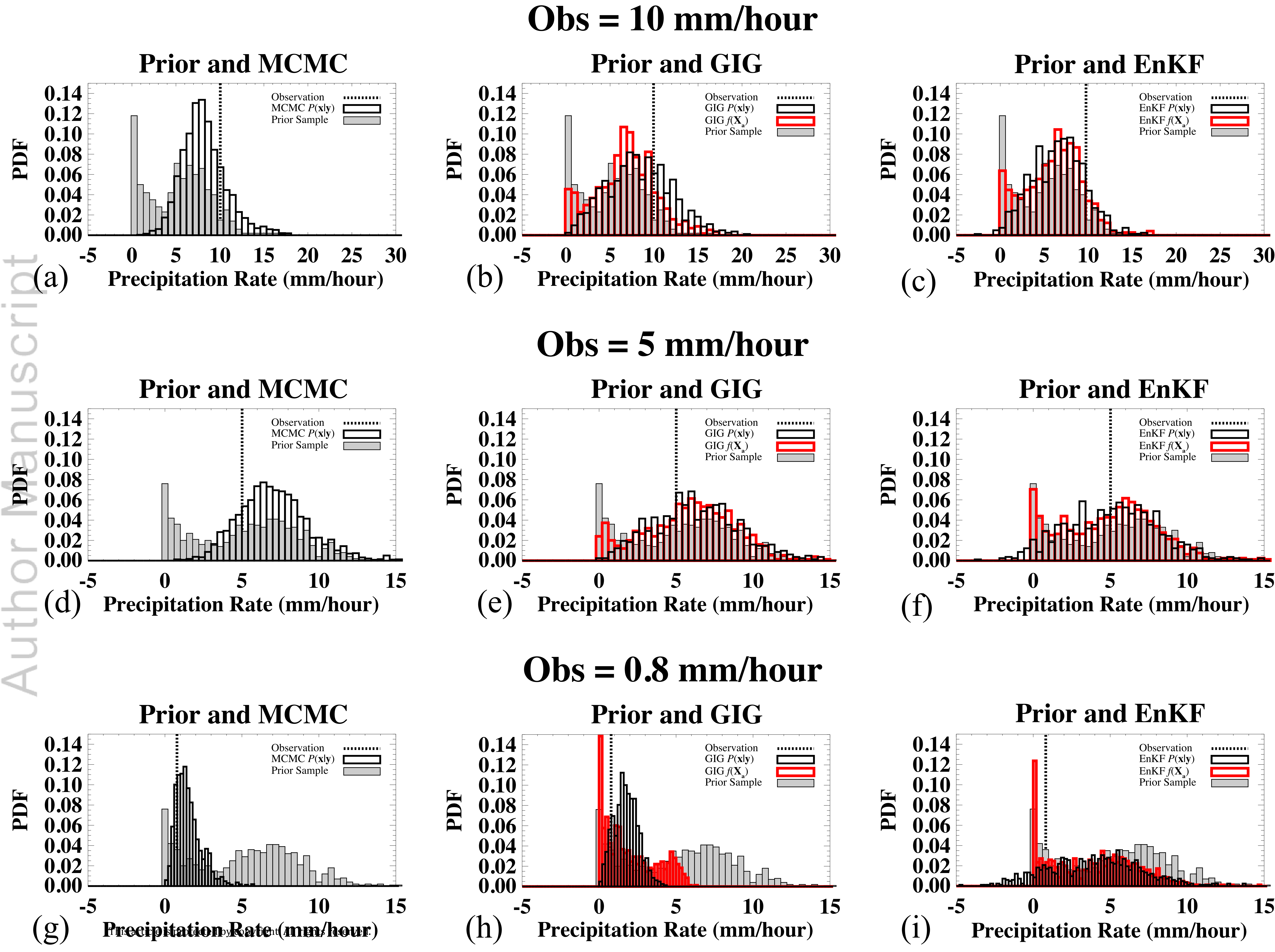

State-Space Posterior, GIG Regression
10.0 mm hour-1
5.0 mm hour-1
0.8 mm hour-1
(a)
(b)
(c)
State-Space Posterior, GIG Outer Loop
10.0 mm hour-1
5.0 mm hour-1
0.8 mm hour-1
(d)
(e)
(f)
State-Space Posterior, MCMC
10.0 mm hour-1
5.0 mm hour-1
0.8 mm hour-1
(g)
(h)
(i)
q_c0
N_0r
0.05
0.04
0.03
0.02
0.01
0.00

Figure 6. As in Fig. 2, but depicting a comparison between the GIG analysis formed via linear regression from the observation space analysis (a, b, c), the GIG analysis that results from 10 iterations of a weighted outer loop regression (d, e, f), and the MCMC state-space posterior analysis (g, h, i).

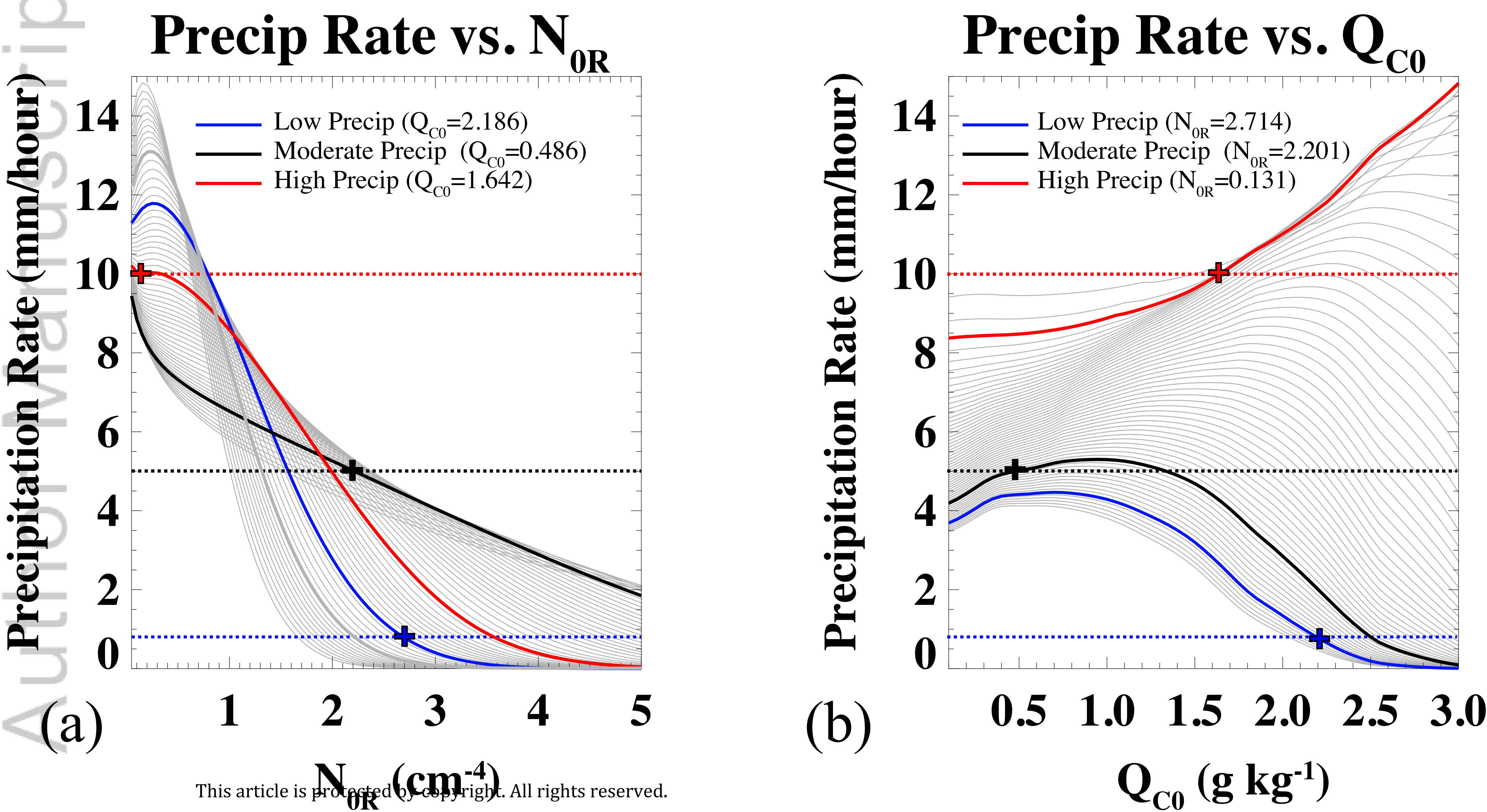

Prior and Obs-Space Posterior, Precipitation = 10.0 mm/hour
Prior and MCMC
Prior and GIG
Prior and EnKF
Observation
MCMC P(x|y)
GIG P(x|y)
EnKF P(x|y)
Prior Sample
PDF
Precipitation Rate (mm/hour)
(a)
(b)
(c)
Prior and Obs-Space Posterior, Precipitation = 5.0 mm/hour
(d)
(e)
(f)
Prior and Obs-Space Posterior, Precipitation = 0.8 mm/hour
(g)
(h)
(i)

