Leveraging Machine Learning to Predict the Autoconversion Rates from Satellite Data

Maria Carolina Novitasari
Department of Electronic and Electrical Engineering
University College London
m.novitasari@ucl.ac.uk

Johannes Quaas
Leipzig Institute for Meteorology
Universität Leipzig
johannes.quaas@uni-leipzig.de

Miguel R. D. Rodrigues
Department of Electronic and Electrical Engineering
University College London
m.rodrigues@ucl.ac.uk

Abstract

One way of reducing the uncertainty involved in determining the radiative forcing of climate change is by understanding the interaction between aerosols, clouds, and precipitation processes. This can be studied using high-resolution simulations such as the ICOsahedral Non-hydrostatic Large-Eddy Model (ICON-LEM). However, due to the extremely high computational cost required, this simulation-based approach can only be run for a limited amount of time within a limited area. To address this, we developed new models using emerging machine learning approaches that leverage a plethora of satellite observations providing long-term global spatial coverage. In particular, our machine learning models are capable of capturing the key process of precipitation formation which greatly control cloud lifetime, namely autoconversion rates – the term used to describe the collision and coalescence of cloud droplets responsible for raindrop formation. We validate the performance of our models against simulation data, showing that our models are capable of predicting the autoconversion rates fairly well.

1 Introduction

Climate change represents one of the most pressing challenges afflicting our societies, with governments worldwide, non-governmental organizations, and other stakeholders placing increasing emphasis on measures to successfully tackle and mitigate its adverse effects. A pressing scientific challenge relates to the need to understand aerosol-cloud interactions because these currently represent one of the major sources of uncertainty in determining the radiative forcing responsible for climate change projections (Intergovernmental Panel on Climate Change 2021 - IPCC [2021]).

While aerosols directly affect climate through scattering and absorbing the solar radiation, they also affect climate indirectly through altering the properties of clouds which impacts Earth's energy budget. In particular, increases in aerosol concentrations can lead to a decrease in the cloud droplet radii which increase the albedo of clouds (radiative forcing due to aerosol-cloud interactions, RFaci - Twomey [1974]). Simultaneously, the smaller size of cloud droplets will hence reduce precipitation efficiency and increase the lifetime of cloud (one important adjustment to RFaci - Albrecht [1989], Gryspeerdt et al. [2019], Bellouin et al. [2020]).

Understanding the complex interactions within the Earth (i.e., the interaction between aerosols, clouds, and precipitation) is possible through climate models. However, considering the complexity of the climate system, it is unfeasible to explain all processes occurring on Earth simultaneously. Climate

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

models numerically solve the differential equations describing the fluid dynamics of the atmosphere and ocean on a discrete grid. Processes that are smaller than the grid-scale of the model cannot be solved directly and are thus poorly represented (IPCC [2021]). Therefore, in order to better represent small-scale processes in the atmosphere, there has been a shift towards developing high-resolution models with smaller grid cells. Newest developments allow for very-high-resolution simulations in realistic, weather prediction mode, albeit over limited spatial domains and for short time periods only. One of the high-resolution simulations that can be used to simulate small-scale processes in the atmosphere is the ICON-LEM (Dipankar et al. [2015], Heinze et al. [2017]). However, to run such a model for 1 simulated hour, it would take 13 hours on 300 computer nodes, costing about EUR 100,000 per simulated day (Costa-Surós et al. [2020]). Due to its large computational costs, it is impractical to run such a model in larger spatial domains or for a longer period of time.

Meanwhile, the use of machine learning in climate science has not yet been fully explored, despite its tremendous revolutionary impact on many fields of research (Rolnick et al. [2019]). The availability of large datasets coupled with increased computing power has accelerated the development of machine learning, leading to impressive results in solving various challenging problems (Reuther et al. [2019]). It is conjectured that these recent advances in machine learning will enable climate science to make big leaps forward, specifically to better understand the impact of aerosol-cloud-precipitation interactions on climate.

The overarching objective of this study is therefore to contribute to our understanding of aerosol-cloud-precipitation-climate interactions by leveraging emerging machine learning approaches. In particular, we developed machine learning models capable of capturing one aspect of precipitation formation that leverage a plethora of satellite observations – which offer global spatial coverage up to several decades – in order to estimate the autoconversion rates. There have been a variety of parameterizations of autoconversion rates developed in the past, such as Berry [1967], Kessler [1969], Manton and Cotton [1977], Baker [1993], Khairoutdinov and Kogan [2000], Liu and Daum [2004], Seifert and Beheng [2006], among numerous others. Attempts to predict autoconversion rates by using machine learning have also been widely undertaken in the last few years, such as Seifert and Rasp [2020], Chiu et al. [2021], Alfonso and Zamora [2021], etc. Existing studies are primarily concerned with estimating autoconversion rates in climate models, whereas our study estimates the autoconversion rates directly from satellite observations, which has not been attempted by other studies.

2 Proposed Approach: Framework, Datasets, and Machine Learning Models

We present a novel approach for extracting autoconversation rates directly from actual satellite observations. Our general framework can be seen in Figure 1. The left side of the image depicts our dataset and how we handle the data from a climate science perspective in order to obtain the input-output pairs, as explained further in Section 2.1. The parts within the long dash dotted line of the image represent our machine learning framework which will be explained further in Section 2.2.

