

Review

Multi-Model Ensemble Approaches in Air Quality Prediction: A Comprehensive Review from Chemical Transport Models to Hybrid Machine Learning

Elena Chianese  and Angelo Riccio * 

Department of Science and Technology, Parthenope University of Naples, Centro Direzionale, Isola C4, 80143 Naples, Italy; elena.chianese@uniparthenope.it

* Correspondence: angelo.riccio@uniparthenope.it; Tel.: +39-081-547-6613

Abstract

Over the past two decades, air-quality prediction has moved from a mainly single-model paradigm toward ensemble systems that make explicit use of diversity across models, observations, and data streams. This review connects developments that are often treated separately: chemical transport model (CTM) ensembles, tree-based and hybrid machine learning ensembles, deep learning architectures, physics-informed neural networks, and distributed approaches such as federated learning. Evidence summarized from recent systematic reviews and coordinated modeling initiatives indicates that, within comparable validation settings, ensembles often outperform individual models for PM_{2.5}, PM₁₀, O₃, NO₂, CO, and SO₂ across a broad range of spatial scales and standard error metrics, including RMSE, MAE, and correlation. Operational CTM ensembles, such as the Copernicus Atmosphere Monitoring Service (CAMS) European system with eleven regional models, improve both forecast skill and uncertainty characterization for ozone and particulate matter. In data-driven applications, tree-based ensembles (Random Forest, gradient boosting, XGBoost, LightGBM) and hybrid deep architectures (CNN–LSTM models, attention-based multi-branch networks, graph neural networks) now form a core part of the state of the art for AQI (Air Quality Index) and particulate-matter estimation from structured and multi-source data. Reported performance can be very high on well-structured tabular datasets, with R^2 values above 0.99 in selected benchmarks and RMSE reductions of 23–45% relative to classical statistical baselines in multi-modal studies; however, these values are not directly interchangeable because pollutant type, prediction horizon, monitoring density, and validation design differ among studies. This review proposes a practical taxonomy of ensemble strategies and uses it to explain why diversity, rather than model count alone, is central to reliable air-quality prediction. Drawing on coordinated European and North American model-evaluation initiatives (AQMEII, HTAP) and on case studies in topographically and meteorologically complex Italian regions (the Po Valley, the Naples metropolitan area, and Campania), we show that effective ensemble design requires a balance among diversity, redundancy, computational feasibility, and interpretability. On the basis of a structured narrative synthesis, the main research gaps concern physics-informed and explainable ensemble frameworks, transferable and adaptive models, standardized benchmarks, severe-pollution-episode forecasting, and scalable distributed architectures. Open questions include how to design compact non-redundant CTM sub-ensembles and how to couple deep learning with chemical-transport physics in next-generation operational systems.



Academic Editor: Qin'geng Wang

Received: 30 May 2026

Revised: 4 July 2026

Accepted: 11 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: air quality; multi-model ensemble; chemical transport models; machine learning; deep learning; Bayesian model averaging; data assimilation; CAMS; AQMEII; Po Valley; Naples

1. Introduction

Air pollution remains one of the most consequential environmental and public-health challenges of the 21st century. Exposure to fine particulate matter (PM_{2.5}), ozone (O₃), nitrogen dioxide (NO₂), and other gaseous pollutants is associated with higher cardiovascular and respiratory mortality, reduced cognitive function, and substantial economic costs [1–3]. Reliable estimates and forecasts of pollutant concentrations are therefore not only a scientific objective; they are also a practical requirement for air-quality management, epidemiological exposure assessment, and evidence-based policy.

For many years, air-quality prediction has relied primarily on deterministic chemical transport models (CTMs), which combine atmospheric dynamics, emission inventories, chemical mechanisms, and physical parameterizations [4,5]. These models have matured into operational tools used worldwide. At the same time, they remain affected by structural uncertainties: emissions are imperfectly known, chemical mechanisms are simplified, meteorological inputs are uncertain, and sub-grid physical processes must be parameterized. Because no single CTM can represent all these processes with equal fidelity in all regions and seasons, multi-model ensembles have become a natural and increasingly necessary extension of deterministic modeling.

