Machine Learning towards a Global Parameterisation of Atmospheric New Particle Formation and Growth

Theodoros Christoudias
christoudias@cyi.ac.cy
Climate and Atmosphere Research Center
The Cyprus Institute
Nicosia, Cyprus

Mihalis A. Nicolaou
m.nicolaou@cyi.ac.cy
Computation-based Science and
Technology Research Center
The Cyprus Institute
Nicosia, Cyprus

Abstract

New particle formation (NPF) and growth in the atmosphere affects climate, weather, air quality, and human health. It is the first step of the complex process leading to cloud condensation nuclei (CCN) formation. Even though there is a wealth of observations from field measurements (in forests, high-altitude, polar regions, coastal and urban sites, aircraft campaigns), as well as laboratory studies of multi-component nucleation (including the CLOUD chamber at CERN), and improved nucleation theories, the NPF parameterisations in regional and global models are lacking. These deficiencies make the impacts of aerosols one of the highest sources of uncertainty in global climate change modelling, and associated impacts on weather and human health. We propose to use Machine Learning methods to overcome the challenges in modelling aerosol nucleation and growth, by ingesting the data from the multitude of available sources to create a single parameterisation applicable throughout the modelled atmosphere (troposphere and stratosphere at all latitudes) that efficiently encompasses all input ambient conditions and concentrations of relevant species.

1 Introduction

Clouds and aerosols continue to contribute the largest uncertainty to estimates and interpretations of the Earth's changing energy budget [26]. Atmospheric aerosols have direct and indirect effects on Earth's climate and impacts on public health [10]. Aerosols originate from several natural and anthropogenic sources. New particle formation (NPF), the gas-to-particle conversion of atmospheric vapours, is a major source of secondary aerosols that can act as cloud condensation nuclei (CCN) and further affect the climate [25]. NPF has been observed in boreal forests, coastal, agricultural, and urban areas, including polluted megacities [16, 17, 15, 6]. NPF in the atmosphere profoundly affects climate, weather, air quality, and human health. It is the first step of the complex process leading to cloud condensation nuclei (CCN) formation. It is estimated that about 40–70% of CCN globally originate from NPF [20, 30].

Even though there is a wealth of observations from field measurements (in forests, high-altitude, polar regions, coastal and urban sites, aircraft campaigns) [2, 9, 1, 3], as well as laboratory studies of multi-component nucleation (including the CLOUD chamber at CERN) [11, 12, 19], and improved nucleation theories, there is still limited understanding and the NPF parameterisations in regional and global models of the atmosphere are lacking. These deficiencies make the impacts of aerosols the highest source of uncertainty in modelling global climate change (IPCC), and associated impacts on weather and human health. Understanding NPF is imperative to reduce uncertainties in climate projections and to tackle urban air quality problems.

Tackling Climate Change with Machine Learning workshop at NeurIPS 2020.

In large-scale models of the atmosphere, which consider aerosol dynamics to project climate change, it is necessary to use parameterised nucleation rates for computational efficiency. Presently, atmospheric models rely on simple parameterisations that are typically polynomial fits to the dependency of measured NPF rates as a function of vapour concentration (and airborne ions) and are only valid for the environments and conditions that match each observation site [4]. To date, measurements contribute to disparate parameterisations with limited spatial applicability, even though they relate to the same physical processes. More detailed theoretical models including additional species (e.g. nano-Köhler theory [14]) were also shown to have limitations and are too computationally expensive to be incorporated into high-resolution models of the atmosphere and climate. A consistent parameterisation in the scope of atmospheric modelling, with predictive capacity and computational efficiency, has so far proven to be elusive. The introduction of machine learning methods in this field is limited to using random forest regression of atmospheric model data to a-posteriori derive measured CCN [22], and automating the manual process of observed event identification based *only* on particle size distributions [8], with no inference or additional insights.