2.1 Datasets

We use datasets from ICON-LEM output from a simulation of the conditions over Germany on 2 May 2013. We analyse a time period approximately corresponding to the overpass times of the polar-orbiting satellites considered in our study, i.e., the time from 09:55 UTC to 13:20 UTC. Distinct cloud regimes occurred in the model domain on the chosen day (2 May 2013), allowing for the investigation of quite different elements of cloud formation and evolution (Heinze et al. [2017]). There are 150 altitude levels in the vertical grid of this data, with the grid reaching the top of the model at 21 kilometers. ICON-LEM is simulated at 3 different horizontal resolutions, in two-way nests, namely 625, 312 and 156 m. Our study focuses on ICON-LEM with the highest resolution, 156 m.

As part of building our testing and training datasets, we need to calculate the groundtruth – cloud-top autoconversion rates – using ICON-LEM output. As ICON-LEM uses the two-moment microphysical parameterization of Seifert and Beheng [2006] to measure autoconversion rates, we generate autoconversion rates based on this method (see Appendix A.1). We then extract the cloud-top autoconversion rates by selecting the autoconversion rates where the cloud optical thickness exceeds 1. This optical

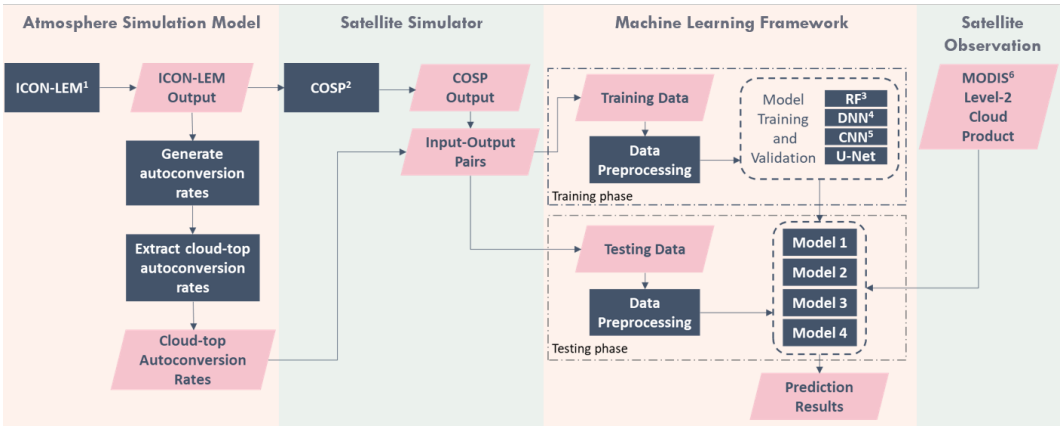

Figure 1: General framework. [1] ICOsahedral Non-hydrostatic Large-Eddy Model; [2] Cloud Feedback Model Intercomparison Project Observation Simulator Package; [3] Random Forest; [4] Deep Neural Network;[5] Convolutional Neural Network; [6] Moderate Resolution Imaging Spectroradiometer

thickness corresponds to the penetration depth for retrievals of cloud microphysical quantities by the optical satellite sensors considered in our study.

To obtain our training input, we feed the ICON-LEM simulation output to the Cloud Feedback Model Intercomparison Project Observation Simulator Package (COSP). COSP is an integrated satellite simulator developed by the Cloud Feedback Model Intercomparison Project (CFMIP) which provide model intercomparison and model–satellite comparison (Bodas-Salcedo et al. [2011], Swales et al. [2018]). In this case, COSP was used to fit the simulation model (ICON-LEM) into satellite format, considering satellite retrieval conditions and limitations. The output of COSP is then combined with cloud-top autoconversion rates to form input-output data pairs where COSP output corresponds to input and cloud-top autoconversion rates correspond to output.

As for the satellite observation data displayed on the right side of Figure 1, we use Collection 6 of level-2 (L2) cloud product of Terra and Aqua MODIS (Platnick et al. [2017], Platnick et al. [2018]) – MOD06 and MYD06, respectively. MODIS L2 cloud product contains both cloud-top properties and cloud optical properties at pixel level. The overpass time of the Terra satellite is approximately 10.30 h, the one of the Aqua satellite, 13.30 h local time.

2.2 Machine Learning Models

In this study, in order to train machine learning prediction models, we use COSP diagnostics output as input. The machine learning model input includes cloud water path liquid, cloud particle size liquid, cloud optical thickness, and cloud droplet number concentration, since these quantities are the cloud microphysical state parameters typically obtained from satellite retrievals (Platnick et al. [2017], Grosvenor et al. [2018]). We also extract spatial information from COSP and incorporate it into our inputs. The machine learning model output, which also serves as the machine learning model groundtruth, is the cloud-top autoconversion rates derived from ICON-LEM. For convenience, we will refer to cloud-top autoconversion rates as autoconversion rates for the remainder of this study.

As a standard practice in machine learning, we split our data into a training set (containing 80% of original data) and a testing set (containing 20% of the original data). During both training and testing phases, we normalise the input and output variables using logarithmic and standard scales. Refer to Appendix A.2 for more details on data preprocessing.

We then selected, trained, and tested various machine learning models to predict desired output from input, including random forest (RF - Breiman [2001]), deep neural network (DNN - Schmidhuber [2015]), convolutional neural network (CNN - LeCun et al. [2015]), and U-Net (Ronneberger et al. [2015]). However, other machine learning methods could potentially be employed. Details of each model are provided in Appendix B.

Table 1: Evaluation of autoconversion prediction results on satellite simulator (COSP) based on Log R-Squared (Log R^2) , Mean Absolute Error (MAE), Mean Squared Error (MSE), Structural Similarity Index (SSIM), and Peak signal-to-noise ratio (PSNR).

Method	Log R^2	MAE ($\mathrm{kg\,m^{-3}\,s^{-1}}$)	MSE ($\mathrm{kg\,m^{-3}\,s^{-1}}$)	SSIM	PSNR (dB)
RF	**66.18%**	6.75e-09	4.21e-15	96.66%	38.78
DNN	66.14%	**6.68e-09**	**3.92e-15**	**96.80%**	**39.08**
CNN	65.09%	6.72e-09	4.09e-15	96.70%	38.91
U-Net	63.55%	7.81e-09	6.37e-15	96.19%	36.98

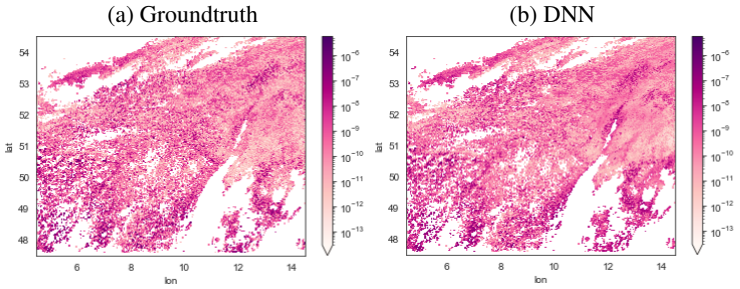

Figure 2: A comparison of the autoconversion rates predicted by DNN with groundtruth.

The performance of each model is then evaluated by calculating the Log R-Squared, MAE, MSE, SSIM, and PSNR on the testing data. In the early stages of testing, we compared the performance of our models against the simulation data. Finally, we use our best model to predict the autoconversion rate directly from satellite data.

3 Experimental Results

Autoconversion on Simulation Models (ICON/COSP) This set of experiments seeks to establish whether one can predict autoconversion rates from simulated satellite data. The results of the first series of experiments are shown in Table 1. These findings show that all models produce good outcomes, with SSIM index exceeding 96%. In terms of MAE, MSE, SSIM, and PSNR, DNN outperforms other models. Despite the fact that U-Net is slightly less accurate, the differences between the performance of the four approaches are minor. Figure 2 shows the visualization of autoconversion rates predicted using DNN compared to groundtruth. As can be seen from the figure, visually, the model is capable of identifying the key features of the groundtruth to a good extent, indicating that the model is performing fairly well.

Autoconversion on Satellite Observation (MODIS) This set of experiments seeks to establish whether one can predict autoconversion rates from real satellite data, i.e., this stage involves testing our models on satellite data. Among the 4 models, we choose the best model to test on MODIS data, which is the DNN model. Our model is tested on MODIS Aqua over Germany at 13:20 UTC with a specific area of interest. The latitude ranges from 47.50° N to 54.50° N. The longitude ranges between 5.87° E and 10.00° E.

Despite the fact that the prediction results of the satellite cannot be directly compared to the simulation model – as despite its high quality, the ICON-LEM simulation does not perfectly place the clouds in the exact right positions – the MODIS autoconversion rates predictions demonstrate statistical agreement with the COSP autoconversion rates, as shown in Table 2. The median, 25th and 75th percentiles of the autoconversion rate of COSP compared to MODIS are quite close, although there is a slight difference in the mean values, but this is believed to be due to discrepancies in the COSP and MODIS variables. This indicates that our approach is capable of estimating autoconversion rates directly from satellite data. Additional results can be found in Appendix C.

Table 2: Mean, standard deviation, median, 25th and 75th percentiles of COSP and MODIS variables: liquid water path (LWP), cloud effective radius (CER), cloud optical thickness (COT), cloud droplet number concentration (N_c), and autoconversion rates (Aut).

	Mean	Standard Deviation	25th Percentile	Median	75th Percentile
LWP COSP ($\mathrm{g\,m^{-2}}$)	73.7	128	10.3	30.8	82.8
LWP MODIS ($\mathrm{g\,m^{-2}}$)	113.0	265	17.0	37.0	98.0
CER COSP (μm)	10.80	5.06	7.34	9.65	13.00
CER MODIS (μm)	12.30	7.28	7.75	9.40	13.90
COT COSP	9.53	13.30	1.59	4.87	11.90
COT MODIS	14.50	24.10	2.15	5.83	17.40
Nc COSP ($\mathrm{cm^{-3}}$)	178	205	45.3	108	236
Nc MODIS ($\mathrm{cm^{-3}}$)	177	179	38.3	124	265
Aut COSP ($\mathrm{kg\,m^{-3}\,s^{-1}}$)	1.77e-08	1.32e-07	2.66e-11	2.12e-10	1.85e-09
Aut MODIS ($\mathrm{kg\,m^{-3}\,s^{-1}}$)	6.09e-08	5.74e-07	2.19e-11	1.02e-10	1.19e-09

4 Conclusion

In this study, we explored how machine learning could help unravel the key process of precipitation formation for liquid clouds, the autoconversion process. This process is a key in better understanding the response of clouds to anthropogenic aerosols (Mülmenstädt et al. [2020]). Future research should include more comparisons with alternative approaches as well as data from other climate models or satellite observations.