For this reason, CTM configuration is itself a modeling problem rather than a purely technical pre-processing step. Practical CTM setup requires choices about the horizontal and vertical grid, spin-up period, boundary and initial conditions, meteorological driver, land-use and deposition data, gas-phase chemical mechanism, aerosol module, photolysis scheme, planetary-boundary-layer and surface-layer parameterizations, emission inventories, temporal profiles, speciation factors, plume-rise treatment, and dust, fire, biogenic, ship, or traffic emission modules. These choices are commonly constrained by process knowledge, operational experience, and retrospective evaluation against monitoring networks, but they remain uncertain and can interact non-linearly. ML cannot remove the need for physically consistent CTM design, but it can assist parameter selection and calibration through surrogate modeling, Bayesian optimization, emission inversion, sensitivity screening, neural emulators for expensive chemistry or aerosol processes, and post-processing layers that identify systematic configuration-dependent biases.

The logic of combining models is well established in numerical weather prediction and climate science, where ensembles are indispensable for seasonal forecasting and long-term projections [6–8]. Air quality modeling adopted the same principle through early coordinated multi-CTM exercises in Europe, which later fed into AQMEII (Air Quality Model Evaluation International Initiative) and the CAMS regional ensemble service [9,10].

A parallel transformation has taken place in data-driven modeling. The growth of environmental monitoring networks, satellite products, reanalyses, and computational resources has stimulated rapid development of machine learning (ML) and deep learning (DL) approaches for air quality prediction [11–14]. Random Forests remain widely used in environmental applications because they handle nonlinear interactions, are relatively resistant to overfitting, and provide strong generalization in tabular settings [15]. Boosted-tree methods, especially XGBoost and LightGBM, further improved scalability and regularization, setting strong benchmarks in many environmental prediction tasks [16,17]. Together with gradient boosting machines [18,19], these methods now dominate many

tabular applications for PM and Air Quality Index (AQI) estimation [20], while hybrid architectures combining recurrent, convolutional, graph-based, and attention mechanisms have become central to spatio-temporal forecasting. Distributed learning and edge-computing approaches are also beginning to reshape multi-site air quality monitoring and ensemble prediction [21–24]. In this data-driven context, ensemble learning became a central ML paradigm after the introduction of bootstrap aggregating, which showed that aggregating multiple predictors can reduce variance and improve robustness [25].

Despite this rich literature, CTM-based ensembles and ML/DL ensembles are still too often discussed in separate communities. This review brings the two perspectives into a single framework, links them to the broader theory of ensemble methods, and focuses on the practical questions that matter for operational design: how much diversity is useful, when redundancy becomes harmful, and how model combinations can remain interpretable enough to support public decisions.

Review Scope and Methodology

This article is a structured narrative review rather than a PRISMA-style systematic review, statistical meta-analysis, or formal risk-of-bias assessment. The evidence base was assembled through iterative searches of Web of Science, Scopus, Google Scholar, ScienceDirect, PubMed, IEEE Xplore, ACM Digital Library, Copernicus/EGU publications and operational documentation from CAMS and related modeling initiatives. Searches covered the period 2000–2026, with earlier foundational papers retained when they introduced ensemble-learning, Bayesian model-averaging, trajectory-classification, or air-quality-model-evaluation concepts used later in the review. Representative search strings combined terms such as “air quality”, “chemical transport model”, “multi-model ensemble”, “CAMS”, “AQMEII”, “Bayesian model averaging”, “machine learning”, “deep learning”, “graph neural network”, “physics-informed neural network”, “model output statistics”, “data assimilation”, and “federated learning”.

