

Observational constraints on warm cloud microphysical processes using machine learning and optimization techniques

Article

Published Version

Creative Commons: Attribution-Noncommercial 4.0

Open Access

Chiu, J. C. ORCID: <https://orcid.org/0000-0002-8951-6913>,
Yang, C. K. ORCID: <https://orcid.org/0000-0001-7685-4481>,
Van Leeuwen, P. J. ORCID: <https://orcid.org/0000-0003-2325-5340>,
Feingold, G. ORCID: <https://orcid.org/0000-0002-0774-2926>,
Wood, R. ORCID: <https://orcid.org/0000-0002-1401-3828>,
Blanchard, Y. ORCID: <https://orcid.org/0000-0003-0515-3041>,
Mei, F. ORCID: <https://orcid.org/0000-0003-4285-2749>
and Wang, J. (2021) Observational constraints on warm cloud microphysical processes using machine learning and optimization techniques. *Geophysical Research Letters*, 48 (2). e2020GL091236. ISSN 0094-8276 doi: <https://doi.org/10.1029/2020GL091236> Available at <https://centaur.reading.ac.uk/94901/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1029/2020GL091236>

Publisher: American Geophysical Union

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Geophysical Research Letters

RESEARCH LETTER

10.1029/2020GL091236

Key Points:

- Machine-learning trained by in situ data constrains autoconversion and accretion rates with uncertainty of 15% and 5%, respectively
- There is a surprising relation between autoconversion rate and drizzle number concentration that significantly improves parameterizations
- The exponent of autoconversion rate dependence on cloud number concentration is 0.75, lower than that in existing parameterizations

Supporting Information:

- Supporting Information S1

Correspondence to:

J. C. Chiu,
Christine.Chiu@colostate.edu

Citation:

Chiu, J. C., Yang, C. K., van Leeuwen, P. J., Feingold, G., Wood, R., Blanchard, Y., et al. (2021). Observational constraints on warm cloud microphysical processes using machine learning and optimization techniques. *Geophysical Research Letters*, 48, e2020GL091236. <https://doi.org/10.1029/2020GL091236>

Received 14 JUL 2020

Accepted 7 DEC 2020

© 2020. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by-nc/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Observational Constraints on Warm Cloud Microphysical Processes Using Machine Learning and Optimization Techniques

J. Christine Chiu¹ , C. Kevin Yang¹ , Peter Jan van Leeuwen^{1,2} ,
Graham Feingold³ , Robert Wood⁴ , Yann Blanchard⁵ , Fan Mei⁶ , and Jian Wang⁷ 

¹Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA, ²Department of Meteorology, University of Reading, Reading, UK, ³NOAA Earth System Research Laboratory, Boulder, CO, USA, ⁴Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA, ⁵Department of Earth and Atmospheric Sciences, ESCER Centre, University of Quebec at Montreal, Montreal, QC, Canada, ⁶Pacific Northwest National Laboratory, Richland, WA, USA, ⁷Center for Aerosol Science and Engineering, Department of Energy, Environmental and Chemical Engineering, Washington University in Saint Louis, Saint Louis, MO, USA

Abstract We introduce new parameterizations for autoconversion and accretion rates that greatly improve representation of the growth processes of warm rain. The new parameterizations capitalize on machine-learning and optimization techniques and are constrained by in situ cloud probe measurements from the recent Atmospheric Radiation Measurement Program field campaign at Azores. The uncertainty in the new estimates of autoconversion and accretion rates is about 15% and 5%, respectively, outperforming existing parameterizations. Our results confirm that cloud and drizzle water content are the most important factors for determining accretion rates. However, for autoconversion, in addition to cloud water content and droplet number concentration, we discovered a key role of drizzle number concentration that is missing in current parameterizations. The robust relation between autoconversion rate and drizzle number concentration is surprising but real, and furthermore supported by theory. Thus, drizzle number concentration should be considered in parameterizations for improved representation of the autoconversion process.

Plain Language Summary Drizzle has been a key element of research, because its formation modulates cloud properties and evolution, and affects the water cycle of the Earth. Since drizzle formation involves cloud droplets of all sizes, it requires extensive computational time. Hence, we often use simplified methods in weather and climate prediction models to obtain a bulk estimate of how fast and how many cloud droplets collide with each other or collide with bigger drops to form drizzle. However, many models continue to have inadequate representation of drizzle formation, calling for the need to improve these simplified methods. We introduce new methods to estimate the rate of those microphysical processes, capitalizing on aircraft measurements and recent advances in machine-learning techniques. Our techniques outperform the current methods significantly. Importantly, our analyses reveal that the rate of drizzle formation via collisions between cloud drops is related to drizzle drop number concentration itself, which is missing in the existing methods. This relation occurs because drizzle drop number concentration provides information on the stage of evolution of cloud size distribution during drizzle formation. Although this is not a causal relationship, it is important to incorporate this relation into models for better prediction of drizzle formation.

1. Introduction

Warm rain formation plays a crucial role in determining the properties and life cycle of marine boundary layer clouds, and has significant impacts on radiative and hydrological budgets. Yet, many global weather and climate models continue to produce rain too frequently over oceans (e.g., Stephens et al., 2010), too light (Ahlgren & Forbes, 2014; Jing et al., 2017), or too heavy (Abel & Boutle, 2012; Bodas-Salcedo et al., 2008). The intermodel spread in precipitation rate in the southeast Pacific, one of the major marine boundary layer cloud decks, can be an order of magnitude (Wyant et al., 2015). The model discrepancy and spread in precipitation are linked to diverse issues, such as rain drop size distributions, and the representation of

boundary layer, autoconversion, and accretion processes. The effects of autoconversion on precipitation can change surface temperature prediction significantly (Golaz et al., 2013).

Many warm rain parameterizations for autoconversion and accretion processes have been developed in the past (e.g., Beheng, 1994; Berry, 1968; Kessler, 1969; Khairoutdinov & Kogan, 2000; Liu & Daum, 2004; Seifert & Beheng, 2001; and many others). These parameterizations have been reviewed critically (Lee & Baik, 2017; Liu & Daum, 2004; Wood, 2005), and confronted with in situ observations (Hsieh et al., 2009; Wood, 2005). By applying in situ size-resolved cloud measurements to the continuous collection equation, the two aforementioned observational studies showed that parameterized accretion rates generally agree with in situ data, but parameterized autoconversion rates can be significantly different from observational estimates. While these results are encouraging and informative, there has been little follow-up observational work. It remains unclear how to maximize the use of observations for improving understanding and model representations of these microphysical processes, and how to extend constraints from in situ to remote sensing platforms that can provide continuous observations in various cloud regimes on a global scale.

The objectives of this study are manifold. Instead of evaluating existing parameterizations, here we use in situ observations to build machine-learning (ML) models to “predict” autoconversion and accretion rates. Since translating ML results to physical formulations is not trivial and remains an active research area, we also perform nonlinear optimizations to fill the gap and to quantify the relationships of autoconversion and accretion with cloud/drizzle properties. These results are compared and contrasted with widely used parameterizations, and the implications are discussed.

