

Machine Learning-Based Prediction of New Particle Formation Events from Terpene Chemistry and Anthropogenic Emission Interactions

Darren J. Woods

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
darrenwoods@buffalo.edu

Edwin Marsh

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
edwin.marsh948@uab.edu

Fengzhu Du

Department of Computer Science, George Mason University, Fairfax, VA, USA.
du660@gmu.edu

Abstract

Atmospheric new particle formation is a central process connecting gas-phase chemistry, aerosol dynamics, cloud condensation nuclei abundance, and regional climate forcing. In environments influenced by both biogenic terpene emissions and anthropogenic pollutants, the formation of new particles emerges from complex chemical and physical interactions that are poorly captured by traditional deterministic parameterizations. Recent observational evidence has highlighted the importance of heavy terpenes and their modulation by anthropogenic emissions in remote high-altitude regions, yet operational prediction of such events remains an unresolved scientific and computational challenge. This article examines the design of machine learning-based prediction systems for new particle formation events with a focus on system architecture, data infrastructure, feature engineering, model selection, deployment, robustness, fairness, and policy implications. We present a system-level analysis that situates machine learning within broader atmospheric observation and modeling infrastructures, emphasizing the structural trade-offs between interpretability, data efficiency, generalization, and operational resilience. The discussion addresses the integration of ground-based aerosol and precursor observations, laboratory chamber data, satellite retrievals, emission inventories, and reanalysis fields, and considers how these heterogeneous data streams can be organized into coherent learning environments. The paper further evaluates the governance requirements for predictive systems that may inform air quality management and environmental policy, particularly across regions with uneven observational capacity. Rather than proposing a single algorithmic solution, we argue that reliable machine learning prediction for new particle formation requires a socio-technical infrastructure capable of sustaining data quality, model monitoring, uncertainty communication, and equitable scientific benefit.

Keywords

new particle formation; terpene chemistry; anthropogenic emissions; machine learning; atmospheric data infrastructure; model robustness; environmental governance; socio-technical systems.

1. Introduction

Atmospheric aerosol particles exert a substantial influence on the Earth system through aerosol-cloud interactions and direct radiative effects, yet their formation and evolution remain among the largest uncertainties in climate assessment. New particle formation from gaseous precursors is a particularly important mechanism because it can produce a large fraction of cloud condensation nuclei at regional and global scales. Understanding the formation and growth rates of ultrafine atmospheric particles has therefore been a long-standing research priority, with field observations and laboratory experiments providing the empirical foundation for current theoretical frameworks [1]. The scientific challenge is not merely to characterize individual chemical pathways but to predict when and where nucleation will occur under dynamically changing atmospheric conditions. This prediction problem is deeply structural because it links molecular-scale chemistry, mesoscale transport, land surface emissions, and anthropogenic activity patterns.

The chemistry underlying new particle formation is broadly understood to involve sulfuric acid, ammonia, amines, and a spectrum of oxidized organic compounds. However, the relative importance of these precursors varies across environments and time. For example, measurements in the free troposphere have demonstrated that chemical composition and timing can jointly determine whether new particles are formed, emphasizing the need for models that respect temporal and spatial context [2]. In biogenic environments, oxidation of volatile organic compounds can produce low-volatility secondary organic species that participate directly in particle formation and growth [3]. Chamber and field studies have further shown that oxidation products of biogenic emissions can contribute to atmospheric nucleation, sometimes in the presence of very low sulfuric acid concentrations [4]. These findings complicate the conventional paradigm that sulfuric acid is the dominant controlling species everywhere, and they motivate the development of predictive approaches capable of representing multiple precursor families simultaneously.