To make this structured search transparent, Table 1 summarizes the search workflow used to identify and organize the literature. The strings were adapted to the syntax of each database; for example, combinations of (“air quality” OR PM_{2.5} OR ozone OR NO₂) AND (ensemble OR “multi-model” OR “model output statistics” OR “data assimilation” OR “machine learning” OR “deep learning” OR “physics-informed” OR federated) were used for broad retrieval, while narrower searches added CAMS, AQMEII, Bayesian model averaging, source apportionment, SHAP, PINN, Snapshot Ensembling, or CTM surrogate. Inclusion required direct relevance to ensemble air-quality prediction, model combination, CTM post-processing, source-resolved hybrid modeling, or uncertainty-aware forecasting. Exclusion criteria were: purely descriptive monitoring papers without a predictive or model-combination component; studies without sufficient validation description; indoor-only applications unless they introduced a transferable ensemble method; and papers whose performance claims could not be connected to a pollutant, forecast horizon, spatial scale, or validation setting.

Studies were prioritized when they (i) evaluated at least one ensemble, hybrid, or model-combination strategy for pollutant concentration, AQI, exposure, source attribution, or impact assessment; (ii) reported validation metrics or operationally relevant diagnostics; (iii) contributed a methodological framework for ensemble diversity, weighting, uncertainty quantification, explainability, transferability, or distributed learning; or (iv) represented an operational or coordinated intercomparison system. Studies with insufficient methodological description, unclear validation design, or no direct relevance to ensemble air-quality modeling were used only for background context, not for performance synthesis.

Table 1. Structured narrative review workflow used to build the evidence map and taxonomy.

Workflow Element	Implementation in This Review
Sources searched	Web of Science, Scopus, Google Scholar, ScienceDirect, PubMed, IEEE Xplore, ACM Digital Library, Copernicus/EGU publications, CAMS documentation, and coordinated-model-evaluation reports.
Period	2000–2026, with earlier foundational papers retained for ensemble learning, BMA, trajectory classification, and air-quality-model evaluation.
Core search themes	CTM ensembles; perturbed physics; statistical and tree-based ML; hybrid decomposition; CTM–ML post-processing; source-resolved hybrid modeling; deep spatio-temporal learning; PINNs; federated/distributed learning.
Screening criteria	Methodological relevance, explicit validation or operational diagnostic value, contribution to ensemble diversity/weighting/uncertainty/explainability/transferability, and relevance to regulated pollutants or AQI/exposure metrics.
Synthesis variables	Diversity source, information source, combination rule, pollutant, spatial and temporal scale, validation design, computational burden, interpretability, transferability, and operational maturity.

Evidence was synthesised qualitatively by grouping studies according to the mechanism that creates diversity (independent CTMs, perturbed physics, bootstrap/boosting, hybrid decomposition, CTM–ML post-processing, deep spatio-temporal architectures, physics-informed learning, or distributed/federated learning), the source of information used, the combination rule, the pollutant target, the spatial scale, and the validation strategy. Reported R^2 , RMSE, MAE, correlation, and skill-score values were retained as indicative, study-specific benchmarks. They were not statistically pooled because pollutant type, forecast horizon, monitoring-network density, spatial scale, training-data volume, and cross-validation protocol differ substantially among studies.

The taxonomy in Table 2 was developed iteratively from this evidence map and checked against existing systematic reviews and operational CTM practice. Research gaps were then derived by triangulating three sources: recurring limitations reported in the reviewed literature, empty or weakly populated cells in the method–application matrix, and the authors’ expert assessment of operational feasibility. The Italian examples were selected as illustrative case studies because they combine complex terrain, strong local emissions, Saharan-dust influence, CAMS/AQMEII continuity, and documented applications of data fusion, post-processing, trajectory ensembles, and high-resolution source apportionment. International CTM, ML, and distributed-learning examples are included throughout the review to support broader interpretation.

The review pursues five objectives:

1. To propose a unified taxonomy of ensemble approaches for air quality, spanning physical CTM ensembles, statistical and ML ensembles, hybrid schemes, source-aware post-processing, physics-informed neural components, and distributed/federated architectures, and to make explicit how this taxonomy extends reviews centered on single algorithmic families or monitoring pipelines;
2. To critically synthesize quantitative evidence on performance gains and their drivers across pollutants and spatial scales;
3. To highlight the theoretical foundations of ensemble diversity, with emphasis on shared bias and redundancy, drawing on work carried out within the AQMEII framework;
4. To discuss Italian case studies, including CAMS post-processing over the Po Valley and source apportionment in Naples and Campania, as examples of multi-model and data-fusion approaches in complex environments;

5. To identify research gaps and future directions, especially interpretable, transferable, physics-informed, and federated ensemble systems.

2. A Taxonomy of Ensemble Approaches in Air Quality

Building on recent systematic reviews [12,13], operational practice, and the evidence-mapping procedure described above, the following taxonomy organizes ensemble approaches for air quality modeling into eight major categories, summarized in Table 2. The categories were defined by the source of ensemble diversity, the physical or statistical information used by the members, and the way member predictions are combined.

The numerical performance values reported in Figure 1 and Table 2 are not directly comparable across studies. A site-level hourly AQI model evaluated with a random split, a regional CTM ensemble evaluated against independent stations, and a satellite-fusion model evaluated under spatial cross-validation answer different validation questions. In particular, very high R^2 values from dense, well-structured tabular datasets should not be extrapolated to severe episodes, new cities, or sparse monitoring networks without temporal, spatial, or external validation. Values reported in Table 2 should therefore be read as contextual examples, not as directly comparable league-table entries.

When a cited ML study reports an exceptionally high coefficient of determination, for example $R^2 \approx 0.99$, the value is interpreted here only together with the concentration range of the pollutant target, the number and spatial distribution of monitoring stations, the balance between background and exceedance records, and the validation split used. These descriptors are not always reported in the primary literature. In such cases, the present review treats the result as an upper benchmark for interpolation within the analyzed dataset, not as evidence of robust performance under sparse monitoring, imbalanced severe-episode distributions, or external deployment.

This taxonomy is intentionally broader than recent reviews focused mainly on either spatio-temporal ML/DL prediction or ML/IoT monitoring pipelines [12,26]. Its unifying feature is not the algorithmic family alone, but the mechanism by which predictive diversity and physical information enter the ensemble. Thus, independent CTMs, perturbed-physics CTM members, MOS/EMOS bias correction, source-contribution features, tree ensembles, deep spatio-temporal networks, PINN solvers or surrogates, and federated systems are placed on the same design axis. This makes explicit which approaches create diversity through independent physics, which through resampling or optimization, which through source-resolved physical diagnostics, and which through distributed data ownership or physics-constrained learning.

The taxonomy also clarifies an important distinction between pollutant-concentration forecasting and AQI classification. Reviews focused on deep learning stress that modern architectures should be evaluated according to the correlation structure they are designed to capture: temporal models alone are insufficient when dynamic spatial transport is important, while purely spatial models miss lagged emission and meteorological effects [27]. Conversely, systematic reviews of outdoor air-quality ML show that ensemble learning and regression dominate estimation tasks, whereas neural networks and support-vector methods have been used more often for forecasting tasks [13]. The distinction between estimation and forecasting tasks matters for ensemble design: the optimal combination of components depends on whether the target is a gridded concentration field, a station forecast, an AQI class, or an exposure metric.

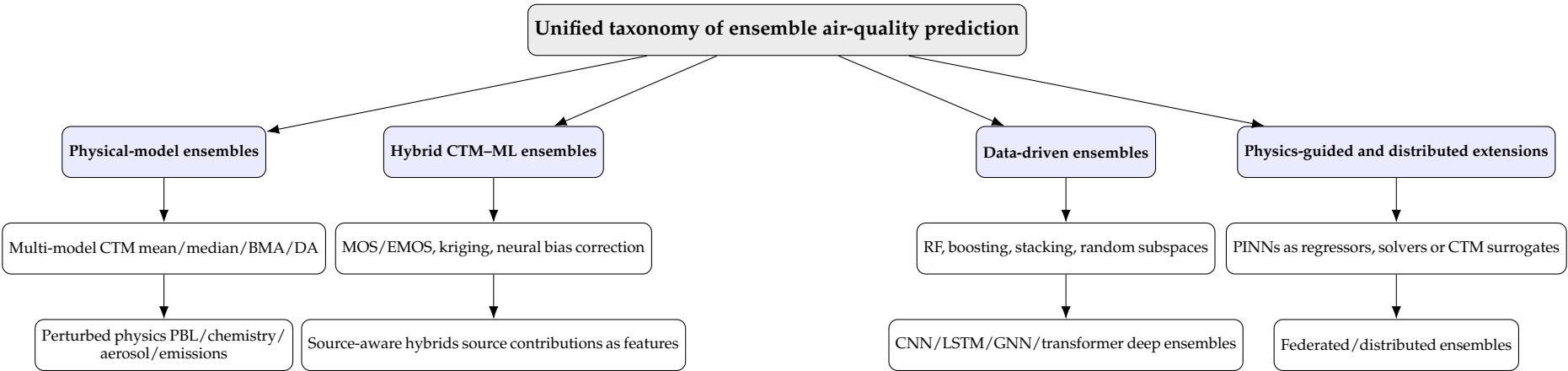


Figure 1. Conceptual taxonomy of ensemble approaches for air-quality prediction. The diagram complements Table 2 by showing how CTM, hybrid, data-driven, physics-informed, and distributed methods are connected by the common design principle of creating useful, non-redundant diversity.

Table 2. Taxonomy of ensemble methods in air quality prediction. Abbreviations : CTM: Chemical Transport Model; BMA: Bayesian Model Averaging; ML: Machine Learning; DL: Deep Learning; FL: Federated Learning; RF: Random Forest; GNN: Graph Neural Network; LSTM: Long Short-Term Memory; CNN: Convolutional Neural Network; GRU: Gated Recurrent Unit; PINN: Physics-Informed Neural Network; MOS: Model Output Statistics; NWP: Numerical Weather Prediction; DA: Data Assimilation; SHAP: SHapley Additive exPlanations; NDE: Neural Differential Equation.

Category	Description, Scale & Combination	Representative Examples	Target Pollutants	Application and Advantages	Main Limitations	Key References
Multi-model CTM ensemble	Collections of independent, full-chemistry CTMs combined via mean, median, BMA, or DA post-processing. Scale: regional–continental (~10–50 km); hemispheric/global. Combination: unweighted mean; ensemble median; BMA; rank-histogram calibration; optimal-subset selection	CAMS (11 models: CHIMERE, DEHM, EMEP, EURAD-IM, GEM-AQ, LOTOS-EUROS, MATCH, MINNI, MOCAGE, MONARCH, SILAM); AQMEII; HTAP; MICS-Asia; ENSEMBLE/ETEX	PM _{2.5} , PM ₁₀ , O ₃ , NO ₂ , SO ₂ , CO	Applications: operational O ₃ /PM forecasts (4-day horizon); uncertainty quantification; health and crop impact assessment. Advantages: physical consistency; structural diversity across chemistry/PBL schemes; calibrated probabilistic output; ensemble mean reduces RMSE and bias vs. individual models	High computational cost (multiple full runs); shared-input correlation limits effective diversity; BMA weights non-portable across scenarios; LULC mask inconsistencies between models	Galmarini et al. [28]; Solazzo et al. [10,29]; Riccio et al. [30]; Peuch et al. [9]; Colette et al. [31]; Baklanov et al. [5]

Table 2. Cont.