Acknowledgments and Disclosure of Funding

We would like to thank two anonymous reviewers for their feedback and comments. This research receives funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI).

References

IPCC. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen,. *Cambridge Univ. Press*, (In Press):3949, 2021. URL `https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC{_}AR6{_}WGI{_}Full{_}Report.pdf`.

S. Twomey. Pollution and the planetary albedo. *Atmospheric Environment (1967)*, 8(12):1251–1256, 1974. ISSN 0004-6981. doi: https://doi.org/10.1016/0004-6981(74)90004-3. URL `https://www.sciencedirect.com/science/article/pii/0004698174900043`.

B. A. Albrecht. Aerosols, cloud microphysics, and fractional cloudiness. *Science*, 245(4923):1227–1230, 1989. ISSN 0036-8075. doi: 10.1126/science.245.4923.1227. URL `https://science.sciencemag.org/content/245/4923/1227`.

E. Gryspeerdt, T. Goren, O. Sourdeval, J. Quaas, J. Mülmenstädt, S. Dipu, C. Unglaub, A. Gettelman, and M. Christensen. Constraining the aerosol influence on cloud liquid water path. *Atmospheric Chemistry and Physics*, 19(8):5331–5347, 2019. doi: 10.5194/acp-19-5331-2019. URL `https://acp.copernicus.org/articles/19/5331/2019/`.

N. Bellouin, J. Quaas, E. Gryspeerdt, S. Kinne, P. Stier, D. Watson-Parris, O. Boucher, K. S. Carslaw, M. Christensen, A.-L. Daniau, J.-L. Dufresne, G. Feingold, S. Fiedler, P. Forster, A. Gettelman, J. M. Haywood, U. Lohmann, F. Malavelle, T. Mauritsen, D. T. McCoy, G. Myhre, J. Mülmenstädt, D. Neubauer, A. Possner, M. Rugenstein, Y. Sato, M. Schulz, S. E. Schwartz, O. Sourdeval, T. Storelvmo, V. Toll, D. Winker, and B. Stevens. Bounding global aerosol radiative forcing of climate change. *Reviews of geophysics.*, 58(1), 2020. ISSN 8755-1209. doi: doi:10.1029/2019RG000660.

A. Dipankar, B. Stevens, R. Heinze, C. Moseley, G. Zängl, M. Giorgetta, and S. Brdar. Large eddy simulation using the general circulation model icon. *Journal of Advances in Modeling Earth Systems*, 7(3):963–986, 2015. doi: https://doi.org/10.1002/2015MS000431. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015MS000431.

R. Heinze, A. Dipankar, C.C. Henken, C. Moseley, O. Sourdeval, S. Trömel, X. Xie, P. Adamidis, F. Ament, H. Baars, C. Barthlott, A. Behrendt, U. Blahak, S. Bley, S. Brdar, M. Brueck, S. Crewell, H. Deneke, P. Di Girolamo, R. Evaristo, J. Fischer, C. Frank, P. Friederichs, T. Göcke, K. Gorges, L. Hande, M. Hanke, A. Hansen, H.C. Hege, C. Hoose, T. Jahns, N. Kalthoff, D. Klocke, S. Kneifel, P. Knippertz, A. Kuhn, T. van Laar, A. Macke, V. Maurer, B. Mayer, C.I. Meyer, S.K. Muppa, R.A.J. Neggers, E. Orlandi, F. Pantillon, B. Pospichal, N. Röber, L. Scheck, A. Seifert, P. Seifert, F. Senf, P. Siligam, C. Simmer, S. Steinke, B. Stevens, K. Wapler, M. Weniger, V. Wulfmeyer, G. Zängl, D. Zhang, and J. Quaas. Large-eddy simulations over Germany using ICON: a comprehensive evaluation. *Q. J. R. Meteorol. Soc.*, 143(702):69–100, 2017. doi: 10.1002/qj.2947.

M. Costa-Surós, O. Sourdeval, C. Acquistapace, H. Baars, C. Carbajal Henken, C. Genz, J. Hesemann, C. Jimenez, M. König, J. Kretzschmar, N. Madenach, C. I. Meyer, R. Schrödner, P. Seifert, F. Senf, M. Brueck, G. Cioni, J. F. Engels, K. Fieg, K. Gorges, R. Heinze, P. K. Siligam, U. Burkhardt, S. Crewell, C. Hoose, A. Seifert, I. Tegen, and J. Quaas. Detection and attribution of aerosol–cloud interactions in large-domain large-eddy simulations with the icosahedral non-hydrostatic model. *Atmospheric Chemistry and Physics*, 20(9):5657–5678, 2020. doi: 10.5194/acp-20-5657-2020. URL https://acp.copernicus.org/articles/20/5657/2020/.