2. In Situ Cloud Measurements

In situ cloud measurements were taken from the Aerosol and Cloud Experiments in the Eastern North Atlantic (ACE-ENA) field campaign, deployed by the Atmospheric Radiation Measurement (ARM) user facility. The aircraft flew near the ARM site on Graciosa Island during two intensive operational periods in June–July 2017 (IOP1) and January–February 2018 (IOP2). The cloud types sampled in ACE-ENA are mainly marine stratocumulus, with some scattered or precipitating cumulus. Measurements from three cloud probes were merged to form combined drop size distributions (DSD). Cloud droplets are defined as those with radii smaller than 25 μm , and drizzle drops are defined as those with radii larger than 25 μm and up to 400 μm . The choice of the cloud/drizzle separation threshold is appropriate for marine stratocumulus (Khairoutdinov & Kogan, 2000; Kogan, 2013). We also define a cloudy sample when cloud water content $q_c \geq 0.01 \text{ g m}^{-3}$. Based on this definition, a total of $\sim 93,000$ in situ cloudy DSDs comprise 11% drizzle-free DSDs and 89% drizzling DSDs after data screening (see Table S1). The smallest drizzle water content (q_r) observed in the drizzling DSDs is on the order of 10^{-5} g m^{-3} , and thus we set it as the threshold for drizzle delineation, denoted as $q_{r,\text{crit}}$. Property distributions from individual days for cloudy samples are shown in Figures 1a–1d.

3. Parameterization Derived From Machine Learning Techniques

The in situ DSDs from ACE-ENA are used to develop 2 ML models. The first ML model, dubbed “*initiation model*,” uses two inputs (q_c , N_c) for predicting autoconversion rate (P_{au}) in drizzle-absent conditions. The second, dubbed “*standard model*,” uses four inputs (q_c , N_c , q_r , N_r) for predicting P_{au} and accretion rate (P_{ac}) in drizzling conditions. As discussed further below, we use the initiation model to generate nonzero q_r and N_r values. Once q_r and N_r exist, the standard model is superior and used to better predict P_{au} and P_{ac} .

Both models use an Artificial Neural Network. It is a deep feed forward neural network (Schmidhuber, 2015) comprising eight hidden, fully connected layers with 1,024 nodes in each layer. All input and output variables are transformed to their logarithmic forms. Since these input variables have rather different magnitudes, we normalized them using their mean and standard deviation. We used LeakyReLU (Mass et al., 2013) as our activation function. Additionally, the training was performed by the Adam optimizer (Kingma & Ba, 2015), based on a loss function defined as the mean squared error between the true value and the prediction.

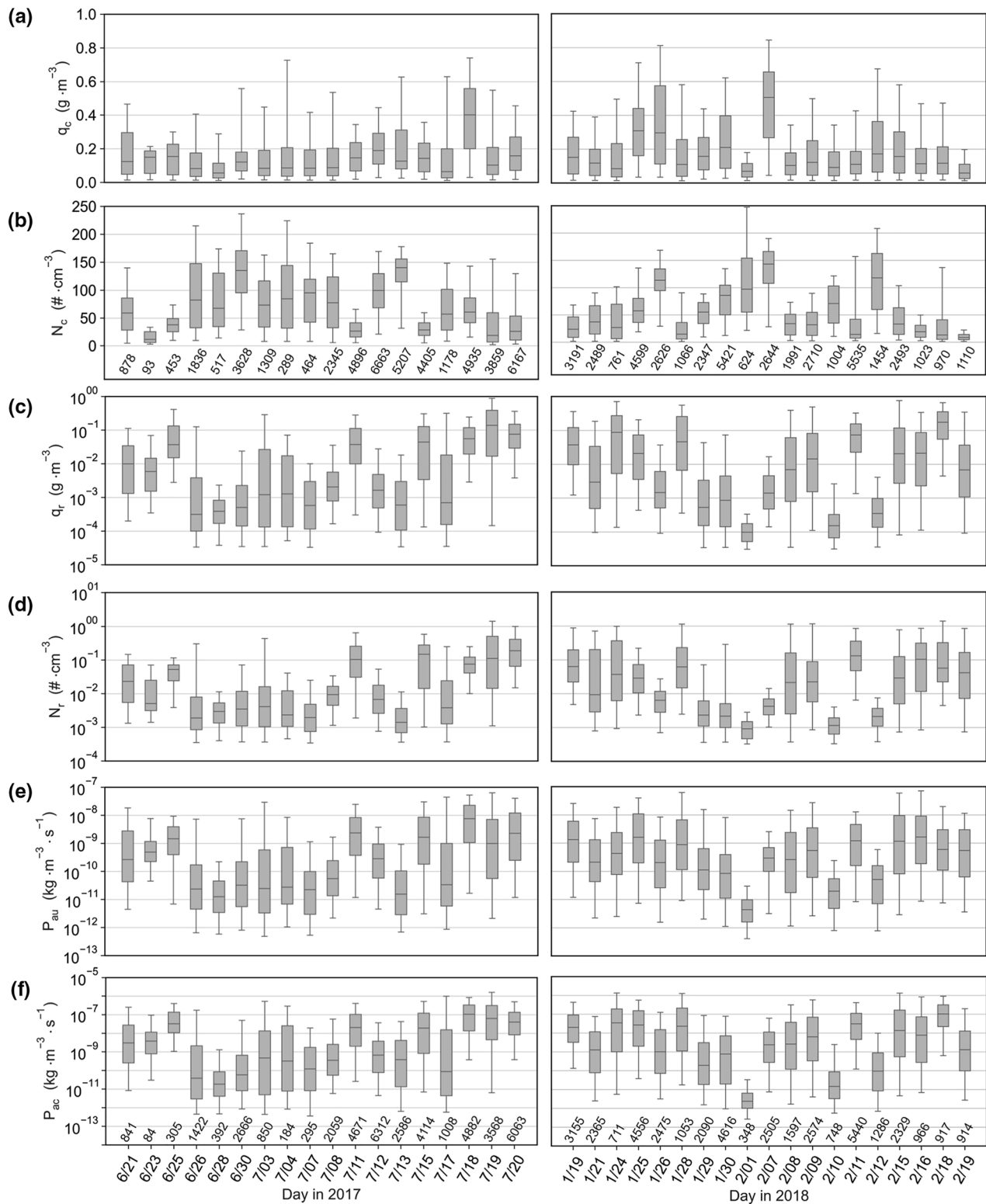


Figure 1. Box plots of in situ (a) cloud water content (q_c), (b) cloud droplet number concentration (N_c), (c) drizzle water content (q_r), and (d) drizzle drop number concentration (N_r) observed in IOP1 (left panels) and IOP2 (right panels) from the ARM campaign at the Azores. The bottom and top of each box represent the 25% and 75% quartiles, and the line inside the box represents the median. The whiskers mark the 5th and 95th percentiles. Panels (e) and (f) are calculated autoconversion rate (P_{au}) and accretion rate (P_{ac}) from observed drop size distributions using the stochastic collection equation formulated as a two-moment bin model (see Section 3). Based on 1-s measurements, the sample size for each day is listed in (b) for all conditions, and (f) for drizzling conditions. Flights on June 29 and July 6, 2017 were excluded due to data availability. ARM, atmospheric radiation measurement; IOP, intensive operational period.

The training data sets for the two models are different but originate from the same pool of data points. The pool was generated by using the in situ DSDs as the initial conditions and propagating the DSDs forward in time with the stochastic collection equation (SCE). We used the two-moment bin model of Tzivion et al. (1987) to compute P_{au} and P_{ac} directly from the explicit drop-drop interaction terms at 1-s time steps for 10 min. The bin model uses the Hall (1980) kernel. The 10-min time period is based on the typical in-cloud residence time (Feingold et al., 1996). Since our focus is on clouds, we exclude noncloudy data points from the pool.

For the initiation ML model, the training data set is based on data points generated from the initially drizzle-free DSDs in the pool. To ensure that DSDs used for the initiation model are absolutely drizzle free, we exclude DSDs that contain cloud droplets in the instrument size bin (17.5–22.5 μm radius) proximate to the cloud/drizzle boundary (i.e., 25 μm radius), based on the uncertainty of 1.5–5 μm in in situ size measurements (Glienke & Mei, 2019, 2020). We do retain all DSDs that do not have droplets in that bin initially but produce nonzero q_r in 5 s. These are practical choices that are as inclusive as possible of DSDs and also facilitate the ML. As shown in Figures 2a and 2b, the initiation model has the 25th and 75th errors ranging between –60% and 80%.

Note that these initially drizzle-free DSDs generate q_r ranging between 10^{-18} and 10^{-9} g m^{-3} in 5 s. The lower bound, 10^{-18} g m^{-3} , is equivalent to one single drizzle drop with a radius of 25 μm in a 10 km \times 10 km \times 500 m cloudy volume. The upper bound, 10^{-9} g m^{-3} , is at least four orders of magnitude smaller than any in situ measured q_r . Therefore, although the initiation model generates nonzero q_r , these numbers are very small and should still be considered as nondrizzling in any practical sense. We emphasize that the initiation ML model is mainly used as a gateway to the standard ML model that requires nonzero q_r as input.

For the standard model, we sample all the cloudy points every 5 s from the pool to form the training data set, as long as their $q_r \geq 10^{-18}$ g m^{-3} . This threshold is based on the q_r magnitude that can be initiated by the aforementioned initiation model, ensuring that the q_r range between two models overlap as much as possible. This leads to a total of ~ 10.7 M data points. From the rest of the pool that are not sampled for training, we randomly selected 2.5 M points for testing. The ratio between the training and the testing is about 4, similar to the standard practice in ML. Additionally, both the training and testing data sets contain $\sim 25\%$ data that have a ratio of P_{au} to $P_{ac} \geq 1$ (i.e., in the early stages of drizzle formation), and 75% data that have a ratio < 1 .

Figures 2c, 2d, 3a and 3b show the performance of the standard ML model. For both P_{au} and P_{ac} , the majority of the data points fall on the 1:1 line in the scatter plots, confirming the appropriateness of the neural network. The uncertainty is 15% for P_{au} , and 5% for P_{ac} . The good performance on the testing data set indicates that the ML model does not suffer from overfitting. Predicting P_{au} and P_{ac} for these 2.5 M points takes about 100 s using a single Intel Xeon E5-2697V4 processor. Note that training P_{au} and P_{ac} separately using the same input yielded similar results (see Table S2).

Results from the standard ML model are further compared with those from the parameterization proposed in Khairoutdinov and Kogan (2000), dubbed the KK parameterization. KK describes P_{au} and P_{ac} as:

$$P_{au} = 7.42 \times 10^{13} q_c^{2.47} N_c^{-1.79} \rho^{-1.47}, \quad (1)$$

$$P_{ac} = 67 q_c^{1.15} q_r^{1.15} \rho^{-1.3}, \quad (2)$$

where ρ is the air density and all variables are in SI units. As shown in Figures 2e, 2f, 3c, and 3d, the KK parameterization predicts P_{ac} reasonably well, with errors smaller than 40%, but significantly overestimates low P_{au} and slightly underestimates high P_{au} . The performance of KK is consistent with the findings in Wood (2005).