Category	Description, Scale & Combination	Representative Examples	Target Pollutants	Application and Advantages	Main Limitations	Key References
Perturbed-physics ensemble	Single CTM or NWP-chemistry model run repeatedly with systematically varied parameterizations (PBL, chemistry mechanism, aerosol module, emission factor). Scale: regional to urban (~1–10 km). Combination: unweighted mean; ANOVA variance decomposition; performance-based weighting	WRF-Chem/CMAQ/CAMx multi-physics suites; WRF with 7 PBL schemes (MYNN2, MYJ, YSU, ACM2; multi-mechanism CMAQ (CB6, SAPRC07); perturbed-emission CAMS sensitivity runs	PM _{2.5} , PM ₁₀ , O ₃ , NO ₂ ; PBL meteorology	Applications: sensitivity analysis; PBL-scheme inter-comparison; regional climate–chemistry sensitivity. Advantages: lower setup and code-maintenance cost per deliberately limited members; isolates parametric uncertainty; reveals dominant parameterization driver of spread; useful as precursor to full multi-model ensemble	Parameterization choices may be correlated; does not capture structural uncertainty from different dynamical cores; limited pollutant diversity unless multiple mechanisms tested	Tyagi et al. [32]; Luecken et al. [33]; Baklanov et al. [5]; Kim et al. [34]; Zhang et al. [35]
Statistical/ML ensemble (tree-based)	Ensembles of tree-based or classical statistical learners trained on observational/reanalysis feature sets. Scale: point/site to regional (multi-station, satellite-fusion global 1 km). Combination: bootstrap aggregating (RF); boosting (sequential error reweighting); stacking with meta-learner; SHAP-guided feature selection; performance-weighted voting	RF; XGBoost; LightGBM; CatBoost; GBM; optimized SVR; stacking (GB + CatBoost + XGB + LGB; R ² = 0.997, Taiwan 74-station); Bayesian-optimized SVR (Kaggle Air Quality, R ² = 0.9994); SERM stacked ensemble for Delhi AQI with SHAP attribution (PM _{2.5} and CO dominant); RFRK (RF + kriging, USA PM _{2.5} , 14 yr); GW-GBM (geographically weighted, China)	PM _{2.5} , PM ₁₀ , AQI, O ₃ , NO ₂ , SO ₂ , CO; traffic CO ₂ ; nanoparticles (PM ₁ , UFP); sulfate	Applications: PM/AQI estimation from tabular data; land-use regression (LUR/LURF); satellite-AOD downscaling; source identification. Advantages: best overall R ² for tabular data (≈0.79–0.997); implicit feature selection; SHAP interpretability; low inference cost; robust to outliers	Degrades for pollution peaks; requires dense observation networks; limited spatio-temporal extrapolation; tabular focus limits gridded output	Singh et al. [36]; Rybarczyk and Zalakeviciute [13]; Hu et al. [37]; Di et al. [38]; Chen et al. [39]; Agbehadji and Obagbuwa [12]; Ansari and Quaff [20]; Özüpak et al. [40]; Singh et al. [41]; Sarkar et al. [21]

Table 2. Cont.

Category	Description, Scale & Combination	Representative Examples	Target Pollutants	Application and Advantages	Main Limitations	Key References
Hybrid statistical–time-series ML	Pipelines coupling classical decomposition (ARIMA, EMD, VMD, STL, CEEMDAN) with ML/DL regressors; linear persistence modeled by decomposition, nonlinear residuals learned by ML/DL. Scale: site-specific to multi-city. Combination: per-component DL/ML prediction; aggregation of IMF component forecasts; error-correction stacking	ARIMA + gradient boosting; VMD-BiLSTM ($R^2 = 0.992$); ESWT-NLSTM ($R^2 = 0.990$); STL-XGBoost/LightGBM, Almaty ($R^2 > 0.98$ for $PM_{2.5}$ and $R^2 > 0.83$ for CO); ARIMA–ANN/ARIMA–SVM and CatBoost for $PM_{2.5}$ in Bangladesh (CatBoost best across Dhaka, Narayanganj, Gazipur); CEEMDAN-VMD-GRU; DNN-nonlinear ARIMA; SAC-QBSO-BiLSTM	$PM_{2.5}$, PM_{10} , CO, O_3 , AQI	Applications: multi-step-ahead AQI/PM forecasting (hourly to daily); seasonal anomaly capture; pollution peak detection. Advantages: reduces forecast error 10–30% vs. pure ARIMA or pure ML; captures low-frequency trend and high-frequency variability jointly; improved long-range dependency modeling	Decomposition-induced data-leakage risk; hyperparameter sensitivity; higher pipeline complexity; non-trivial interpretation of IMF contributions	Shahriar et al. [42]; Duan et al. [43]; Liu et al. [44]; Qi et al. [45]; Naizabayeva et al. [46]; Sarkar et al. [21]
Hybrid CTM–ML ensemble (MOS / DA)	CTM ensemble outputs used as physical predictors; ML layer (MLP, GBM, SPDE-INLA, Bayesian fusion) corrects gridded fields against observations; encompasses MOS, bias-correction neural nets, and variational/ensemble DA. Scale: regional to urban (~1–10 km gridded). Combination: EMOS; Bayesian variational approx. (KL-divergence); kriging of residuals; MOS regression; neural bias-correction; conformal intervals; EnKF; 4D-Var	CAMS-MOS (EMOS + SPDE-INLA, 2024; $r > 0.9$, near-zero bias); Bayesian hierarchical DA of CAMS PM_{10} Italy (169 stations, 2017; $R^2 \approx 0.83$); MLP on CAMS for Italian PM_{10} (413 stations, 4-day); RFRK (RF + CTM, USA); WRF–CAMx–PSAT (Campania)	$PM_{2.5}$, PM_{10} , O_3 , NO_2 , SO_2 , CO	Applications: operational improvement of CAMS forecasts; near-zero bias gridded products; source apportionment; WHO-threshold exceedance maps. Advantages: combines CTM physical consistency with ML flexibility; near-zero bias achievable; no CTM re-run at inference; gridded uncertainty maps	Structural CTM errors propagate unless corrected; MOS/BMA weights are scenario-dependent; dense observations required for training; scale mismatch between CTM grid and point observations	Chianese et al. [47]; Riccio and Chianese [48]; Riccio et al. [49,50]; Casciaro et al. [51]; Fioravanti et al. [52]; Baklanov et al. [5]