Recent field evidence from the southeastern Qinghai-Tibet Plateau has further shown that heavy terpenes and their interactions with anthropogenic emissions can play a pivotal role in initiating particle formation under high-altitude conditions [5]. This finding is significant because it demonstrates that even remote and relatively pristine regions may experience new particle formation events shaped by transported or locally emitted anthropogenic species interacting with abundant biogenic terpene chemistry. Such interactions are not easily captured by current regional or global aerosol models, which often rely on simplified organic nucleation parameterizations and coarse emission inventories. The observational evidence therefore creates both an opportunity and an obligation for the atmospheric science community to develop learning-based systems that can ingest high-dimensional precursor and environmental data and produce reliable event-level predictions. This article addresses that challenge from a systems perspective, focusing on the architectural, infrastructural, and governance dimensions of machine learning-based prediction.

2. Background and Related Work

Mechanistic research over the past two decades has substantially advanced the conceptual understanding of atmospheric nucleation. Laboratory experiments have demonstrated that

multicomponent nucleation involving sulfuric acid, ammonia, and biogenic vapors can explain many observed formation events, but the exact contribution of each precursor remains highly variable [6]. Ion-induced nucleation has also been shown to produce pure biogenic particles under controlled conditions, indicating that charged clusters and organic molecules can sustain particle formation even in the absence of strong inorganic acid sources [7]. A key development has been the recognition that oxidation of monoterpenes and sesquiterpenes can produce extremely low volatile organic compounds, which condense onto nascent clusters and stabilize their growth [8]. These low-volatility compounds appear particularly important during the early growth phase, bridging the gap between molecular clusters and larger particles that can act as cloud condensation nuclei [9]. Comprehensive reviews of nucleation and growth have further documented the interactions between inorganic acids, bases, and organic vapors, emphasizing that the boundary between nucleation and growth is not sharp in real atmospheric systems [10].

Within this mechanistic landscape, anthropogenic emissions can either suppress or enhance new particle formation. Sulfur dioxide, nitrogen oxides, ammonia, and aromatic hydrocarbons alter the production rates of key nucleating and condensing species, while primary particles can increase the condensation sink and thereby inhibit nucleation. The net effect depends on the relative timing of emission changes, oxidant availability, temperature, relative humidity, and existing aerosol surface area. In many coastal and continental regions, new particle formation is most probable when air masses combine biogenic organic precursors with modest levels of anthropogenic sulfur or nitrogen species. The Qinghai-Tibet Plateau example is particularly instructive because it reveals that heavy terpenes may interact with transported anthropogenic emissions to produce nucleation under conditions that were previously considered unfavorable [5].

Machine learning has emerged as a complementary strategy to mechanistic modeling because it can identify nonlinear and nonstationary relationships in observational data without requiring an explicit representation of every chemical pathway. Deep learning methods have been applied to a range of Earth system science problems, often revealing emergent structures that were not anticipated by existing theory [11]. However, the application of machine learning to atmospheric chemistry is constrained by data sparsity, measurement heterogeneity, and the need to maintain physical consistency. In the case of new particle formation, the target variable is often an event classification derived from particle size distribution measurements, while predictors include gas-phase concentrations, meteorological fields, and markers of emission influences. The construction of a robust learning system therefore requires not only algorithmic sophistication but also careful attention to data provenance, harmonization, and uncertainty.

3. System Architecture and Data Infrastructure

A credible machine learning system for new particle formation prediction must be embedded within a broader atmospheric data infrastructure. Historically, atmospheric composition data have been collected and curated by separate research groups, monitoring networks, and satellite agencies, resulting in fragmented and heterogeneous archives. The emergence of benchmark datasets tailored for air quality machine learning has helped to standardize access to ground-level observations and model outputs, but aerosol-specific benchmarks remain comparatively underdeveloped [12]. Similarly, large databases of surface ozone and related reactive gases have supported reproducible machine learning studies, yet they do not capture the full complexity of aerosol formation processes [13]. To address this gap, proposals for a

global Earth observatory have emphasized the need for coordinated atmospheric measurement systems, open data policies, and shared cyberinfrastructure [14]. Within such an observatory, new particle formation prediction could operate as a service that consumes real-time observations from field stations, chamber facilities, and satellite platforms.