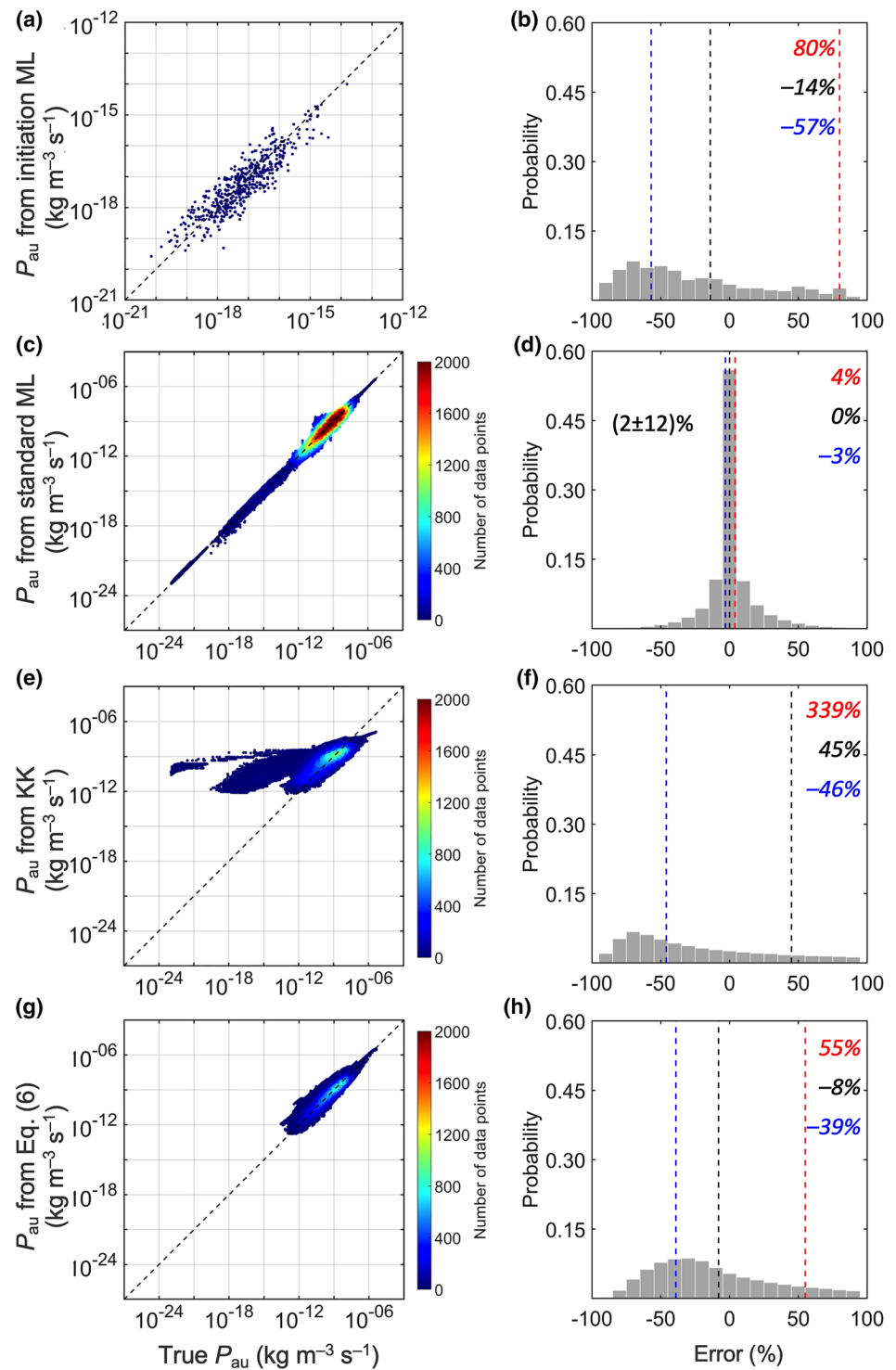


Figure 2. Plots of the predicted versus the true autoconversion rates from the testing data set, using (a) the initiation and (c) standard machine-learning model, (e) the KK parameterization, and (g) Equation 6. (a) is a scatter plot, while the number of data points in all others are indicated by color. The corresponding histograms of errors (%) from these individual methods are shown in (b, d, f, and h), respectively. The blue, black, and red dashed lines, respectively, represent the 25th, 50th, and 75th percentiles of the data. The corresponding errors for these lines are denoted in each subplot in their own color. For the KK parameterization, the 75th percentile is not plotted because it is out of the axis range. For the standard ML model, the mean error (%) and the mean absolute deviation (%; using the mean as the center point) are denoted in (d).

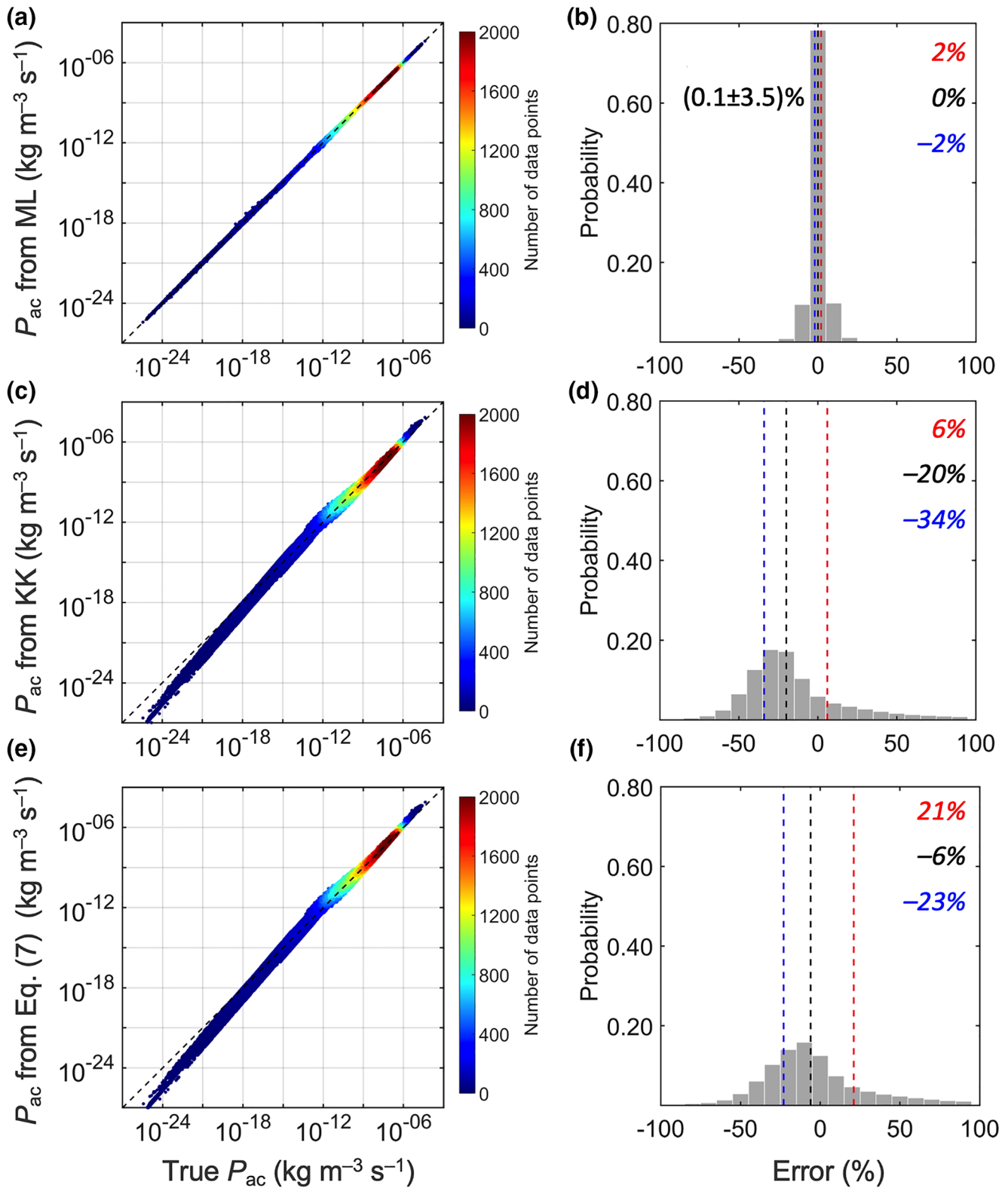


Figure 3. Same as Figure 2, but for accretion rate. Panels (e and f) are based on Equation 7, using 2.5 M data points. Note that estimates from KK and Equation 7 are so similar that the density scatter plots do not show any obvious differences.

4. Parameterizations Based on a Simple Form

While the ML model can be used to predict P_{au} and P_{ac} successfully, extracting the bulk dependencies that shed light on the underlying physics, is not straightforward. To characterize the physical relationships, we assume that process rate, P , representing either P_{au} or P_{ac} , can be parameterized as:

$$P = kq_c^a N_c^b q_r^c N_r^d. \quad (3)$$

This form is similar to the KK parameterization (Equations 1 and 2), and the parameterization in Liu and Daum (2004), dubbed the LD parameterization, in which

$$P_{au} \propto q_c^3 N_c^{-1}. \quad (4)$$

Unlike the KK and LD parameterizations, we include all observables on the right-hand side of Equation 3 to cover any possible dependencies, but will examine if all are necessary.

To find the optimal values for parameter $a-d$ and k in Equation 3, we take the logarithm of both sides, leading to

$$\log P = \log k + a \log q_c + b \log N_c + c \log q_r + d \log N_r. \quad (5)$$

This linear equation allows us to set up a least squares problem, solved with a limited memory quasi-Newton method (L-BFGS; Byrd et al., 1995) with wide bounds on the parameter values. The bounds are wide enough to ensure the best values do not end up close to the bounds. Uncertainty estimates are obtained from the inverse of the Hessian. The mean value of k has been corrected for the fact that k is implicitly assumed to be lognormally distributed because of the log transformation between Equations 3 and 5.

