‘Machine Learning Shines a Laser Light on Murky Aerosol Picture’

Estimation of CCN Concentration from Lidar Observations Using Neural Networks

Lan Gao¹ and Jens Redemann¹

Acknowledgement:
1. School of Meteorology, University of Oklahoma
   Emily Lenhardt, Benting Chen, Feng Xu
2. NASA Langley Research Center
   Richard Ferrare, Chris Hostetler, Marta Fenn, Johnathan Hair
Motivation

Global observations of aerosol-cloud-precipitation-climate interactions

Daniel Rosenfeld¹, Meirvat O. Andreae², Ari Asmi³, Mian Chin⁴, Gerrit de Leeuw⁵,⁶, David P. Donovan⁷, Ralph Kahn⁸, Stefan Kinne⁹, Niku Kivekäs⁵,⁶, Markku Kulmala⁵, William Lau¹, K. Sebastian Schmidt¹, Tanja Suni¹, Thomas Wagner¹⁰, Martin Wild¹⁰, and Johannes Quaas¹¹

¹Institute of Earth Sciences, The Hebrew University of Jerusalem, Israel, ²Biogeochemistry Department, Max Planck Institute for Chemistry, Mainz, Germany, ³Department of Physics, University of Helsinki, Helsinki, Finland, ⁴Earth Science Division, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA, ⁵Atmospheric Composition Research Unit, Finnish Meteorological Institute, ⁶ZEUS, Institute of Atmospheric Physics, Freie Universität Berlin, ⁷National Center for Atmospheric Research, Boulder, Colorado, USA, ⁸Institute of Atmospheric Physics, Freie Universität Berlin, ⁹Atmospheric Chemistry Section, Max Planck Institute for Chemistry, ¹⁰Max Planck Institute for Meteorology, Hamburg, Germany, ¹¹Max Planck Institute for Chemistry, Mainz, Germany

An urgent need for global observations of CCN(S) by remote sensing follows from these considerations. Because the microphysical and radiative effects of aerosols act simultaneously on a given cloud population and change the thermodynamic environment of cloud formation and the microphysical processes of the cloud development (Rosenfeld et al., 2008a), the CCN(S) field should be observed simultaneously with aerosol light scattering and absorption properties. Since the effects of light scattering (cooling of the ground surface) and absorption (cooling at the ground combined with heating aloft) have different impacts on atmospheric stability, they must be observed independently. Here, quantitative measures of absorption are especially important.

- Column-effective aerosol quantities may not be relevant to aerosol-cloud interaction.
- The uncertainty of CCN-AOD parameterization can be very large depending on:
  - Types of aerosol
  - Vertical distribution
  - Humidity response of light scattering
  - Spatial-temporal variability

Stier, 2016: “…71 % of the area of the globe shows correlation coefficients between CCN0.2 % at cloud base and aerosol optical depth (AOD) below 0.5, i.e. AOD variability explains only 25 % of the CCN variance” – model-based, self-consistent.
Developed an optimization approach to retrieve the speciated aerosol profiles from lidar and polarimeter observations.

Used aerosol reanalysis product to estimate the bulk hygroscopicity parameter for aerosol mixture.

Applied κ-Köhler theory to calculate CCN concentration.

Limitations – AOS retrieval simulation “lessons learned”

- Great dependence on a priori information (aerosol size distribution and chemical composition) to retrieve CCN
- Computationally very expensive.
Methodology

- Collocate HSRL-2 and in-situ measured CCN from multiple campaigns.
  - ACTIVATE, CAMP²EX, DISCOVER-AQ, ORACLES
- Train neural networks for different sets of lidar observables (e.g., NASA AOS, EarthCARE).
  - HSRL-2: $3\beta + 2\alpha + 3\delta$
  - HSRL-1: $2\beta + 1\alpha + 2\delta$
  - EarthCARE: $1\beta + 1\alpha + 1\delta$
  - Simulated-Elastic-Backscatter (SEBL): $2\beta + 2\delta$
- Evaluate model prediction using in-situ measured CCN.
  - Correlation coefficient (R)
  - Mean absolute error (MAE)
  - Mean relative error (MRE)
### Aircraft observations (lidar and in situ) from multiple campaigns

<table>
<thead>
<tr>
<th>Lidar observables</th>
<th>EXT$<em>{355}$, EXT$</em>{532}$, BSC$<em>{355}$, BSC$</em>{532}$, BSC$<em>{1064}$, DEPO$</em>{355}$, DEPO$<em>{532}$, DEPO$</em>{1064}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancillary</td>
<td>Altitude (ALT), Relative humidity (RH), Temperature (T)</td>
</tr>
<tr>
<td>In situ</td>
<td>CCN concentration ($N_{ccn}$) at 0.4% SS</td>
</tr>
</tbody>
</table>
How to Deal with the black box...