We propose to instead use state-of-the-art data science methods to combine all available measurements of NPF and growth, and associated (vapour concentrations, presence of ions, meteorological conditions) from the multitide of souces (Sec. 2) into a single hyper-parameter model that can provide consistent (at all altitudes and environments/latitudes) and efficient (once trained, the computational load of inference permits to be incorporated in GCMs) NPF simulation. Machine Learning methods can overcome the challenges in simulating aerosol nucleation and growth by ingesting the data from the multitude of available sources to create a single parameterisation applicable throughout the atmosphere (troposphere and stratosphere) that efficiently encompasses all ambient conditions and concentrations of relevant species. At the same time, we can gain insight into the physical processes undeprinning NPF and growth and the complex interaction of the different components. Such insights can drive further research, e.g., by guiding future observation campaigns.

2 Data Aggregation

Measurement campaigns of NPF and growth collocated with ambient conditions measurements include in situ ground station, tower, and aircraft observations. Additional multi-component nucleation measurement data are available from chamber experiments: **(i)** In situ condensation particle counters (CPCs) for 22 ground station locations from the EBAS [28] database over the years 1972–2009. **(ii)** IAGOS CARIBIC detailed and extensive measurements during long distance flights deploying a modified airfreight container with automated scientific apparatus. Using a passenger Airbus A340-600 from Lufthansa in total more than 550 flights are succesfully completed. **(iii)** The NASA Atmospheric Tomography Missions (ATom) deploying an extensive gas and aerosol payload on the NASA DC-8 aircraft for systematic, global-scale sampling of the atmosphere, profiling continuously from 0.2 to 12 km altitude in all 4 seasons over a 4-year period. **(iv)** The Aerosol, Cloud, Precipitation, and Radiation Interactions and Dynamics of Convective Cloud Systems (ACRIDICON) dataset by the DLR High Altitude and Long Range Aircraft (HALO). **(v)** Chamber measurements, in particular the CERN CLOUD experiment.

3 Machine Learning Application

In the following, we propose two pathways for globally analysing the disparate sources of data described in Sec. 2. While not necessarily disjoint, the first is focused on obtaining *accurate* models that can be integrated in general circulation models (GCMs) in a black-box fashion, facilitating future climate predictions and reducing uncertainty. In contrast, the second is oriented towards discovering insights of the underlying physical process that can drive future observation campaigns and enhance the understanding of the process by domain experts.

Global Parameterisation for NPF from ambient conditions and species concentrations. Current approaches result in disparate NPF parameterisations with specific applicability in terms of initial conditions [4]. In contrast, we propose the adoption of machine learning approaches to create a hyperparametrization of new particle formation and growth applicable throughout the atmosphere (from lower troposphere to higher levels of the stratosphere), that is accurate *outside* the current limits of the input data span of current parametrizations. The speedy test-time evaluation of machine learning models can lead to computationally efficient techniques that can be readily integrated in

GCMs for multi-decadal simulations. To do so, data will be ingested from multiple data sources (Sec. 2), with each source covering specific portions of the problem phase space, whilst describing the same underlying physical process.

The above problem can be considered in the context of supervised learning, for example by considering a classification problem with NPF being a binary target variable, or a regression problem with a goal of predicting particle formation and growth. Deep learning methods [18] for supervised learning have proven successful in a multitude of scientific and engineering tasks, and can be therefore evaluated in this context. Such pre-trained models (e.g., convolutional / recurrent neural network architectures) can then be incorporated straight-forwardly into GCMs. In the context of transfer learning [24, 27], domain adaptation techniques (e.g., fine-tuning) can be utilized to transfer knowledge between various data domains, that is to compensate for covariate shifts between measurements in different conditions (e.g., between chamber experiments and in situ ground station measurements). It is noted that while accuracy is the primary metric to optimize for in this setting, interpretability and explainability is crucial for developing further intuition into the physical processeses underlying NPF. To this end, methods that attempt to explain the decision making carried out by Deep Networks can be employed [21, 5]. However, interpretability is still very much an open problem in deep learning, and alternative approaches can be explored as discussed in what follows.