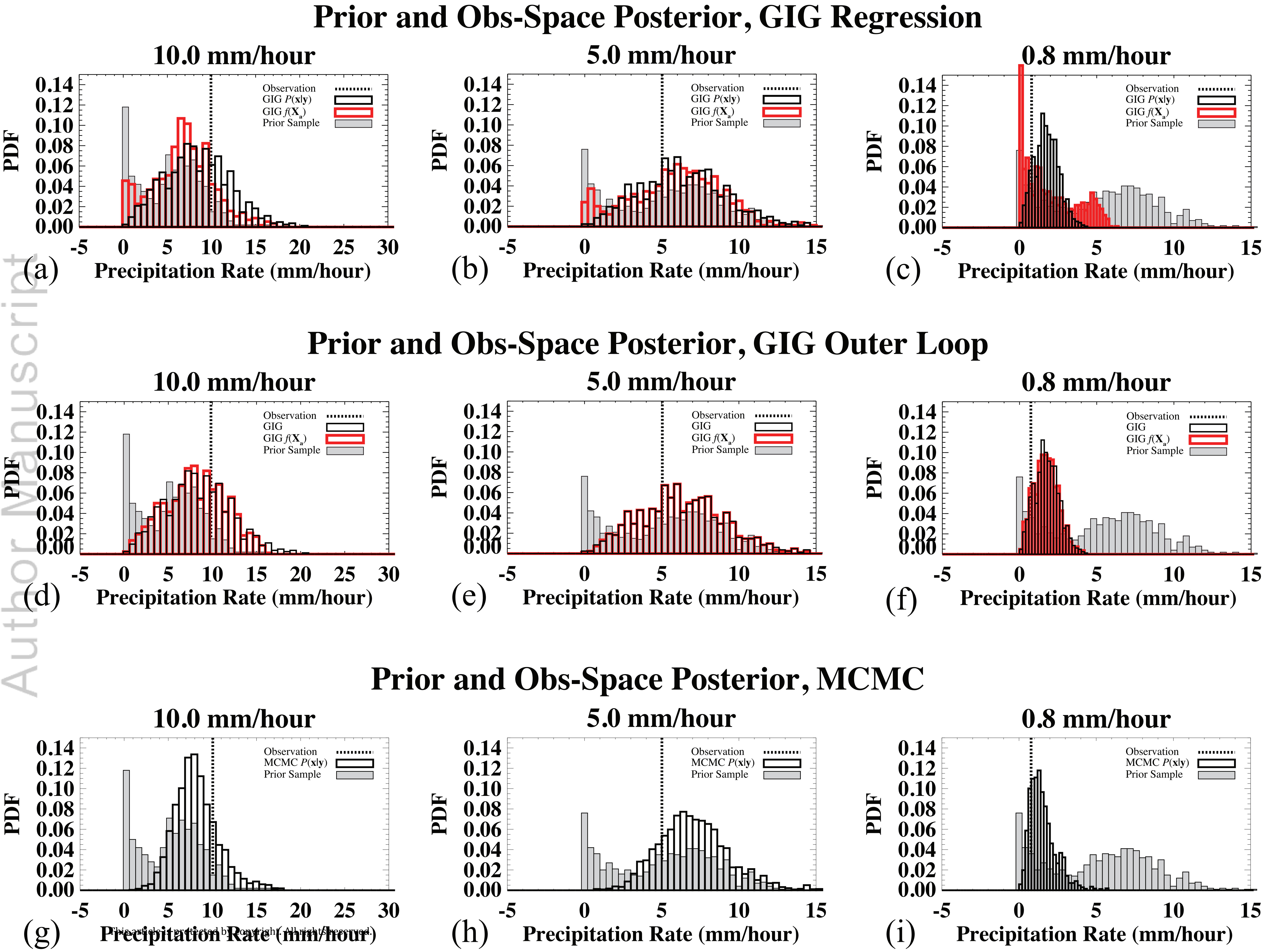

MCMC
GIG
EnKF
10.0 mm hour-1
5.0 mm hour-1
0.8 mm hour-1
qc0
N0r
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
0.05
0.04
0.03
0.02
0.01
0.00