The architectural challenge is to design a system that can integrate data streams with very different temporal resolutions, spatial scales, and uncertainty structures. Ground-based aerosol instruments such as scanning mobility particle sizers provide size distribution measurements at high time resolution, while gas-phase chemical ionization mass spectrometers detect precursor molecules at trace levels. These observations are often available only at a small number of research-grade stations. Satellite retrievals offer broader spatial coverage but are limited in their ability to resolve near-surface precursor concentrations and the smallest particle sizes. Atmospheric reanalysis products provide continuous meteorological context but do not directly represent the chemical complexity of terpene oxidation. Consequently, a layered architecture is required, in which raw observations are harmonized into a common data model, quality-controlled, gap-filled, and then transformed into features suitable for machine learning.

This architecture must also address the tension between centralized and federated data management. Centralized repositories simplify model training and reproducibility, but they require sustained funding and institutional coordination. Federated approaches allow individual research groups to retain custody of their data while sharing learned model updates or derived features, which may be more acceptable in international collaborations. However, federated learning introduces additional complexity in ensuring statistical representativeness and preventing systematic biases. Hybrid models in which high-level data products are centralized while raw instrument data remain distributed may offer a practical compromise. Regardless of the specific design, the infrastructure must support metadata standards that describe instrument type, calibration history, detection limits, and data quality flags, because these metadata are essential for interpreting model outputs and diagnosing performance degradation.

4. Feature Engineering for Terpene Chemistry and Anthropogenic Emission Interactions

Feature engineering is the critical bridge between raw atmospheric measurements and machine learning algorithms. For new particle formation prediction, the predictor space should include variables that reflect the availability of nucleating and condensing species, the thermodynamic state of the atmosphere, the preexisting aerosol sink, and the history of the air mass. Biogenic emission models such as those describing global terrestrial isoprene and monoterpene fluxes can provide spatially explicit estimates of terpene precursor availability under varying light and temperature conditions [15]. However, the actual atmospheric concentrations of monoterpenes, sesquiterpenes, and their oxidation products depend not only on emission rates but also on transport, dilution, and oxidative processing. Features derived from volatility basis set representations can help encode the transformation of organic molecules across volatility classes, linking precursor chemistry to condensable vapors and secondary organic aerosol mass [16]. Such representations reduce the dimensionality of the organic mixture while preserving essential physicochemical information about volatility and oxidation state.

Anthropogenic influence can be represented through features describing sulfur dioxide, nitrogen oxides, ammonia, particulate surface area, and tracers of combustion or industrial

activity. Because the interaction between terpene chemistry and anthropogenic emissions is central to recent field findings, machine learning models should be provided with features that explicitly capture the co-occurrence and temporal alignment of these influences. For example, lagged sulfur dioxide and monoterpene concentrations can reveal whether nucleation tends to follow periods of increased oxidant availability after emission pulses. Diurnal, seasonal, and synoptic-scale features are also important, because new particle formation often exhibits strong temporal regularities related to solar radiation, boundary layer dynamics, and vegetation activity. Temporal aggregation over multiple time windows allows the model to learn characteristic precursor accumulation times without requiring a prior specification of the relevant timescales.

Feature selection and dimensionality reduction must be approached with care because atmospheric datasets often contain highly correlated predictors. Tree-based methods naturally handle nonlinear interactions and correlated inputs, while neural networks may require more extensive preprocessing to avoid overfitting. The interpretability of the resulting model depends heavily on the features chosen. If the system is intended to support scientific discovery, features should retain clear physical meaning and be traceable to original measurements. If the primary goal is operational event prediction, then abstract learned representations may be acceptable provided that their outputs are calibrated and validated. In both cases, feature engineering should be treated as a dynamic process that evolves as new instruments, emission inventories, and chemical mechanisms become available.

5. Machine Learning Modeling Approaches and Structural Trade-offs

Several families of machine learning models are applicable to new particle formation prediction, and each entails distinct structural trade-offs. Gradient-boosted decision trees offer strong performance on tabular datasets, require relatively little hyperparameter tuning, and can provide feature importance measures that aid scientific interpretation. However, they may struggle to exploit long-range temporal dependencies unless explicit lagged features are carefully constructed. Deep neural networks, including recurrent and temporal convolutional architectures, can learn sequential patterns from time series data, but they demand larger labeled datasets and more computational resources. Attention-based models have shown promise in Earth system applications by allowing the model to weight different time steps and sensor channels according to their relevance, yet they can be difficult to interpret. Hybrid models that combine mechanistic constraints with learned components can preserve physical consistency, but they require careful design to avoid imposing overly restrictive assumptions.

Interpretable machine learning is especially important in atmospheric science because model outputs may inform field campaigns, emission policy, or public health interventions. The meteorological community has developed a range of interpretability tools that can be adapted to aerosol prediction, including partial dependence analysis, Shapley values, and case-based reasoning [17]. These techniques allow researchers to examine how a model responds to changes in terpene concentration, anthropogenic tracer levels, or condensation sink, thereby building confidence in the learned relationships. At the same time, interpretability can conflict with predictive accuracy, since highly expressive models may learn interactions that are difficult to summarize in simple terms. A system-level approach requires deciding in advance which trade-offs are acceptable for a given deployment context.

The selection of a machine learning model must also consider the structure of the prediction problem itself. New particle formation events can be formulated as binary classification, multiclass event intensity prediction, or regression of particle formation and growth rates.

Classification approaches are appealing for operational early warning systems because they produce discrete decisions, but they require a defensible threshold for defining an event. Regression approaches retain more information but may be dominated by measurement noise during non-event periods. In addition, spatial and temporal cross-validation must be designed to prevent data leakage from adjacent time steps or from the same measurement site appearing in both training and test partitions. The performance of operational systems can degrade substantially when models are transferred to regions with different emission regimes, vegetation types, or meteorological conditions, and this generalization gap remains a central challenge for machine learning in weather and climate applications [18].

6. Deployment, Robustness and Monitoring

Deployment of machine learning-based new particle formation prediction systems requires a robust operational framework that extends well beyond model training. In an operational context, the system must tolerate missing sensor data, instrument drift, sudden changes in emission patterns, and the appearance of novel atmospheric conditions not represented in the training record. Continuous monitoring of model performance is essential, but label availability may be delayed because event identification depends on quality-controlled aerosol size distribution measurements. A practical deployment strategy therefore separates a real-time prediction layer from a delayed evaluation layer. The real-time layer can issue warnings using available predictors, while the evaluation layer periodically assesses model skill after new observations have been validated. This separation also enables retraining cycles that incorporate recent data and prevent silent model degradation.

The computational and organizational sustainability of such systems is a key concern. Machine learning for weather and climate applications has highlighted the challenges of maintaining long-term data pipelines, managing model versioning, and ensuring reproducibility across platforms [19]. For aerosol prediction, these challenges are amplified by the scarcity of reference-grade observations and the high cost of maintaining advanced mass spectrometers and particle sizing instruments. The environmental footprint of model training and inference should also be considered, particularly if large ensembles or retraining schedules are adopted. Lightweight models with careful feature selection may be more appropriate for regional operational centers with limited computing capacity, while centralized research systems can afford more compute-intensive experiments. Sustainability therefore depends on aligning model complexity with the available cyberinfrastructure and long-term funding.

A notable case illustration is the deployment of continuous monitoring in high-altitude or remote regions, where instrument maintenance is logistically difficult and data transmission may be intermittent. In such settings, the machine learning system must be capable of graceful degradation, producing probabilistic outputs with clearly communicated uncertainty rather than categorical predictions that may be misleading when input data are incomplete. Ensemble approaches can partially address this by combining models trained on different subsets of predictors, thereby providing predictions even when some sensor channels fail. However, ensembles increase computational cost and may obscure the underlying failure modes. An alternative is to design modular architectures in which separate submodels handle precursor chemistry, meteorology, and aerosol state, with a fusion layer that propagates uncertainty. Such modularity can improve robustness and interpretability while maintaining operational flexibility.