‘Garbage in, garbage out’ – computer science

Before machine learning:
- Remove data that has large uncertainties
  - Lidar
    - ✔ Negative lidar observables
    - ✔ Aerosol depolarization ratio greater than 1
  - In situ
    - ✔ CCN below 10 cm⁻³
    - ✔ ABS below 0.1 Mm⁻¹

Algorithm selection:
- Supervised regression learning problem with large number of numerical features.
- Fully-Connected Neural Network (FCNN) regression model

Architecture setup:
- Training data: 70%, Testing data: 30%
- 10-fold cross validation
- Hyperparameters are tuned iteratively during the training using Bayesian optimization

---

(https://xkcd.com/1838/)
ML-predicted CCN concentrations from HSRL-2 without and with reanalysis

Without reanalysis:
- $y = 0.88x + 75.08$
- $\text{MAE}_{\text{all}} = 128.45\%$
- $\text{MRE}_{\text{all}} = 20.39\%$
- $R = 0.92$\hspace{0.5cm} N = 2961$
- $N_{<20\%} = 2058$ (70%)$
- N_{<50\%} = 2508$ (85%)

With reanalysis:
- $y = 0.94x + 38.74$
- $\text{MAE}_{\text{all}} = 77.32\%$
- $\text{MRE}_{\text{all}} = 12.24\%$
- $R = 0.97$\hspace{0.5cm} N = 2961$
- $N_{<30\%} = 2510$ (85%)$
- N_{<50\%} = 2725$ (92%)
Test for model performance with incomplete in situ data

With reanalysis:

\[ y = 0.59x + 193.65 \]

\[ MAE_{\text{all}} = 148.10 \quad MRE_{\text{all}} = 23.56\% \]

\[ R = 0.83 \quad N = 2961 \]

\[ N_{\leq \pm 30\%} = 2282 \quad (77\%) \]

\[ N_{\leq \pm 50\%} = 2560 \quad (86\%) \]

Grey-shaded areas excluded in training

With reanalysis:

\[ y = 0.89x + 100.22 \]

\[ MAE_{\text{all}} = 137.73 \quad MRE_{\text{all}} = 21.86\% \]

\[ R = 0.92 \quad N = 2961 \]

\[ N_{\leq \pm 30\%} = 1966 \quad (66\%) \]

\[ N_{\leq \pm 50\%} = 2324 \quad (78\%) \]
CCN for HSRL-2, HSRL-1 and EarthCARE observables

\[ 3\beta + 2\alpha + 3\delta \]

Without reanalysis

- Density
- 1:1 line
- Linear fit
- ±30% error
- ±50% error

With reanalysis

- Density
- 1:1 line
- Linear fit
- ±30% error
- ±50% error

\[ 2\beta + 1\alpha + 2\delta \]

Without reanalysis

- Density
- 1:1 line
- Linear fit
- ±30% error
- ±50% error

With reanalysis

- Density
- 1:1 line
- Linear fit
- ±30% error
- ±50% error

\[ 1\beta + 1\alpha + 1\delta \]

Without reanalysis

- Density
- 1:1 line
- Linear fit
- ±30% error
- ±50% error

With reanalysis

- Density
- 1:1 line
- Linear fit
- ±30% error
- ±50% error
CCN predicted by ML-model for ER-2 flight across Southeast Atlantic - Aug. 26, 2016

- Tropical thin Cirrus
- Upper tropospheric aerosol
- Patchy altocumulus clouds
- Broken low clouds
- Biomass burning aerosol plume
## Mean Absolute (Relative) Error of CCN and ABS predictions for all and pristine conditions

<table>
<thead>
<tr>
<th>Predictor Data set →</th>
<th>HSRL-2 observables</th>
<th>HSRL-2 observables + Reanalysis Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor Indicator →</td>
<td>Mean Absolute Error (Relative)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full range</td>
<td>Pristine 0&lt;CCN&lt;100</td>
</tr>
<tr>
<td>CCN [1/cm³]</td>
<td>128.45 (20.39%)</td>
<td>101.11 (181.03%)</td>
</tr>
</tbody>
</table>
- CCN profile at lidar product grid when lidar observables are available.
We successfully trained machine learning models to perform CCN predictions using HSRL-2 lidar data from four airborne campaigns.

Our algorithm shows its capability to predict vertically-resolved CCN concentration with an approximate relative uncertainty of 20% with lidar observables, and uncertainty can be further reduced by adding reanalysis constraints.

For pristine conditions, CCN retrieval errors are much higher (partly due to sparse training data).

We expect that the machine learning-based algorithm can substantially enhance the campaign data product and benefit future satellite mission.
The computation in this paper was performed at the Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU).