A novel Python based large eddy simulation capability designed to generate training data for machine learning of microphysical process rates

Kyle Pressel, Colleen Kaul, Jacob Sphund,
Jiwen Fan, Mikhail Ovchinnikov, Po-Lun Ma

2020 ESMD-E3SM PI-Meeting

Motivation

- A goal of the EAGLES project is to develop **better representations of cloud water autoconversion** as part of the project's larger goal of providing improved parameterization of **aerosol cloud interaction** (ACI).
- Models have relied on **curve fits** relating cloud water mass and droplet number concentration to autoconversion rate based on **limited numbers of numerical simulations** of clouds with **explicit representation** of cloud microphysical process. For example:

Khairoutdinov and Kogan (2000):

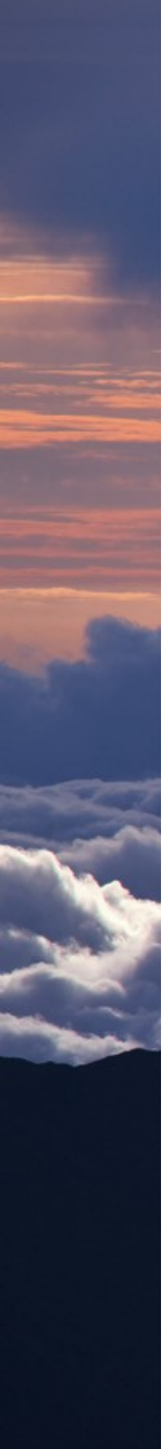
$$\left(\frac{\partial q_r}{\partial t}\right)_{auto} = 1350 q_c^{2.47} N_c^{-1.79}$$

Specialized for stratocumulus

Kogan (2013):

$$\left(\frac{\partial q_r}{\partial t}\right)_{auto} = \left(7.98 \times 10^{10}\right) q_c^{4.22} N_c^{-3.01}$$

Specialized for shallow cumulus



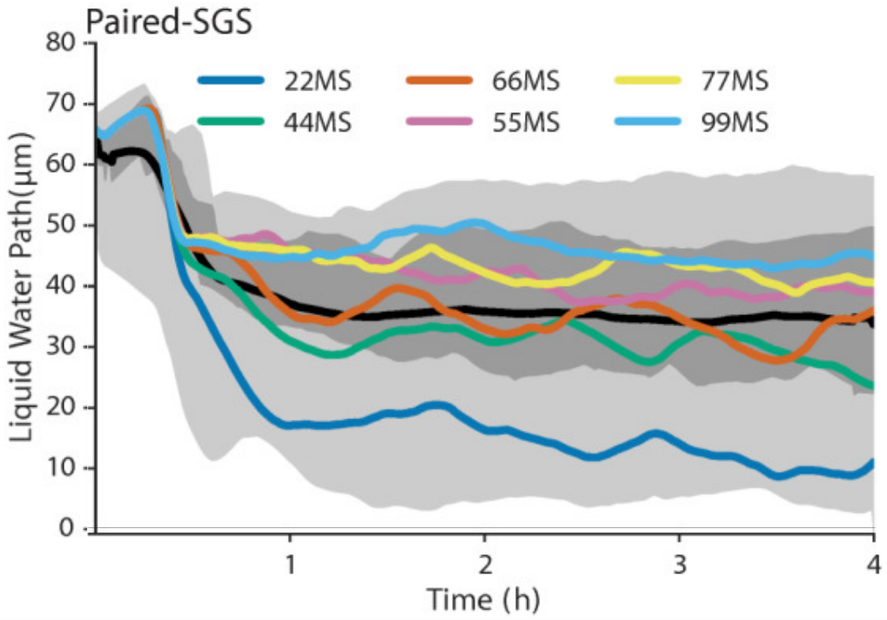
EAGLES adopts a strategy of using machine learning to improve parameterization of autoconversion

- EAGLES uses **large eddy simulations** (LES) with **spectral bin microphysics** (SBM) that explicitly represents cloud water autoconversion for **diverse cloud types** to train **deep neural network** (DNN) representations of autoconversion.
- The machine learning approach taken by EAGLES in many ways parallels previous strategies based on parametric optimization of simple functional forms to explicitly simulated microphysical tendencies, but **replaces the simple functional** forms with much more general **deep neural networks**.
- However, the **accuracy** and **generality** of this approach depends on the **fidelity** and **diversity** of the training data set.