Discovering insights to better process understanding. While building accurate global machine learning models is crucial for reducing the uncertainty in climate projections, discovering further insights regarding the new particle formation process can improve the understanding of the process, and can further guide future observation campaigns. In this light, we propose to help domain scientists by identifying interactions (*interdependencies*) between nucleating species (e.g., H2SO4, volatile organic compounds, HOMs) and ambient conditions (e.g., temperature, humidity), to gain further insight into the new particle formation process, and disentangle the factors that contribute to particle formation. Methods based on tensor decomposition [13] can be evaluated either in an unsupervised manner, to capture the principal modes of variation of the data along with the structure of the underlying interactions in a core tensor, as well as in a supervised learning context [23], where high-order interactions and multiplicative interactions [7] can be captured quickly, often in linear time. Such techniques consist of multi-linear generalizations of linear algorithms, and therefore preserve interpretability while increasing the model capacity. Finally, tree-based methods such as random forests, and the recently proposed Neural-backed Decision trees [29], have been shown to provide similar performance to state-of-the-art deep learning methods, while providing a detailed breakdown of the decision rules that lead to a predictive result. This family of methods can provide detailed insights on the interdependencies of species concentrations as well as the effect of ambient conditions on phenomena related to NPF.

4 Conclusions

Using Machine Learning methods can overcome the long-standing challenges in understanding and simulating aerosol nucleation and growth by ingesting the data from diverse sources into a unified, global, multi-component parameterisation, valid throughout the atmosphere. This in turn will decrease the largest uncertainty in climate projections and provide a tool to effectively tackle air quality problems caused by urbanisation and population growth.

References

[1] Meinrat O Andreae, Otávio C Acevedo, A Araùjo, Paulo Artaxo, Cybelli GG Barbosa, HMJ Barbosa, J Brito, Samara Carbone, Xuguang Chi, BBL Cintra, et al. The Amazon Tall Tower Observatory (ATTO): overview of pilot measurements on ecosystem ecology, meteorology, trace gases, and aerosols. *Atmospheric Chemistry and Physics*, 15(18):10723–10776, 2015.

[2] CAM Brenninkmeijer, P Crutzen, F Boumard, T Dauer, B Dix, R Ebinghaus, D Filippi, H Fischer, H Franke, U Frieß, et al. Civil aircraft for the regular investigation of the atmosphere based on an instrumented container: The new caribic system. *Atmospheric Chemistry and Physics*, 7(18):4953–4976, 2007.

[3] Miikka Dal Maso, Markku Kulmala, Ilona Riipinen, Robert Wagner, Tareq Hussein, Pasi P Aalto, and Kari EJ Lehtinen. Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland. *Boreal environment research*, 10(5):323, 2005.

[4] Eimear M Dunne, Hamish Gordon, Andreas Kürten, João Almeida, Jonathan Duplissy, Christina Williamson, Ismael K Ortega, Kirsty J Pringle, Alexey Adamov, Urs Baltensperger, et al. Global atmospheric particle formation from cern cloud measurements. *Science*, 354(6316):1119–1124, 2016.

[5] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[6] Song Guo, Min Hu, Misti L Zamora, Jianfei Peng, Dongjie Shang, Jing Zheng, Zhuofei Du, Zhijun Wu, Min Shao, Limin Zeng, et al. Elucidating severe urban haze formation in China. *Proceedings of the National Academy of Sciences*, 111(49):17373–17378, 2014.

[7] Siddhant M Jayakumar, Wojciech M Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2019.

[8] Jorma Joutsensaari, Matthew Ozon, Tuomo Nieminen, Santtu Mikkonen, Timo Lähivaara, Stefano Decesari, M Cristina Facchini, Ari Laaksonen, and Kari EJ Lehtinen. Identification of new particle formation events with deep learning. 2018.

[9] Astrid Kiendler-Scharr, Jürgen Wildt, Miikka Dal Maso, Thorsten Hohaus, Einhard Kleist, Thomas F Mentel, Ralf Tillmann, Ricarda Uerlings, Uli Schurr, and Andreas Wahner. New particle formation in forests inhibited by isoprene emissions. *Nature*, 461(7262):381–384, 2009.

[10] Ki-Hyun Kim, Ehsanul Kabir, and Shamin Kabir. A review on the human health impact of airborne particulate matter. *Environment international*, 74:136–143, 2015.