We use the testing data set with 2.5 M points to derive the parameters in Equation 3. The in situ q_c and q_r have an uncertainty of 30%, while N_c and N_r have an uncertainty of 50% and 20%, respectively (Glienke & Mei, 2019, 2020; Mei et al., 2020). These uncertainties are accounted for in Equation 3, leading to an additive error of 0.8 in Equation 5. Since it is possible that not all the variables constrain the solution, we systematically reduce the number of variables and adjust the additive error accordingly in the minimization.

Table 1 summarizes the parameter estimates and error statistics for predicting P_{au} and P_{ac} , based on the testing data set but with $q_r \geq q_{r,crit}$. The reason for this restriction on q_r is that the power laws are unable to fit a range of q_r spanning 18 orders of magnitudes. As a result, the sample size was reduced from 2.5 to 2.3 M. If we must predict P_{au} from q_c and N_c alone, as do existing parameterizations, the corresponding exponents are about 2.90 and -1.69 , respectively. Our q_c exponent is closer to LD's value ($a = 3$) than KK's ($a = 2.47$), and our N_c exponent is closer to KK's value ($b = -1.78$) than LD's ($b = -1$). As shown in Table 1, adding q_r into parameterizations produces a better correlation in P_{au} predictions. However, in general, the parameterizations involving N_r tend to perform best and have smaller errors. Once N_r is considered in the physical relationship, the exponents for both q_c and N_c , that is, the sensitivity of the P_{au} to these two variables, is reduced. Interestingly, we also find that the exponents of N_r and N_c are nearly reciprocal. The key role of N_r in autoconversion rate is counter-intuitive and will be discussed in the next section.

For accretion, the key roles of q_c and q_r are consistent with existing parameterizations. Taking the KK parameterization as an example, the exponent for q_c and q_r is 1.15, which is very close to our exponents, 1.148 for q_c and 1.159 for q_r .

Based on the performance in errors in Table 1 and the minimum number of variables required to reach a high correlation, we select the following power laws for predicting P_{au} and P_{ac} :

$$P_{au} = 16.8 q_c^{2.015} N_c^{-0.746} N_r^{0.640}, \quad (6)$$

$$P_{ac} = 69.5 q_c^{1.148} q_r^{1.159}. \quad (7)$$

Table 1

Optimal Parameters for Representing Autoconversion and Accretion Rates in the Form of $P = kq_c^a N_c^b q_r^c N_r^d$, where P is in $\text{kg m}^{-3} \text{s}^{-1}$, q_c in kg m^{-3} , N_c in m^{-3} , q_r in kg m^{-3} , and N_r in m^{-3} , Using 2.3 M Data Points

Corr.	Error (%)		k	a	b	c	d
	25th/50th/75th						
Autoconversion							
Nonturbulent conditions							
0.96	−39/−8/55		2.44 ± 0.05	2.0681 ± 0.0007	$−0.7760 \pm 0.0007$	$−0.1285 \pm 0.0004$	0.7844 ± 0.0005
0.96	−39/−9/55		5.9 ± 0.1	1.9839 ± 0.0008	$−0.7496 \pm 0.0007$	$−0.0642 \pm 0.0005^a$	0.7043 ± 0.0006
0.92	−52/−9/90		$(164 \pm 2) \text{ E7}$	2.2742 ± 0.0007	$−1.0930 \pm 0.0006$	0.3177 ± 0.0002	—
0.93	−49/−9/77		$(71.1 \pm 1.4) \text{ E7}$	2.5247 ± 0.0009	$−1.0548 \pm 0.0008$	0.4185 ± 0.0004^a	—
0.96	−39/−8/55		16.8 ± 0.3	2.0150 ± 0.0006	$−0.7461 \pm 0.0006$	—	0.6403 ± 0.0003
0.88	−59/−8/129		$(375 \pm 4) \text{ E12}$	2.8957 ± 0.0005	$−1.6945 \pm 0.0004$	—	—
Turbulent conditions with a dissipation rate of $400 \text{ cm}^2 \text{ s}^{-3}$							
0.96	−39/−8/53		11.1 ± 0.2	1.9777 ± 0.0007	$−0.7366 \pm 0.0006$	—	0.6511 ± 0.0003
Nonturbulent conditions, but only using points that $P_{au} \geq P_{ac}$							
0.96	−31/1/49		$(201 \pm 7) \text{ E6}$	1.7699 ± 0.0018	$−0.7975 \pm 0.0014$	0.8043 ± 0.0009	—
0.96	−31/1/48		1.22 ± 0.05	1.7656 ± 0.0017	$−0.7929 \pm 0.0013$	—	0.8432 ± 0.0008
0.84	−56/3/129		$(73 \pm 2) \text{ E10}$	2.611 ± 0.0014	$−1.512 \pm 0.001$	—	—
Nonturbulent conditions, but only using initially drizzle-free cloud size distributions							
0.89	−54/−2/113		$(4 \pm 2) \text{ E17}$	4.08 ± 0.02	$−2.25 \pm 0.02$	—	—
Accretion							
Nonturbulent conditions							
0.996	−22/−3/22		$(94.8 \pm 1.5) \text{ E5}$	1.4030 ± 0.0007	$−0.3147 \pm 0.0006$	1.3069 ± 0.0004	$−0.2389 \pm 0.0004$
0.997	−25/−3/29		$(89 \pm 1) \text{ E3}$	1.4159 ± 0.0006	$−0.3018 \pm 0.0005$	1.1172 ± 0.0001	—
0.960	−64/27/244		$(631 \pm 9) \text{ E−4}$	1.9603 ± 0.0006	$−0.6487 \pm 0.0005$	—	1.2153 ± 0.0001
0.996	−23/−6/21		69.5 ± 0.2	1.1476 ± 0.0002	—	1.1587 ± 0.0001	—
Turbulent conditions with a dissipation rate of $400 \text{ cm}^2 \text{ s}^{-3}$							
0.997	−20/−5/17		55.2 ± 0.1	1.1287 ± 0.0003	—	1.1462 ± 0.0001	—

The correlation coefficient (Corr.), the 25th, 50th, and 75th percentile errors (%) are also listed. The rows in bold indicate the set of parameters used for comparisons to the ML model in Figures 2 and 3.

^aInstead of q_r , we use $q_r / (q_r + q_c)$ in the formula, motivated by Seifert and Beheng (2001).

Figures 2g, 2h, 3e, and 3f show that the parameterizations from Equations 6 and 7 perform well, but not as well as the standard ML model. The ML model is therefore the desired choice, but Equations 6 and 7 remain a good option compared to KK.

We also performed the minimization technique on the drizzle initialization data set using only q_c and N_c , resulting in

$$P_{au} = 4 \times 10^{17} q_c^{4.08} N_c^{-2.25}. \quad (8)$$

This is in very close agreement with the analytical expression in Seifert and Beheng (2001) that was based on a gamma distribution for the cloud droplet mass and the Long kernel, in the limit of drizzle initiation.