Our machine learning approach places significant demands on LES infrastructures

- The LES training data must span the range of conditions over which the DNN based parameterization is expected to work.
  - Thus we need many LES simulations with many configurations. So we need an LES with high throughput.
- The LES must provide high fidelity simulations over a wide range of conditions, with minimal tuning. However, we know this is a challenge for LES.
- The LES needs to be easily modified to output custom data needed to generate training data.

Liquid water path in one LES code, changing only the numerics and SGS model, can vary by a factor 6!



LES are known to be highly sensitive to model numerics and SGS models

Figure from Pressel et al. (2017)

A challenge to fulfill the demands of ML within existing widely used LES codes

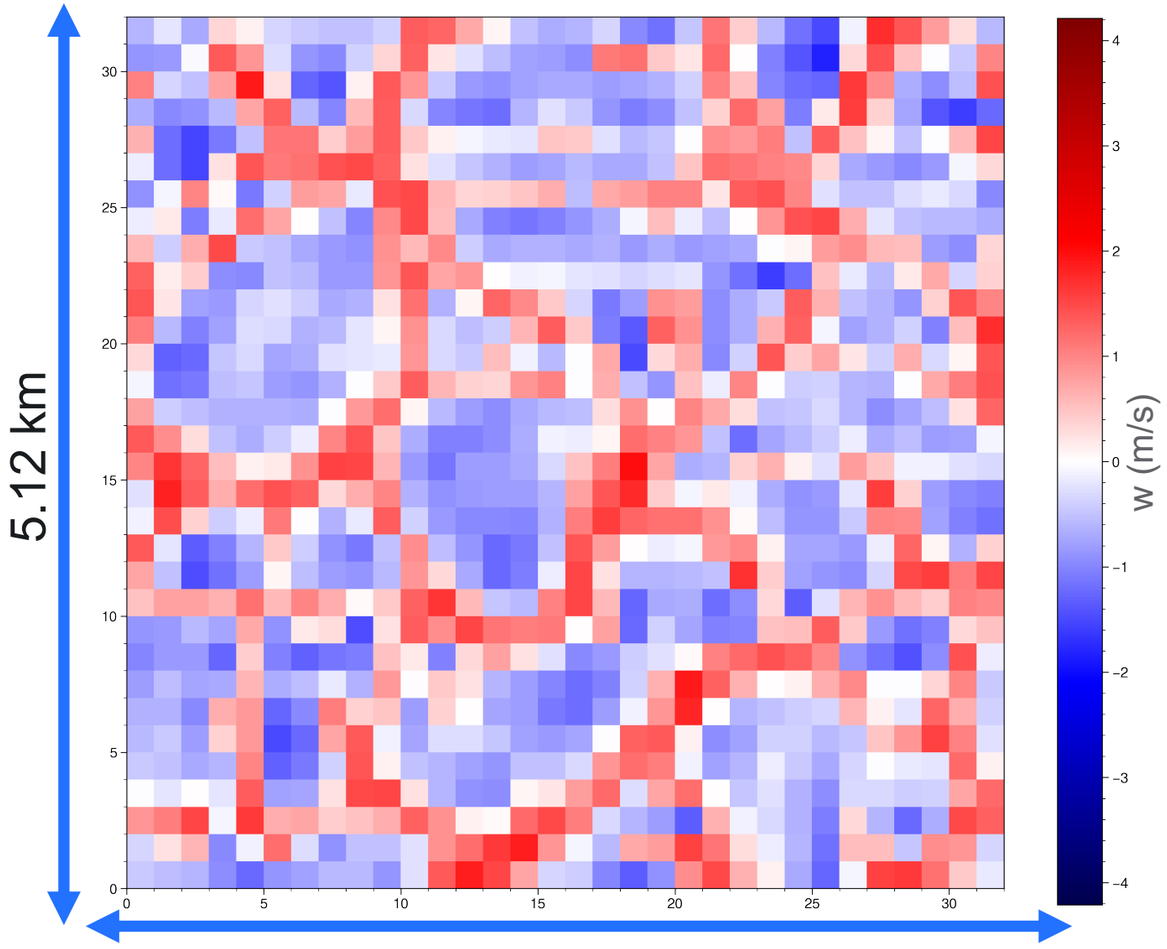
- Weather Research and Forecasting (WRF) model
  - **Compressible dynamical** core with explicit treatment of horizontal acoustics makes it too expensive to perform large ensembles of high resolution LES.
  - Limited flexibility and extensibility to facilitate custom training data output and custom idealized cases.
  - Numerical accuracy limited by vertical grid.
- System for Atmospheric Modeling (SAM)
  - Higher throughput, **anelastic dynamical** core but its numerical options required for general high fidelity LES are limited.
  - Limited flexibility and extensibility to facilitate custom training data output and custom idealized cases.

We decided to write a new LES code from scratch that satisfies the demands of ML. That code is called PINACLES.

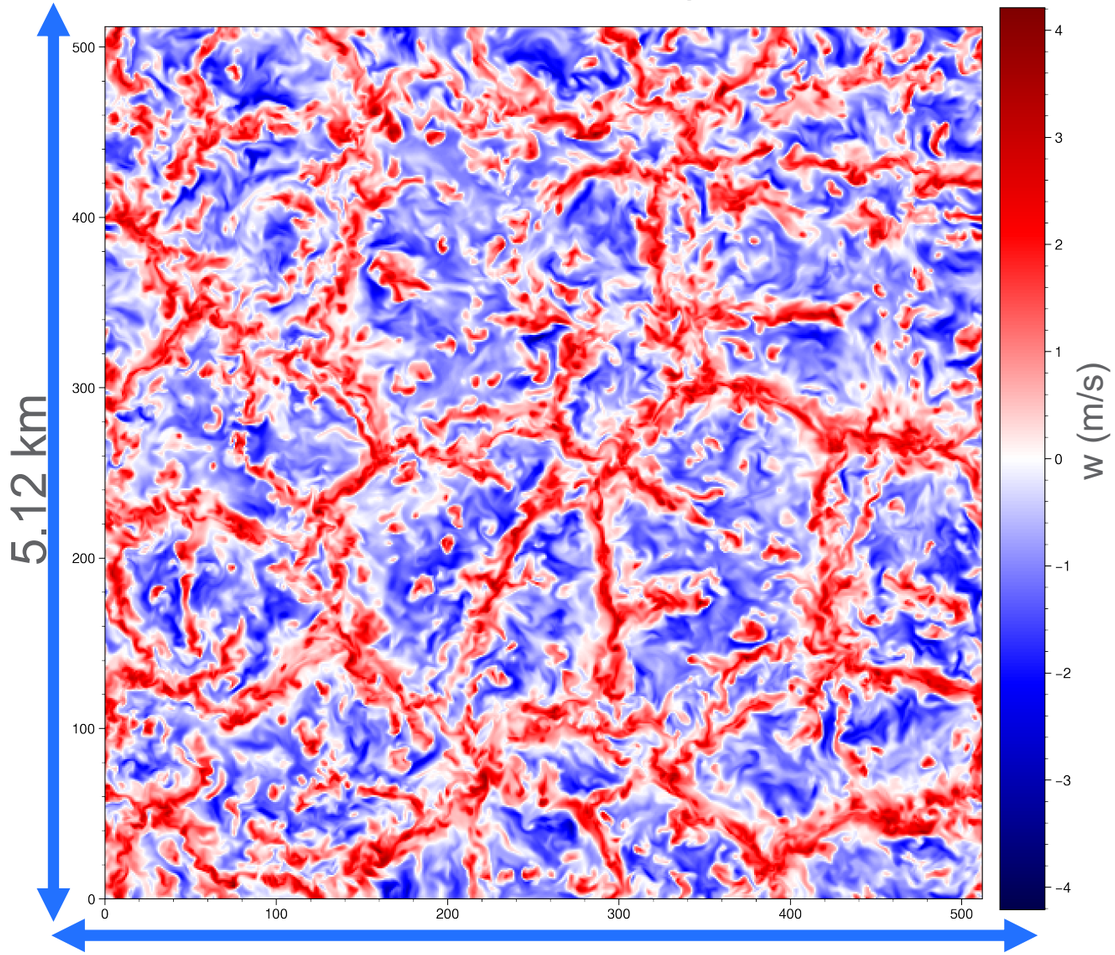
A brief aside: What does using an anelastic dynamical core get you?