7. Fairness, Governance and Policy Implications

The deployment of machine learning-based environmental prediction systems raises important questions of fairness and equity. Training data for new particle formation are heavily concentrated in well-instrumented regions, primarily in Europe, North America, East Asia, and a limited number of high-altitude research sites. As a result, models may perform well in these regions while systematically underestimating or misclassifying events in underrepresented environments, including tropical forests, arid regions, the Arctic, and parts of the Global South. If such models are used to guide air quality management or climate policy, their uneven performance could reinforce existing scientific and infrastructural disparities. Fairness in this context is not simply about equal aggregate accuracy across regions, but about ensuring that the benefits and risks of predictive environmental technology are distributed equitably and that affected communities have access to locally relevant information.

Governance mechanisms are needed to ensure that machine learning predictions for atmospheric processes are transparent, accountable, and scientifically defensible. Satellite monitoring has already demonstrated the value of open data for air quality and health applications, and similar openness should be extended to aerosol prediction systems [20]. Model documentation should describe the training data, feature definitions, performance metrics, uncertainty estimates, and known limitations. Regulatory bodies may require that machine learning outputs used for environmental decision support undergo periodic review and independent validation. In addition, data governance must respect the rights and interests of countries and communities that host monitoring infrastructure, particularly when data are used to train models that may later be commercialized or embedded in international policy frameworks.

Policy implications extend to emission control strategies and climate intervention pathways. If machine learning models reliably identify combinations of biogenic and anthropogenic conditions that lead to strong new particle formation, they could inform short-term measures to reduce precursor emissions or to issue public health advisories about ultrafine particle exposure. However, uncertainty in model predictions must be communicated carefully to avoid false confidence. Decision-makers need not only event probabilities but also estimates of the consequences of different actions, including co-benefits and unintended side effects. This requires close integration between machine learning researchers, atmospheric chemists, exposure scientists, and policy analysts. Ultimately, the value of a machine learning-based prediction system lies not in replacing mechanistic understanding but in organizing diverse data streams into timely, interpretable, and equitable knowledge products that can support more responsive environmental governance.

8. Conclusion

Machine learning-based prediction of new particle formation events from terpene chemistry and anthropogenic emission interactions represents a promising but demanding interdisciplinary frontier. The scientific problem is governed by nonlinear chemical pathways, highly variable emission sources, and atmosphere-biosphere coupling that differs across regions and seasons. Recent evidence from high-altitude environments has underscored the importance of heavy terpenes and their interactions with anthropogenic pollutants, challenging the assumption that remote regions can be treated as chemically simple. A reliable prediction system must integrate ground-based observations, chamber experiments, satellite retrievals, emission inventories, and reanalysis products within a coherent data architecture. It must also balance the competing demands of accuracy, interpretability,

robustness, and equity. This article has argued that the central challenges are not purely algorithmic but are distributed across data infrastructure, model governance, deployment monitoring, and policy integration. Future progress will require sustained investment in reference-grade atmospheric measurements, open and interoperable data systems, and collaborative frameworks that link machine learning experts with atmospheric scientists and decision-makers. Only through such an integrated systems approach can the potential of machine learning be realized for understanding and anticipating one of the most complex processes in the atmospheric aerosol system.