5. The Dependence of Autoconversion on Drizzle Number Concentration N_r

Results in Sections 3 and 4 demonstrate that both P_{au} and P_{ac} are influenced by cloud and drizzle simultaneously. The influence of cloud and drizzle properties on accretion makes sense and is consistent with collision-coalescence theory and existing parameterizations, but the influence of drizzle on autoconversion is less straightforward.

From the definition of autoconversion, one should not expect a causal relationship between N_r and P_{au} . Instead, the dependence of P_{au} on N_r represents the influence on P_{au} from the evolution stage of the cloud DSD, which is related to the appearance of raindrops. Such an influence was first pointed out by Cotton (1972), followed by Seifert and Beheng (2001) who incorporated this associated relationship using the ratio of q_r to total water content (shown as one of the options in Table 1). Zeng and Li (2020) also demonstrated that q_r is a good predictor of the width of the cloud droplet size distribution, and thus P_{au} . However, it remains unclear what is the best form to describe this associated relationship, and whether q_r and N_r are equally effective predictors.

To understand whether N_r contains different information from q_r and whether N_r is an effective predictor for all coalescence regimes, we conducted a number of ML tests (see Table S2). For the regime of $P_{au} \geq P_{ac}$, we found the use of (q_c, N_c, q_r, N_r) remains the best, and the performances from (q_c, N_c, N_r) and (q_c, N_c, q_r) are similar. For all regimes, the use of (q_c, N_c, q_r, N_r) is better than (q_c, N_c, N_r) , and the latter is better than (q_c, N_c, q_r) . These suggest that q_r and N_r contain different information. These also suggest that q_r and N_r are equally effective predictors of the autoconversion-dominant regime, but N_r is a better choice for all regimes. This is understandable as q_r depends also on the accretion rate, while N_r is not affected by accretion.

We further turn to the SCE used to generate the data set to seek a theoretical basis for the associated relationship. The rate of change in the total drop number concentration (N) can be derived from

$$\frac{dN}{dt} = -\frac{1}{2} \int_0^{\infty} \int_0^{\infty} K(x, y) n(x) n(y) dx dy \quad (9)$$

in which x and y represent the mass per drop; $n(\cdot)$ is the drop number concentration as function of mass; and $K(x, y)$ is the collection kernel.

Defining cloud droplets as those drops with mass smaller than x_0 , and rain drops as those with mass larger than x_0 allows us to decompose Equation 9 as:

$$\begin{aligned} \frac{dN}{dt} = & -\frac{1}{2} \int_0^{x_0} \int_0^{x_0} K(x, y) n(x) n(y) dx dy \\ & -\frac{1}{2} \int_0^{x_0} \int_{x_0}^{\infty} K(x, y) n(x) n(y) dx dy \\ & -\frac{1}{2} \int_{x_0}^{\infty} \int_0^{x_0} K(x, y) n(x) n(y) dx dy \\ & -\frac{1}{2} \int_{x_0}^{\infty} \int_{x_0}^{\infty} K(x, y) n(x) n(y) dx dy. \end{aligned} \quad (10)$$

The first integral in Equation 10 denotes the changes in N due to interactions between cloud droplets, that is, via self-collection and autoconversion. The second and third integrals denote interactions between cloud and rain drops, that is, via accretion. The last integral contains interactions between rain drops, that is, via self-collection of rain drops. To proceed analytically, we use the kernels of Long (1974),

$$\begin{aligned} K(x, y) = & k_c (x^2 + y^2), & \text{for } x \text{ and } y < x_0; \\ K(x, y) = & k_r (x + y), & \text{for } x \text{ or } y \geq x_0 \end{aligned} \quad (11)$$

as an example, leading to:

$$\frac{dN}{dt} = \underbrace{-k_c N_c Z_c}_{\text{1st integral}} - \underbrace{k_r (N_c q_r + q_c N_r)}_{\text{2nd \& 3rd integral}} - \underbrace{k_r N_r q_r}_{\text{last integral}}, \quad (12)$$

where Z_c is defined as:

$$Z_c \equiv \int_0^{x_0} x^2 n(x) dx = \int_0^{x_0} x \cdot xn(x) dx = \int_0^{x_0} xq(x) dx, \quad (13)$$

proportional to the radar reflectivity in the Rayleigh regime. As explained above, the underlying processes for each term in Equation 12 are:

$$\begin{aligned} -k_c N_c Z_c &= \left. \frac{dN_c}{dt} \right|_{sc} + \left. \frac{dN_c}{dt} \right|_{au} \\ -k_r (N_c q_r + q_c N_r) &= \left. \frac{dN_c}{dt} \right|_{ac} \\ -k_r N_r q_r &= \left. \frac{dN_r}{dt} \right|_{sc} \end{aligned} \quad (14)$$

where the subscript “sc” denotes self-collection.

Similar to Equation 9, we can perform the integrals for q_c , that is,

$$\left. \frac{dq_c}{dt} \right|_{ac} = - \int_0^{x_0} \int_{x_0}^{\infty} xK(x,y)n(x)n(y)dydx = -k_r \int_0^{x_0} \int_{x_0}^{\infty} x(x+y)n(x)n(y)dydx$$

which leads to

$$\left. \frac{dq_c}{dt} \right|_{ac} = -k_r (N_r Z_c + q_c q_r). \quad (15)$$

Let us define

$$q_c = \bar{x}_c N_c, \quad (16)$$

where \bar{x}_c is the average mass per cloud droplet. Then, we find

$$\frac{dq_c}{dt} = \frac{d(\bar{x}_c N_c)}{dt} = \bar{x}_c \frac{dN_c}{dt} + N_c \frac{d\bar{x}_c}{dt}. \quad (17)$$

Since

$$\begin{aligned} \frac{dq_c}{dt} &= \left. \frac{dq_c}{dt} \right|_{au} + \left. \frac{dq_c}{dt} \right|_{ac}, \text{ and} \\ \frac{dN_c}{dt} &= \left. \frac{dN_c}{dt} \right|_{sc} + \left. \frac{dN_c}{dt} \right|_{au} + \left. \frac{dN_c}{dt} \right|_{ac}, \end{aligned} \quad (18)$$

we can combine Equations 17 and 18 to express the autoconversion process rate as:

$$\begin{aligned}
 P_{au} &= -\left. \frac{dq_c}{dt} \right|_{au} = -\left. \frac{dq_c}{dt} + \frac{dq_c}{dt} \right|_{ac} = -\bar{x}_c \frac{dN_c}{dt} - N_c \frac{d\bar{x}_c}{dt} + \left. \frac{dq_c}{dt} \right|_{ac} \\
 &= -\bar{x}_c \left(\left. \frac{dN_c}{dt} \right|_{sc} + \left. \frac{dN_c}{dt} \right|_{au} + \left. \frac{dN_c}{dt} \right|_{ac} \right) + \left. \frac{dq_c}{dt} \right|_{ac} - N_c \frac{d\bar{x}_c}{dt}.
 \end{aligned} \tag{19}$$

Inserting Equations 14–16 into Equation 19, yields

$$\begin{aligned}
 P_{au} &= \bar{x}_c \left(k_c N_c Z_c + k_r (N_c q_r + q_c N_r) \right) - k_r (N_r Z_c + q_c q_r) - N_c \frac{d\bar{x}_c}{dt} \\
 &= k_c q_c Z_c + k_r \bar{x}_c q_c N_r - k_r N_r Z_c - N_c \frac{d\bar{x}_c}{dt} \\
 &= k_r q_c^2 \frac{N_r}{N_c} + k_c q_c Z_c - k_r N_r Z_c - N_c \frac{d\bar{x}_c}{dt}
 \end{aligned} \tag{20}$$

with $k_r = 5.78 \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-1}$ and $k_c = 9.44 \times 10^9 \text{ m}^3 \text{ kg}^{-2} \text{ s}^{-1}$ (Long, 1974).