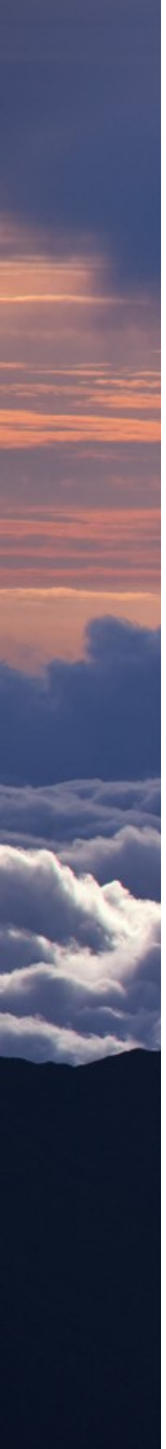
These convective boundary layer simulations are performed on grids of equal extent.

Simulated with maximum resolution allowed by WRF with a timestep of ~1 s, 160 m.



Simulated with maximum resolution allowed by PINACLES with a timestep of ~1 s, 10 m.





Pacific Northwest
NATIONAL LABORATORY

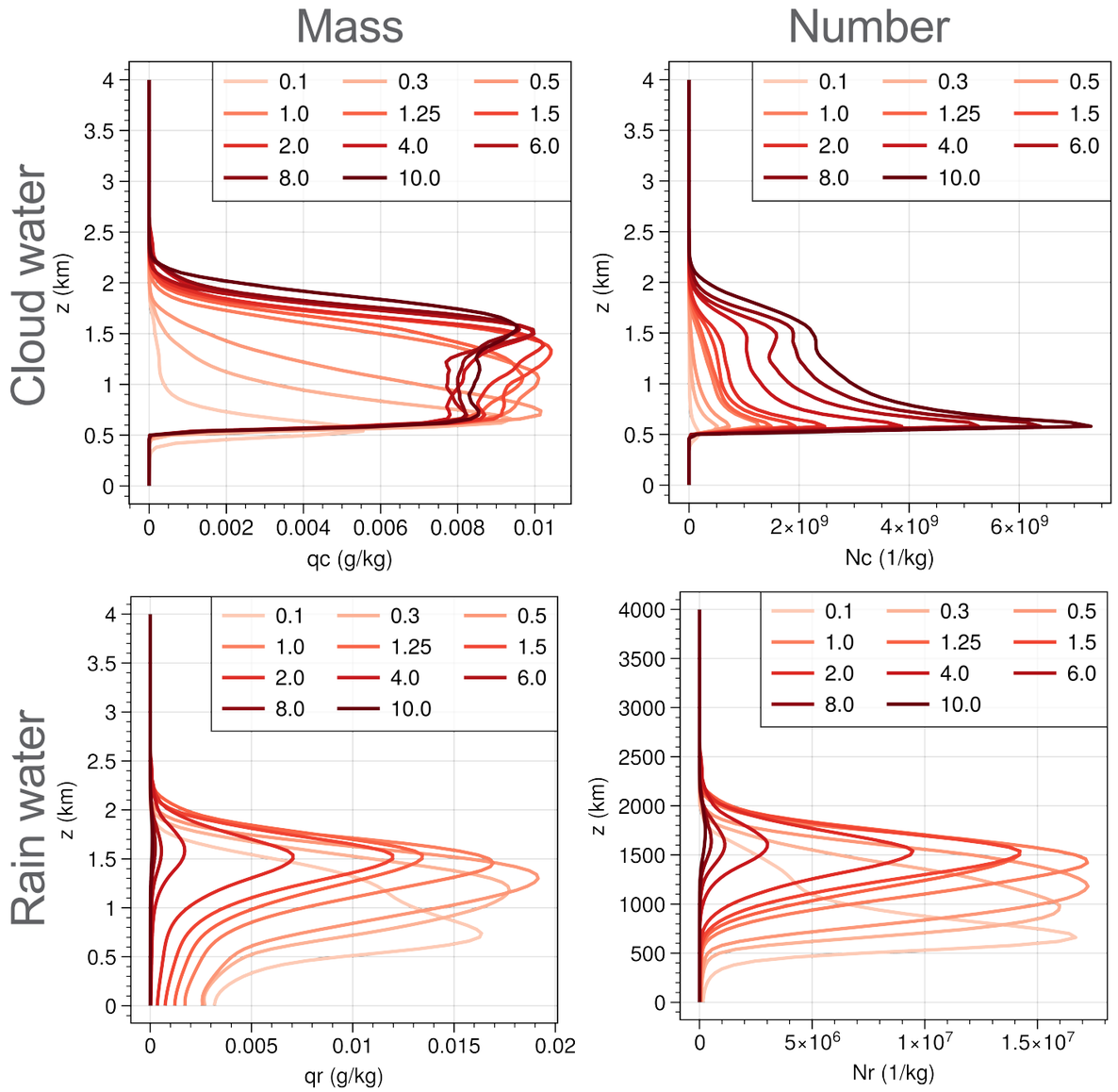
PINACLES: Predicting Interactions of Aerosol and Clouds in Large Eddy Simulations