D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio. Tackling climate change with machine learning, 2019.

A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner. Survey and benchmarking of machine learning accelerators. In *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9, 2019. doi: 10.1109/HPEC.2019.8916327.

E.X. Berry. Cloud droplet growth by collection. *J. Atmos. Sci.*, 24:688–701, 1967. cited By 163.

E. Kessler. *On the Distribution and Continuity of Water Substance in Atmospheric Circulations*, pages 1–84. American Meteorological Society, Boston, MA, 1969. ISBN 978-1-935704-36-2. doi: 10.1007/978-1-935704-36-2_1. URL https://doi.org/10.1007/978-1-935704-36-2_1.

M. J. Manton and W. R. Cotton. Formulation of approximate equations for modeling moist deep convection on the mesoscale., 1977.

M.B. Baker. Variability in concentrations of cloud condensation nuclei in the marine cloud-topped boundary layer. *Tellus, Series B*, 45 B(5):458–472, 1993. doi: 10.3402/tellusb.v45i5.15742. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0027799409&doi=10.3402%2ftellusb.v45i5.15742&partnerID=40&md5=32b5e6e0c7d28fcea4314ce5b2c36538. cited By 46.

M. Khairoutdinov and Y. Kogan. A new cloud physics parameterization in a large-eddy simulation model of marine stratocumulus. *Monthly Weather Review*, 128(1):229–243, 2000. doi: 10.1175/1520-0493(2000)128<0229:ANCPPI>2.0.CO;2. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034027431&doi=10.1175%2f1520-0493%282000%29128%3c0229%3aANCPPI%3e2.0.CO%3b2&partnerID=40&md5=55dbfd32df22bb29686df75f8a20912b. cited By 488.

Y. Liu and P.H. Daum. Parameterization of the autoconversion process. part i: Analytical formulation of the kessler-type parameterization. *Journal of the Atmospheric Sciences*, 61(13):1539–1548, 2004. doi: 10.1175/1520-0469(2004)061<1539:POTAPI>2.0.CO;2. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-2942568878&doi=10.1175%2f1520-0469%282004%29061%3c1539%3aPOTAPI%3e2.0.CO%3b2&partnerID=40&md5=37b9b4266fb3ce56db8d59f4230066ea. cited By 144.

A. Seifert and K. D. Beheng. A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description. *Meteorol. Atmos. Phys.*, 92(1-2):45–66, 2006. ISSN 01777971. doi: 10.1007/s00703-005-0112-4.

A. Seifert and S. Rasp. Potential and Limitations of Machine Learning for Modeling Warm-Rain Cloud Microphysical Processes. *J. Adv. Model. Earth Syst.*, 12(12), 2020. ISSN 19422466. doi: 10.1029/2020MS002301.

J. Christine Chiu, C. Kevin Yang, Peter Jan van Leeuwen, Graham Feingold, Robert Wood, Yann Blanchard, Fan Mei, and Jian Wang. Observational constraints on warm cloud microphysical processes using machine learning and optimization techniques. *Geophysical Research Letters*, 48(2), 1 2021. doi: 10.1029/2020GL091236.

L. Alfonso and J.M. Zamora. A two-moment machine learning parameterization of the autoconversion process. *Atmospheric Research*, 249:105269, 2021. ISSN 0169-8095. doi: https://doi.org/10.1016/j.atmosres.2020.105269. URL https://www.sciencedirect.com/science/article/pii/S0169809520312060.

A. Bodas-Salcedo, M. J. Webb, S. Bony, H. Chepfer, J. L. Dufresne, S. A. Klein, Y. Zhang, R. Marchand, J. M. Haynes, R. Pincus, and V. O. John. COSP: Satellite simulation software for model assessment. *Bull. Am. Meteorol. Soc.*, 92(8):1023–1043, 2011. doi: 10.1175/2011BAMS2856.1.

D.J. Swales, R. Pincus, and A. Bodas-Salcedo. The Cloud Feedback Model Intercomparison Project Observational Simulator Package: Version 2. *Geosci. Model Dev.*, 11(1):77–81, 2018. ISSN 19919603. doi: 10.5194/gmd-11-77-2018.

S. Platnick, K.G. Meyer, M.D. King, G. Wind, N. Amarasinghe, B. Marchant, G.T. Arnold, Z. Zhang, P.A. Hubanks, R.E. Holz, P. Yang, W.L. Ridgway, and J. Riedi. The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples from Terra and Aqua. *IEEE Trans. Geosci. Remote Sens.*, 55 (1):502–525, 2017. doi: 10.1109/TGRS.2016.2610522.

S. Platnick, K.G. Meyer, M.D. King, G. Wind, N. Amarasinghe, B. Marchant, G.T. Arnold, Z. Zhang, P.A. Hubanks, B. Ridgway, and J. Riedi. MODIS Cloud Optical Properties: User Guide for the Collection 6/6.1 Level-2 MOD06/MYD06 Product and Associated Level-3 Datasets. 2018. URL https://modis-atmos.gsfc.nasa.gov/sites/default/files/ModAtmo/MODISCloudOpticalPropertyUserGuideFinal{_}v1.1{_}1.pdf.

D. P. Grosvenor, O. Sourdeval, P. Zuidema, A. Ackerman, M. D. Alexandrov, R. Bennartz, R. Boers, B. Cairns, C. Chiu, M. Christensen, H. Deneke, M. Diamond, G. Feingold, A. Fridlind, A. Hünerbein, C. Knist, P. Kollias, A. Marshak, D. McCoy, D. Merk, D. Painemal, J. Rausch, D. Rosenfeld, H. Russchenberg, P. Seifert, K. Sinclair, P. Stier, B. Van Diedenhoven, M. Wendisch, F. Werner, R. Wood, Z. Zhang, and J. Quaas. Remote sensing of droplet number concentration in warm clouds: A review of the current state of knowledge and perspectives. *Reviews of geophysics.*, 56(2):409–453, 2018. ISSN 8755-1209. doi: doi:10.1029/2017RG000593.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015. ISSN 0893-6080. doi: http://dx.doi.org/10.1016/j.neunet.2014.09.003. URL http://www.sciencedirect.com/science/article/pii/S0893608014002135.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015.

J. Mülmenstädt, C. Nam, M. Salzmann, J. Kretzschmar, T. S. L'Ecuyer, U. Lohmann, P. Ma, G. Myhre, D. Neubauer, P. Stier, K. Suzuki, M. Wang, and J. Quaas. Reducing the aerosol forcing uncertainty using observational constraints on warm rain processes. *Science Advances*, 6(22):eaaz6433, 2020. doi: 10.1126/sciadv.aaz6433. URL https://www.science.org/doi/abs/10.1126/sciadv.aaz6433.

D. Czernia and B. Szyk. Air density calculator, Sep 2021. URL https://www.omnicalculator.com/physics/air-density.

J. Quaas, O. Boucher, and U. Lohmann. Constraining the total aerosol indirect effect in the LMDZ and ECHAM4 GCMs using MODIS satellite data. *Atmos. Chem. Phys.*, 6(4):947–955, 2006. ISSN 16807324. doi: 10.5194/acp-6-947-2006.

A Datasets

A.1 Generate Autoconversion Rates

To generate the autoconversion rates based on Seifert and Beheng [2006] approach, we need mass density of cloud droplets or cloud water content (L_c), mass density of raindrops or rain water content (L_r), cloud droplet number concentration (N_c), and air density (ρ) as inputs. The following is the procedure to calculate each input variable:

1. Air density (ρ)
   Given the air temperature, air pressure, and relative humidity, we can measure the air density using the following procedure ((Czernia and Szyk [2021])):
   (a) Measure the saturation vapor pressure at temperature T using the following equation:
   $$p_v = (6.1078 * 10^{[7.5*T/(T+237.3)]}) * RH \tag{1}$$
   where T is temperature in Celsius and RH is the relative humidity
   (b) Find the pressure dry air by subtracting the total air pressure to vapor pressure as follows: $p_d = p - p_v$
   (c) Estimate the air density using:
   $$\rho = \frac{p_d}{R_d * T} + \frac{p_v}{R_v * T} \tag{2}$$
   where: p_d is the pressure of dry air in Pa, p_v is the water vapor pressure in Pa, T is the air temperature in Kelvins, R_d is the specific gas constant for dry air equal to 287.058 J/(kg·K), and R_v is the specific gas constant for water vapor equal to 461.495 J/(kg·K).
2. Cloud water content (L_c)
   We can calculate L_c using mixing ratio large scale cloud liquid (q_l) and ρ, as follows:
   $$L_c = q_l * \rho \tag{3}$$
3. Rain water content (L_r)
   Since L_r is not available in our dataset, we assume:
   $$L_r = 10\% * L_c \tag{4}$$
4. Cloud droplet number concentration (N_c)
   For simulation model, we derive N_c using L_c, ρ, and cloud effective radius (r_e) – hereinafter N_{c1} – as follows:
   $$L_c = N_{c1} * \frac{4\pi}{3} \frac{\rho_w}{\rho} r_e^3 , \tag{5}$$
   with density liquid water (ρ_w) = 1000 $\mathrm{kg\,m^{-3}}$.
   While for satellite data, following Quaas et al. [2006], given cloud-top effective radius (r_e) and cloud optical depth (τ_c), we can specify the N_c – hereinafter N_{c2} – as follows:
   $$N_{c2} = \alpha * \tau_c^{0.5} r_e^{-2.5} , \tag{6}$$
   with $\alpha = 1.37 * 10^{-5}\ \mathrm{m^{-0.5}}$.

A.2 Data Preprocessing

In order to enhance the quality of our data, we preprocess the data as follows:

1. remove nan and zero elements from the data;
2. remove data with cloud particle size liquid more than 40 µm since 40 µm is used to separate cloud droplets and raindrops as used in Seifert and Beheng [2006];
3. remove data with liquid water content (L_c) less than 10^{-4} to exclude noncloudy data points from the pool
4. normalize input-output data using logarithmic and standard scales, which can be expressed as follows:
   $$x' = \frac{\log(x) - \mu}{\sigma} \tag{7}$$
   where μ is the mean value of $\log(x)$ and σ is the standard deviation of $\log(x)$.