There are a number of interesting features in Equation 20. First, the right-hand side is independent of q_r . Second, the first term on the right-hand side has a remarkable resemblance to our power laws, showing the dependence of P_{au} on N_r and the reciprocal feature between N_c and N_r . Since the separation in Long's kernels was $50 \mu\text{m}$ radii, we extend the derivation of Equation 20 for cases in which the cutoff size is smaller than Long's separation value. As shown in supporting information Text S1, Equation 20 can be rewritten as:

$$P_{au} = k_r q_c^2 \frac{N_r}{N_c} + k_c q_c Z_c - k_r N_r Z_c + \left(k_c q_c \frac{Z_c}{N_c} - k_r \frac{q_c^2}{N_c} - k_c T_c + k_r Z_c \right) \tilde{N}_r - N_c \frac{d\bar{x}_c}{dt}, \tag{21}$$

where T_c is the third moment of the mass distribution, and \tilde{N}_r is the number concentration in the range between x_0 and the drop mass at $50 \mu\text{m}$ radius.

To examine the role of the first term in Equation 21, we calculate all terms for a wide range of size distributions approximated by various combinations of lognormal and gamma distributions with realistic cloud and drizzle properties (Figure S1). Our results show that although the first term alone cannot completely replicate P_{au} , compared to the second, third, \tilde{N}_r -related, and the last term, the first term is closer to P_{au} than each of them in 95%, 100%, 71%, and 100% of all cases, respectively, with the relative contributions of terms dependent on the size distributions. This supports the finding in Table 1 that the first term is a good predictor of P_{au} , and provides evidence as to why our P_{au} estimates show a dependence on N_r , and why the inclusion of N_r in the autoconversion parameterization is beneficial.

6. The Effect of Turbulence

The ML model and power-law parameterizations introduced above are based on SCE calculations in non-turbulent conditions. Since small-scale turbulence can enhance the collection rate (e.g., Ayala et al., 2008; Chen et al., 2018; Grabowski & Wang, 2013; Wang & Grabowski, 2009), we evaluate the turbulence impact by incorporating the enhancement of the collision efficiency tabulated in Wang and Grabowski (2009). Using the enhancement factor under turbulent cloud conditions with a $400 \text{ cm}^2 \text{ s}^{-2}$ dissipation rate, we found that the exponents in power-law relationships did not change significantly (see Table 1). Ignoring the turbulence effects in the ML model leads to $(-15\% \pm 13\%)$ errors in P_{au} and $(-10\% \pm 7\%)$ errors in P_{ac} . The medians of the error histograms are -18% for P_{au} and -7% for P_{ac} . Compared to the median errors introduced by the KK parameterization, which are respectively 45% and -20% for P_{au} and P_{ac} (see Figures 2 and 3), the errors due to turbulence collision effects in our ML model are smaller and can be accounted for if the dissipation rate can be estimated from radar or lidar observations.

7. Summary

We have built machine-learning models to predict autoconversion and accretion rates from cloud and drizzle properties, using cloud probe measurements from the ACE-ENA campaign in the Azores and the stochastic collection equation formulated as a two-moment bin model. Overall, the estimated autoconversion and accretion rates from the machine-learning model agree with the observed rates to within 15% and 5%, respectively. The standard model requires concurrent, separated cloud and drizzle water contents and number concentrations, which can be obtained from in situ observations or retrievals from remote sensing measurements (e.g., Fielding et al., 2015; Mace et al., 2016; Rusli et al., 2017; Wu et al., 2020).

The joint analyses from the machine-learning model and optimization techniques led to a robust dependence of autoconversion on drizzle number concentration. The dependence on drizzle number concentration also shows a reciprocity with the dependence of cloud droplet number concentration. These findings are unexpected, because the autoconversion process represents the coalescence between cloud droplets and is causally only related to cloud properties. However, drizzle number concentration does contain information on the width and evolution of the DSD, and hence indirectly on the autoconversion rate. By using simple collection kernels, we replicate the dependence and reciprocity in theoretical derivations. This implies that these features are physical and can be incorporated to improve parameterizations of autoconversion rate. The power-law parameterizations also suggest that the autoconversion rate relates to cloud droplet number concentration with an exponent of 0.75, that is, smaller than often assumed, which will affect precipitation susceptibility, and therefore warrants further investigation.

Data Availability Statement

ARM data are available online through <http://www.archive.arm.gov>. The work on machine learning used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, under DOE Contract No. DE-AC05-00OR22725. The training and testing data sets, and the machine-learning trained models are available freely in the ARM Archive and in Github (<https://github.com/yang0920colostate/AuAc>).

Acknowledgments

This research was supported by the Office of Science (BER), DOE under Grants DE-SC0021167, DE-SC0013489, DE-SC0020259, and DE-8924302SSC000055. Van Leeuwen was supported by the European Research Council under the CUNDA project 694509.