- A novel, massively parallel LES code developed at PNNL, originally to support the LES needs of the **EAGLES** project.
- Is written entirely in **Python**, accelerated using **Numba**.
- Solves the **anelastic equations** of motion.
- Designed to **enable scientific discovery** by explicitly simulating atmospheric process at scales ranging from tens of centimeters to tens of kilometers.
- Designed to fulfill new and emergent demands made on LES codes, for example by **machine learning,** that are challenging to fulfill with existing models.
- Optimized for **maximum scientific throughput**.
- Achieves maximum scientific throughput by leveraging emergent software engineering strategies to **minimize software complexity** and **development time**.
- **Can adopt existing physics packages** (e.g., microphysics from WRF) to accelerate development.

Generation of ML training data using PINACLES

Pacific Northwest NATIONAL LABORATORY

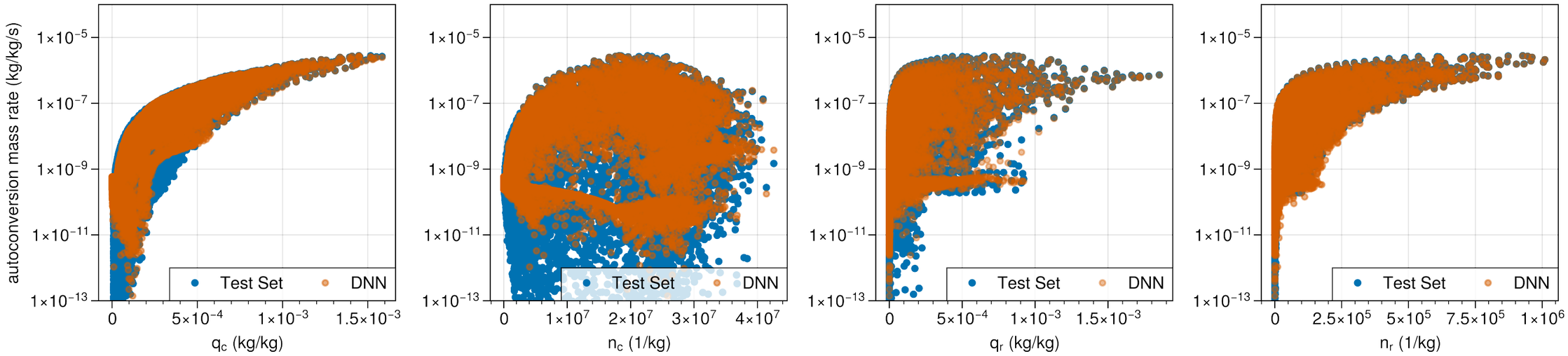
- Training datasets need to span the range of conditions over which the DNN will be applied.
- Here we use PINACLES coupled to the Hebrew University Spectral Bin Model to generate an ensemble of shallow cumulus cloud cases by perturbing the initial aerosol concentration.
- The simulations show expected aerosol cloud interactions, with decreasing (increasing) precipitation (cloud condensate) with increasing aerosol up to a concentration where the effect saturates.
- Cloud properties and microphysical process rates are extracted from the spectral bin model and are used for ML training.



The number corresponding to each color denotes a factor multiplying a baseline initial CCN concentration.

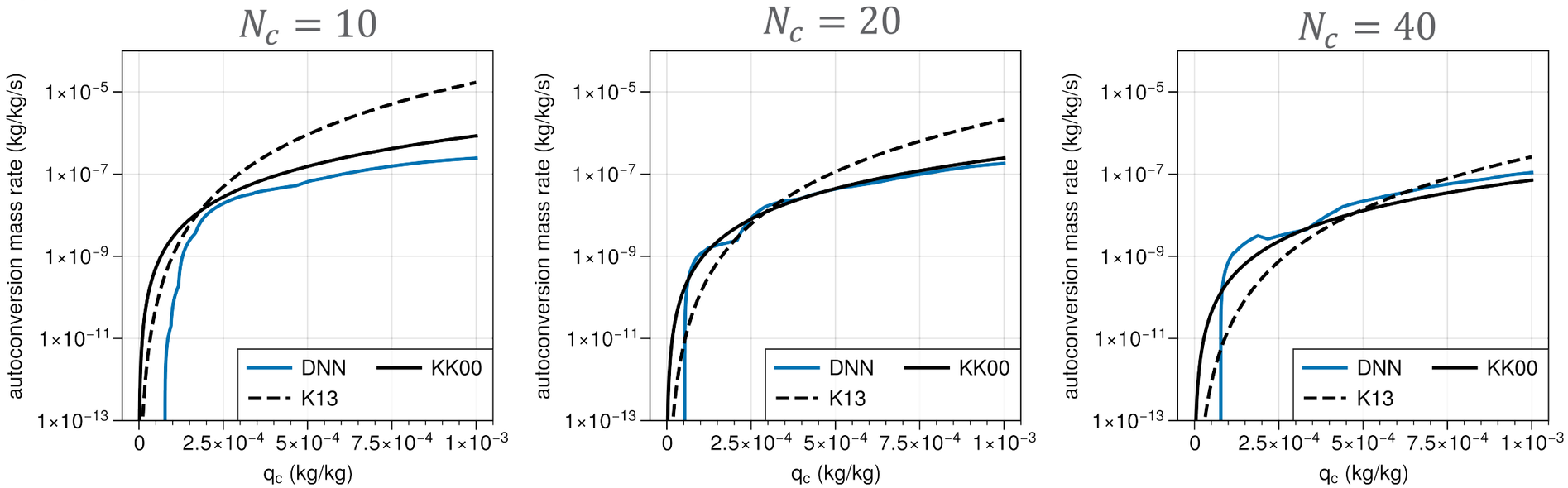
Use of training data for machine learning of autoconversion

- The training data generated by the simulations described in the previous slide are used to train a deep neural network, with three dense hidden layers.
- The training dataset is composed of more than 5 million samples.
- Initial training showed that a DNN mapping of $q_c, N_c, q_r, N_r \rightarrow \left(\frac{\partial q_c}{\partial t}\right)_{auto}, \left(\frac{\partial N_c}{\partial t}\right)_{auto}$ proved effective. This will be extended to prediction of $\left(\frac{\partial N_r}{\partial t}\right)_{auto}$.
- Below the the DNN is evaluated against a test dataset withheld from the training, showing that the DNN is able to capture the relationships between the autoconversion mass tendency and the training variables.



Early comparison of ML based autoconversion to existing, widely used parameterizations

- Below we compare the predictions of autoconversion rate for the Khairoutdinov and Kogan (2000) and Kogan (2013) schemes to the DNN based scheme, for three different cloud droplet number concentrations.
- The DNN scheme requires the number and mass concentration of rain as input. In the plots below we hold these fixed at $N_r = 2e5\ (kg^{-1})$; $q_r = 2.5{\times}10^{-3} (kg\ kg^{-1})$.
- DNN results are qualitatively similar to results from existing parameterizations, with largest differences at low cloud-liquid water concentrations.
- First attempts at training DNNs to represent autoconversion using LES data strongly suggest that it is a viable approach.



Conclusions and next steps

- PINACLES is an entirely new large eddy simulation code written from the ground up to provide the flexibility required to support machine learning workflows.
- PINACLES is coupled to the Hebrew University Spectral Bin microphysics code enabling us to generate the data necessary to train DNNs that represent microphysical process rates.
- Initial DNNs for autoconversion have been trained using data from PINACLES and strongly suggest the viability of the approach.
- PINACLES will be used to simulate a broader range of boundary layer cloud types, increasing the breadth of the training dataset.
- DNNs trained on this data will be implemented and tested in E3SM.