[11] Jasper Kirkby, Joachim Curtius, João Almeida, Eimear Dunne, Jonathan Duplissy, Sebastian Ehrhart, Alessandro Franchin, Stéphanie Gagné, Luisa Ickes, Andreas Kürten, et al. Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature*, 476(7361):429–433, 2011.

[12] Jasper Kirkby, Jonathan Duplissy, Kamalika Sengupta, Carla Frege, Hamish Gordon, Christina Williamson, Martin Heinritzi, Mario Simon, Chao Yan, João Almeida, et al. Ion-induced nucleation of pure biogenic particles. *Nature*, 533(7604):521–526, 2016.

[13] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[14] Jenni Kontkanen, Tinja Olenius, Markku Kulmala, Ilona Riipinen, et al. Exploring the potential of nano-köhler theory to describe the growth of atmospheric molecular clusters by organic vapors using cluster kinetics simulations. *Atmospheric Chemistry and Physics*, 2018.

[15] M Kulmala, V-M Kerminen, T Petäjä, AJ Ding, and L Wang. Atmospheric gas-to-particle conversion: why NPF events are observed in megacities? *Faraday discussions*, 200:271–288, 2017.

[16] M Kulmala, M Dal Maso, JM Mäkelä, L Pirjola, M Väkevä, P Aalto, P Miikkulainen, K Hämeri, and CD O'dowd. On the formation, growth and composition of nucleation mode particles. *Tellus B*, 53(4):479–490, 2001.

[17] Markku Kulmala, Jenni Kontkanen, Heikki Junninen, Katrianne Lehtipalo, Hanna E Manninen, Tuomo Nieminen, Tuukka Petäjä, Mikko Sipilä, Siegfried Schobesberger, Pekka Rantala, et al. Direct observations of atmospheric aerosol nucleation. *Science*, 339(6122):943–946, 2013.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[19] Katrianne Lehtipalo, Chao Yan, Lubna Dada, Federico Bianchi, Mao Xiao, Robert Wagner, Dominik Stolzenburg, Lauri R Ahonen, Antonio Amorim, Andrea Baccarini, et al. Multicomponent new particle formation from sulfuric acid, ammonia, and biogenic vapors. *Science advances*, 4(12):eaau5363, 2018.

[20] J. Merikanto, D. V. Spracklen, G. W. Mann, S. J. Pickering, and K. S. Carslaw. Impact of nucleation on global CCN. *Atmospheric Chemistry and Physics*, 9(21):8601–8616, 2009.

[21] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[22] A. A. Nair and F. Yu. Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements. *Atmospheric Chemistry and Physics*, 20(21):12853–12869, 2020.

[23] Alexander Novikov, Mikhail Trofimov, and Ivan V. Oseledets. Exponential machines. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[24] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[25] John H Seinfeld and Spyros N Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016.

[26] Thomas F Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda Tignor, Simon K Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex, Pauline M Midgley, et al. Climate change 2013: The physical science basis. *Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*, 1535, 2013.

[27] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[28] Tørseth, Kjetil and Aas, Wenche and Breivik, Knut and Fjæraa, Ann Mari and Fiebig, Markus and Hjellbrekke, Anne-Gunn and Lund Myhre, C and Solberg, Sverre and Yttri, Karl Espen. Introduction to the European Monitoring and Evaluation Programme (EMEP) and observed atmospheric composition change during 1972–2009. *Atmospheric Chemistry and Physics*, 12(12):5447–5481, 2012.

[29] Alvin Wan, Lisa Dunlap, Daniel Ho, Jihan Yin, Scott Lee, Henry Jin, Suzanne Petryk, Sarah Adel Bargal, and Joseph E Gonzalez. NBDT: Neural-Backed Decision Trees. *arXiv preprint arXiv:2004.00221*, 2020.

[30] Fangqun Yu and Gan Luo. Simulation of particle size distribution with a global aerosol model: contribution of nucleation to aerosol and CCN number concentrations. *Atmospheric Chemistry & Physics Discussions*, 9(2), 2009.