References

- Abel, S. J., & Boutle, I. A. (2012). An improved representation of the raindrop size distribution for single-moment microphysics schemes. *Quarterly Journal of the Royal Meteorological Society*, *138*, 2151–2162.
- Ahlgren, M., & Forbes, R. (2014). Improving the representation of low clouds and drizzle in the ECMWF model based on ARM observations from the Azores. *Monthly Weather Review*, *142*, 668–685.
- Ayala, O., Rosa, B., Wang, L.-P., & Grabowski, W. W. (2008). Effects of turbulence on the geometric collision rate of sedimenting droplets. Part 1: Results from direct numerical simulation. *New Journal of Physics*, *10*, 075015. <https://doi.org/10.1088/1367-2630/10/7/075015>
- Beheng, K. D. (1994). A parameterization of warm cloud microphysical conversion processes. *Atmospheric Research*, *33*, 193–206.
- Berry, E. X. (1968). Modification of the warm rain process. *Preprints, First national conference on weather modification* (pp. 81–88). Albany, NY: American Meteorological Society.
- Bodas-Salcedo, A., Webb, M. J., Brooks, M. E., Ringer, M. A., Williams, K. D., et al. (2008). Evaluating cloud systems in the Met Office global forecast model using simulated CloudSat radar reflectivities. *Journal of Geophysical Research*, *113*, D00A13. <https://doi.org/10.1029/2007JD009620>
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory Algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.
- Chen, S., Yau, M. K., & Bartello, P. (2018). Turbulence effects of collision efficiency and broadening of droplet size distribution in cumulus clouds. *Journal of the Atmospheric Sciences*, *75*, 203–217. <https://doi.org/10.1175/JAS-D-17-0123.1>
- Cotton, W. R. (1972). Numerical simulation of precipitation development in supercooled cumuli. *Monthly Weather Review*, *100*, 764–784. [https://doi.org/10.1175/1520-0493\(1972\)100<0764:NSOPDI.2.3.CO.2](https://doi.org/10.1175/1520-0493(1972)100<0764:NSOPDI.2.3.CO.2)
- Feingold, G., Kreidenweis, S. M., Stevens, B., & Cotton, W. R. (1996). Numerical simulation of stratocumulus processing of cloud condensation nuclei through collision-coalescence. *Journal of Geophysical Research*, *101*, 21391–21402.
- Fielding, M. D., Chiu, J. C., Hogan, R. J., Feingold, G., Eloranta, E., O'Connor, E. J., & (2015). Joint retrievals of cloud and drizzle in marine boundary layer clouds using ground-based radar, lidar and zenith radiances. *Atmospheric Measurement Techniques*, *8*, 2663–2683. <https://doi.org/10.5194/amt-8-2663-2015>
- Glienke, S., & Mei, F. (2019). *Two-dimensional stereo (2D-S) probe instrument handbook*. Retrieved from https://www.arm.gov/publications/tech_reports/handbooks/doe-sc-arm-tr-233.pdf
- Glienke, S., & Mei, F. (2020). *Fast cloud droplet probe (FCDP) instrument handbook*. Retrieved from https://www.arm.gov/publications/tech_reports/handbooks/doe-sc-arm-tr-238.pdf
- Golaz, J.-C., Horowitz, L. W., & Levy, H. II (2013). Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, *40*, 2246–2251. <https://doi.org/10.1002/grl.5023>
- Grabowski, W. W., & Wang, L.-P. (2013). Growth of cloud droplets in a turbulent environment. *Annual Review of Fluid Mechanics*, *45*, 293–324.

- Hall (1980). A detailed microphysical model within a two-dimensional dynamic framework: Model description and preliminary results. *Journal of the Atmospheric Sciences*, *37*, 2486–2507.
- Hsieh, W. C., Jonsson, H., Wang, L.-P., Buzorius, G., Flagan, R. C., Seinfeld, J. H., & (2009). On the representation of droplet coalescence and autoconversion: Evaluation using ambient cloud droplet size distributions. *Journal of Geophysical Research*, *114*, D07201. <https://doi.org/10.1029/2008JD010502>
- Jing, X., Suzuki, K., Guo, H., Goto, D., Ogura, T., Koshiro, T., & (2017). A multimodel study on warm precipitation biases in global models compared to satellite observations. *Journal of Geophysical Research: Atmospheres*, *122*, 11806–11824. <https://doi.org/10.1002/2017JD027310>
- Kessler, E. (1969). *On the distribution and continuity of water substance in atmospheric circulations. Meteorological monographs (Vol. 10, pp. 1–84)*. Boston, MA: American Meteorological Society. https://doi.org/10.1007/978-1-935704-36-2_1
- Khairoutdinov, M., & Kogan, Y. (2000). A new cloud physics parameterization in a large-eddy simulation model of marine stratocumulus. *Monthly Weather Review*, *128*, 229–243.
- Kingma, D. P., & Ba, J. L. (2015). ADAM: A method for stochastic optimization. *3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*. Retrieved from <http://arxiv.org/abs/1412.6980>
- Lee, H., & Baik, J.-J. (2017). A physically based autoconversion parameterization. *Journal of the Atmospheric Sciences*, *74*, 1599–1616.
- Liu, Y., & Daum, P. H. (2004). Parameterization of the autoconversion process. Part I: Analytical formulation of the Kessler-type parameterizations. *Journal of the Atmospheric Sciences*, *61*, 1539–1548.
- Long, A. B. (1974). On the evaluation of the collection kernel for the coalescence of water droplets. *Journal of the Atmospheric Sciences*, *31*, 1040–1052.
- Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*.
- Mace, G. G., Avey, S., Cooper, S., Lebsock, M., Tanelli, S., & Dobrowalski, G. (2016). Retrieving co-occurring cloud and precipitation properties of warm marine boundary layer clouds with A-Train data. *Journal of Geophysical Research: Atmospheres*, *121*, 4008–4033. <https://doi.org/10.1002/2015JD023681>
- Mei, F., Wang, J., Comstock, J. M., Weigel, R., Krämer, M., Mahnke, C., et al. (2020). Comparison of aircraft measurements during GoAmazon2014/5 and ACRIDICON-CHUVA. *Atmospheric Measurement Techniques*, *13*, 661–684. <https://doi.org/10.5194/amt-13-661-2020>
- Rusli, S. P., Donovan, D. P., & Russchenberg, H. W. J. (2017). Simultaneous and synergistic profiling of cloud and drizzle properties using ground-based observations. *Atmospheric Measurement Techniques*, *10*, 4777–4803. <https://doi.org/10.5194/amt-10-4777-2017>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Seifert, A., & Beheng, K. D. (2001). A double-moment parameterization for simulating autoconversion, accretion and self-collection. *Atmospheric Research*, *59*, 265–281. [https://doi.org/10.1016/S0169-8095\(01\)00126-0](https://doi.org/10.1016/S0169-8095(01)00126-0)
- Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettleman, A., Golaz, J.-C., Bodas-Salcedo, A., et al. (2010). Dreary state of precipitation in global models. *Journal of Geophysical Research*, *115*, D24211. <https://doi.org/10.1029/2010JD014532>
- Tzivion, S., Feingold, G., & Levin, Z. (1987). An efficient numerical solution to the stochastic coalescence equation. *Journal of the Atmospheric Sciences*, *44*(21), 3139–3149.
- Wang, L.-P., & Grabowski, W. W. (2009). The role of air turbulence in warm rain initiation. *Atmospheric Science Letters*, *10*, 1–8.
- Wood, R. (2005). Drizzle in stratiform boundary layer clouds. Part II: Microphysical aspects. *Journal of the Atmospheric Sciences*, *62*, 3034–3050.
- Wu, P., Dong, X., Xi, B., Tian, J., & Ward, D. M. (2020). Profiles of MBL cloud and drizzle microphysical properties retrieved from ground-based observations and validated by aircraft in situ measurements over the Azores. *Journal of Geophysical Research: Atmospheres*, *125*, e2019JD032205. <https://doi.org/10.1029/2019JD032205>
- Wyant, M. C., Bretherton, C. S., Wood, R., Carmichael, G. R., Clarke, A., Fast, J., et al. (2015). Global and regional modeling of clouds and aerosols in the marine boundary layer during VOCALS: The VOCA intercomparison. *Atmospheric Chemistry and Physics*, *15*(1), 153–172.
- Zeng, X., & Li, X. (2020). Two-moment bulk parameterization of the drop collection growth in warm clouds. *Journal of the Atmospheric Sciences*, *77*, 797–811.