Table 3: Random forest hyperparameters

Variable	Value
$n\_estimators$	300
$min\_samples\_split$	15
$min\_samples\_leaf$	8
$max\_features$	sqrt
max_depth	20
$bootstrap$	True

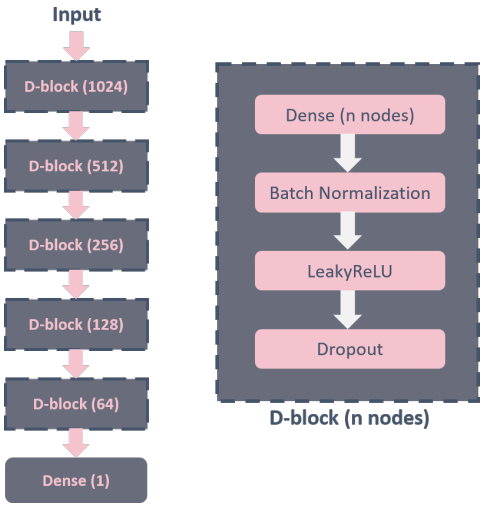

Figure 3: The architecture of our machine learning model using DNN.

B Machine Learning Models

Here, we will examine in more depth each of the models mentioned in Section 2.2. For deep learning approach, the Leaky ReLU serves as the activation function for all the nodes in each layer. While for the loss function, we use mean absolute error (MAE). Furthermore, the training was performed using Adam's optimizer with 32 training samples per batch.

B.1 Model 1: Random Forest

In this experiment, we convert the input data into a dataframe with each column representing each feature and each row representing each data point in a given longitude and latitude. As RF performance is strongly influenced by its hyperparameters, we conduct a random search to find reasonable hyperparameter values. Our final parameters are shown in Table 3.

B.2 Model 2: Deep Neural Network

Within this model, there are five fully connected hidden layers, with 1,024 nodes in the first layer, 512 nodes in the second layer, 256 nodes in the third layer, 128 nodes in the fourth layer, and 64 nodes in the fifth layer. At each hidden layer, a batch normalization with Leaky ReLU as the activation function is performed, followed by a 10% dropout as shown in Figure 3.

B.3 Model 3: Convolutional Neural Network

This model is a simple 1-dimensional CNN model consisting of one convolutional layer with 64 number of filters and one fully connected layer with 64 nodes, as shown in Figure 4.

Figure 4: The architecture of our machine learning model using CNN.

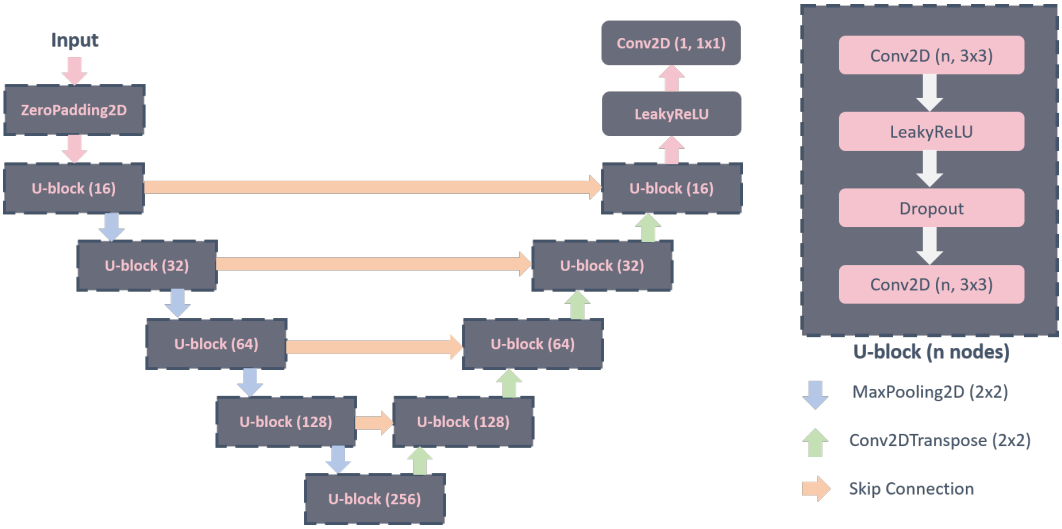

Figure 5: The architecture of our machine learning model using U-Net.

B.4 Model 4: U-Net

This model is based on the U-Net architecture. It consists of two paths, a contracting path and an expansive path. Two convolutions with 3x3 kernel size are applied repeatedly in the contracting path. In between the convolution operations, the Leaky ReLU activation function and dropout operation are applied. This was followed by a 2x2 max pooling operation. At each contracting step, the number of filters is doubled from 16 to 256.

Similarly to the contracting path, each expansive path also involves two convolutions with 3x3 kernel sizes where the Leaky ReLU activation function and dropout operation are applied in between the convolution operations. This was followed by a 2x2 transpose convolution. The number of filters for each expansive path is reduced by half from 256 to 16, which is contrary to a contracting path. In the final layer, there is a Leaky ReLU activation function followed by a convolution operation, as shown in Figure 5.