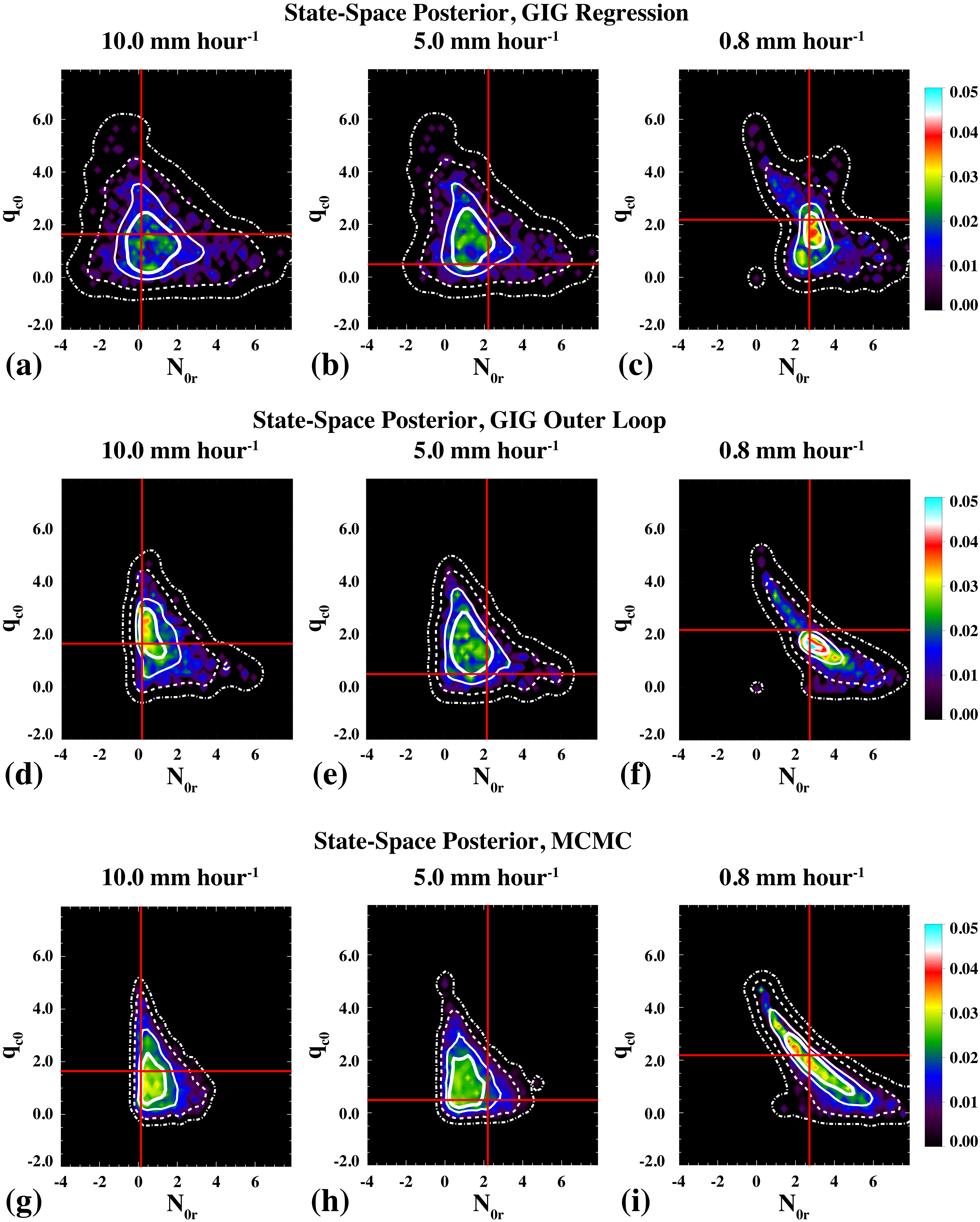


Obs = 10 mm/hour
Prior and MCMC
Prior and GIG
Prior and EnKF
Observation
MCMC P(x|y)
GIG P(x|y)
GIG f(X_a)
EnKF P(x|y)
EnKF f(X_a)
Prior Sample
PDF
Precipitation Rate (mm/hour)
(a)
(b)
(c)
Obs = 5 mm/hour
(d)
(e)
(f)
Obs = 0.8 mm/hour
(g)
(h)
(i)

Precipitation Rate Response Functions




Prior and Obs-Space Posterior, GIG Regression
10.0 mm/hour
5.0 mm/hour
0.8 mm/hour
Observation
GIG P(x|y)
GIG f(X_a)
Prior Sample
PDF
Precipitation Rate (mm/hour)
(a)
(b)
(c)
Prior and Obs-Space Posterior, GIG Outer Loop
10.0 mm/hour
5.0 mm/hour
0.8 mm/hour
Observation
GIG
GIG f(X_a)
Prior Sample
(d)
(e)
(f)
Prior and Obs-Space Posterior, MCMC
10.0 mm/hour
5.0 mm/hour
0.8 mm/hour
Observation
MCMC P(x|y)
Prior Sample
(g)
(h)
(i)



State-Space Posterior, GIG Regression
10.0 mm hour-1
5.0 mm hour-1
0.8 mm hour-1
q_c0
N_0r
(a)
(b)
(c)
State-Space Posterior, GIG Outer Loop
(d)
(e)
(f)
State-Space Posterior, MCMC
(g)
(h)
(i)
0.05
0.04
0.03
0.02
0.01
0.00

Nonlinear Data Assimilation for Clouds and Precipitation using a

Gamma-Inverse Gamma Ensemble Filter

Derek J. Posselt[1]* and Craig H. Bishop[2]

[1] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA*
[2] *Naval Research Laboratory, Monterey, CA, USA*

As shown in the image above, Earth's atmosphere produces a vast spectrum of various precipitating and non-precipitating clouds. Where clouds and rainfall occur, their amounts are always positive definite, and cloud remote sensing uncertainties are often fractional. A recently developed ensemble filter algorithm, the Gamma, Inverse-Gamma, and Gaussian Ensemble Kalman Filter (GIGG-EnKF), allows for fractional observation errors and positive definite quantities. As such, this algorithm represents a promising methodology for cloud and rainfall estimation using remote sensing data.

Corresponding author address: Derek J. Posselt, Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109. E-mail: Derek.Posselt@jpl.nasa.gov