Table 2. Cont.

Category	Description, Scale & Combination	Representative Examples	Target Pollutants	Application and Advantages	Main Limitations	Key References
Deep learning ensemble	Ensembles of DL architectures combining temporal (LSTM, GRU, Bi-RNN), spatial (CNN, GNN), and attention (transformer, self-attention) modules; trained end-to-end on multivariate spatio-temporal data. Scale: urban station to multi-city regional (via GNN); satellite spatial fields (multi-modal fusion). Combination: averaging of DL outputs; stacking with meta-learner; boosting fusion; dropout Monte Carlo; deep ensemble variance; attention-weighted multi-modal fusion	CNN-LSTM; RF-CNN-GRU ($R^2 = 0.87$); Res-GCN-BiLSTM ($R^2 = 0.88$); LSTM-CNN-GNN stacking (MAE 11.30, RMSE 15.60, $R^2 = 0.89$ on Beijing); hybrid CNN-LSTM AQI classifier (F1 $\approx 91\%$); MAST-Net (CNN/LSTM multi-modal fusion using Sentinel-5P, MODIS, Landsat-8, ERA5, and ground stations; RMSE $\downarrow 23\text{--}31\%$); DL-GVMD ($R^2 = 0.999$); transformer + country embedding ($R^2 = 0.98$); hybrid LSTM-SVR	PM _{2.5} , PM ₁₀ , AQI, O ₃ , NO ₂ , SO ₂ , CO, NH ₃	Applications: short- to medium-range AQI forecasting; spatio-temporal PM fields; multi-station joint forecasting; pollution peak detection; multi-modal satellite-ground fusion. Advantages: $R^2 > 0.9$ typical; RMSE improvements 23–45% over baselines; captures nonlinear spatial correlations via GNN; multi-source data fusion	Black-box; high training/inference cost; requires large labelled datasets; degrades for long horizons and OOD events; temporal leakage in 68% of published studies; spatial CV in only 12%	Zhang et al. [27]; Agbehadji and Obagbuwa [12]; Wang et al. [53]; Sarkar et al. [21]; Raja et al. [54]; Madan et al. [55]; Kalaiselvi et al. [56]
Physics-informed NN (PINN) ensemble	NNs whose loss or architecture encodes PDEs governing atmospheric transport (advection–diffusion–reaction, mass conservation); ensembled with DL or CTM; includes Neural ODEs and dual-branch NDEs. Scale: site-specific to regional; designed for cross-domain transfer. Combination: physics-constrained loss; neural ODE integration; hybrid CTM-PINN fusion; SHAP attribution; Bayesian posterior sampling	AirPhyNet (GNN-based diffusion–advection differential-equation network); CTENet (Eulerian ADR-PINN; RMSE $\downarrow 45.8\%$ USA, 21.0% China vs. SOTA); Air-DualODE (open-system BA-DAE physics branch + data-driven Neural ODE branch); PINN-XAI hybrid (PINN + SHAP); Bayesian NNs (probabilistic AQI); physics-constrained LSTM	PM _{2.5} , PM ₁₀ , O ₃ , NO ₂ , CO; generalizable to any CTM-modeled species	Applications: improving generalization to unmonitored regions; reducing training-data requirements; explainable source attribution. Advantages: physical consistency by construction; improved OOD generalization; RMSE improvement 21–46% vs. pure DL; physics as regularizer	Immature at operational scale; PDE implementation complex; potential conflict with data-driven terms at high resolution; limited benchmark comparisons; high development cost	Hettige et al. [57]; Zhang et al. [58]; Liang et al. [59]; Selvam et al. [60]; Raissi et al. [61]; Sarkar et al. [21]