C Results

Autoconversion on Simulation Models (ICON/COSP) The mean of the prediction results compared to the groundtruth over latitude and longitude can be seen in Figure 6. The blue line represents the groundtruth. A random forest model is illustrated by the orange line. The second model, DNN, is illustrated by the green line. The third model, CNN, is illustrated by the red line. The last model, U-Net, is illustrated by the purple line. From this figure, it can be observed that all models are fairly consistent with the groundtruth.

Autoconversion on Satellite Observation (MODIS) The probability density function of COSP and MODIS variables including the autoconversion rates can be seen in Figure 7. It shows that there is a slight difference at the peak of the probability density function of autoconversion rates, where MODIS is slightly higher than COSP. However, that is understandable, as the probability density function's peak in all variables indicates that MODIS is slightly higher than COSP. Generally, the probability density function of the COSP autoconversion rate compared to MODIS is in relatively good agreement.

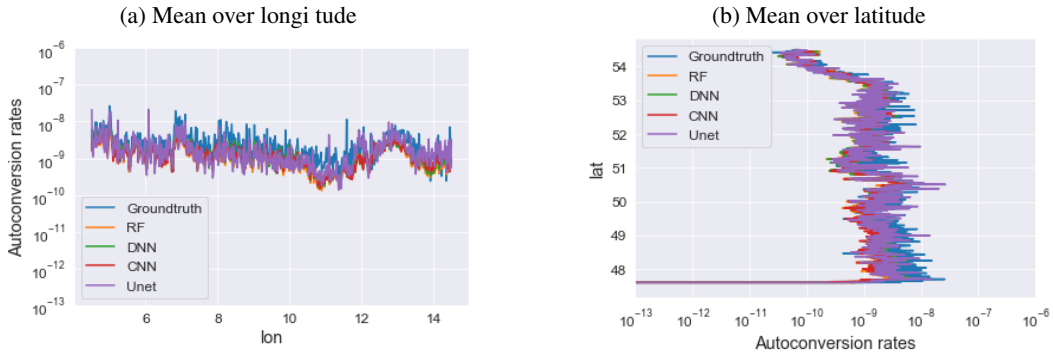

Figure 6: Mean plot of autoconversion rates over latitude and longitude.

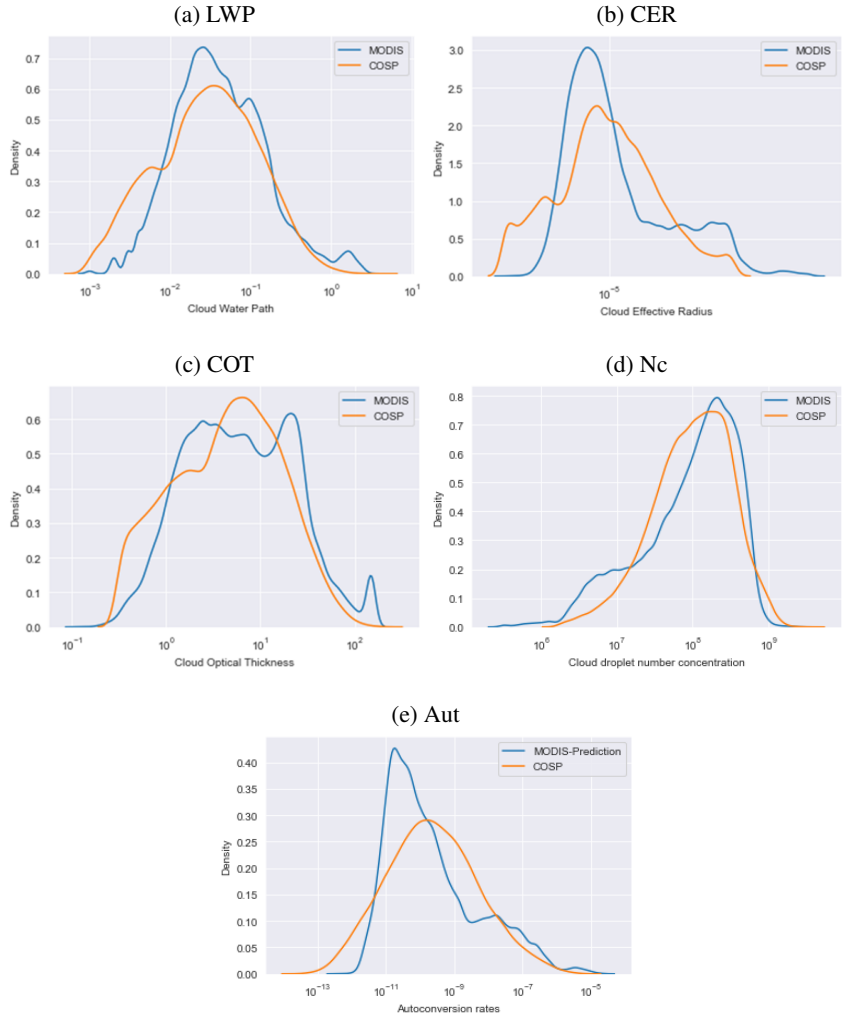

Figure 7: Probability density function of COSP and MODIS variables: liquid water path (LWP), cloud effective radius (CER), cloud optical thickness (COT), cloud droplet number concentration (Nc), and autoconversion rates (Aut).