References

1. Kulmala, M., Vehkamäki, H., Petäjä, T., Dal Maso, M., Lauri, A., Kerminen, V.-M., Birmili, W., & McMurry, P. H. (2004). Formation and growth rates of ultrafine atmospheric particles: A review of observations. *Journal of Aerosol Science*, 35(2), 143-176.
2. Bianchi, F., Tröstl, J., Junninen, H., Frege, C., Henne, S., Hoyle, C. R., ... Baltensperger, U. (2016). New particle formation in the free troposphere: A question of chemistry and timing. *Science*, 352(6289), 1109-1112.
3. Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., ... Mentel, T. F. (2014). A large source of low-volatility secondary organic aerosol. *Nature*, 506(7489), 476-479.
4. Riccobono, F., Schobesberger, S., Scott, C. E., Dommen, J., Ortega, I. K., Rondo, L., ... Baltensperger, U. (2014). Oxidation products of biogenic emissions contribute to nucleation of atmospheric particles. *Science*, 344(6185), 717-721.
5. Liu, Y., Nie, W., Qi, X., Li, Y., Xu, T., Liu, C., ... & Ding, A. (2024). The pivotal role of heavy terpenes and anthropogenic interactions in new particle formation on the southeastern Qinghai-Tibet Plateau. *Environmental Science & Technology*, 58(44), 19748.
6. Lehtipalo, K., Yan, C., Dada, L., Bianchi, F., Xiao, M., Wagner, R., ... Kulmala, M. (2018). Multicomponent new particle formation from sulfuric acid, ammonia, and biogenic vapors. *Science Advances*, 4(12), eaau5363.
7. Kirkby, J., Duplissy, J., Sengupta, K., Frege, C., Gordon, H., Williamson, C., ... Curtius, J. (2016). Ion-induced nucleation of pure biogenic particles. *Nature*, 533(7604), 521-526.
8. Jokinen, T., Berndt, T., Makkonen, R., Kerminen, V.-M., Junninen, H., Paasonen, P., ... Sipilä, M. (2015). Production of extremely low volatile organic compounds from biogenic emissions: Measured yields and atmospheric implications. *Proceedings of the National Academy of Sciences*, 112(23), 7123-7128.
9. Tröstl, J., Chuang, W. K., Gordon, H., Heinritzi, M., Yan, C., Molteni, U., ... Baltensperger, U. (2016). The role of low-volatility organic compounds in initial particle growth in the atmosphere. *Nature*, 533(7604), 527-531.
10. Zhang, R., Khalizov, A., Wang, L., Hu, M., & Xu, W. (2012). Nucleation and growth of nanoparticles in the atmosphere. *Chemical Reviews*, 112(3), 1957-2011.
11. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195-204.

12. Betancourt, C., Stomberg, T. T., Stadtler, S., Schultz, M. G., Roscher, R., & Wiedinmyer, C. (2021). AQ-Bench: A benchmark dataset for machine learning on global air quality. *Earth System Science Data*, 13(6), 3013-3040.
13. Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., ... Ziemke, J. R. (2017). Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations. *Elementa: Science of the Anthropocene*, 5, 58.
14. Kulmala, M. (2018). Build a global Earth observatory. *Nature*, 553(7686), 21-23.
15. Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., & Geron, C. (2006). Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). *Atmospheric Chemistry and Physics*, 6(11), 3181-3210.
16. Donahue, N. M., Kroll, J. H., Pandis, S. N., & Robinson, A. L. (2012). A two-dimensional volatility basis set—Part 2: Diagnostics of organic-aerosol evolution. *Atmospheric Chemistry and Physics*, 12(2), 615-634.
17. McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., ... Williams, J. K. (2017). Using interpretable machine learning in meteorology. *Bulletin of the American Meteorological Society*, 98(10), 2115-2128.
18. Dueben, P.-D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10), 3999-4009.
19. Chantry, M., Christensen, H., Dueben, P., & Palmer, T. (2021). Opportunities and challenges for machine learning in weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200083.
20. Holloway, T., Daigle, H., Krotkov, N. A., Strell, E., Follette-Cook, M., Wiedinmyer, C., ... Martin, R. V. (2020). Satellite monitoring for air quality and health. *Annual Review of Environment and Resources*, 45, 327-359.