Table 2. Cont.

Category	Description, Scale & Combination	Representative Examples	Target Pollutants	Application and Advantages	Main Limitations	Key References
Federated/distributed ensemble	Decentralized ML/DL training across IoT sensors, monitoring stations, or city clients; local models share parameters or gradients with an aggregator rather than full raw-data streams; encompasses horizontal FL, vertical FL, and blockchain-secured FL. Scale: multi-city to national IoT networks. Combination: FedAvg; FedProx (proximal-regularized); loss-adaptive boosting; reputation-weighted aggregation; communication-efficient and secure aggregation	FL-LSTM; FL-BGRU (MSE 40.1); Fed-BiLSTM (RMSE 6.78); FL-DPLA-SSA-LSTM; FL-CRNN (Kanto, Japan); HAFA (Acc. 90.8%); blockchain FL with EMD–Transformer–BiLSTM (Ramadan et al. 2025); short-horizon accuracy 92.5%, $R^2 = 0.89$; FedAvg/FedProx; UAV-enabled FL (India); VGG-19+FL (Acc. 92.3%)	PM _{2.5} , PM ₁₀ , O ₃ , NO ₂ , SO ₂ , CO, CO ₂ , VOC, NH ₃ , AQI	Applications: distributed multi-site AQI forecasting; multi-operator monitoring networks; bandwidth-limited sensor streams; edge/IoT deployment. Advantages: scalable to heterogeneous IoT; reduces raw-data movement and communication burden vs. centralized systems; supports local personalization; improves accuracy vs. non-cooperative local models	Performance gap vs. centralized under non-IID data; high communication overhead; adversarial poisoning risk; node coverage gaps; heterogeneous hardware; multi-pollutant benchmarks lacking	McMahan et al. [62]; Jiang et al. [22]; Sarkar et al. [21]; Nguyen and Zettsu [63]; Ramadan et al. [23]; Alwabli [24]

Practical Guidance for Constructing Ensembles for Unseen Data

The preceding taxonomy is descriptive; operational use also requires guidance on how to build an ensemble that is likely to generalize beyond the calibration dataset. A robust workflow begins by defining the deployment target (station forecast, gridded concentration field, AQI class, exceedance probability, exposure product, or source attribution), the available predictors, and the validation regime. Model selection should then be evaluated with temporal blocks, spatial leave-site-out tests, or an external city/year whenever the intended use involves extrapolation. Random record-wise validation is useful for debugging but is insufficient evidence of unseen-domain skill.

A practical decision sequence is as follows. If physically consistent gridded forecasts are needed and several CTMs are available, a compact multi-CTM ensemble followed by MOS/EMOS, Bayesian fusion, or ML bias correction is usually the strongest operational design because it combines structural physical diversity with observation-based calibration. If only one CTM can be run, perturbed-physics members or source-resolved CTM outputs can provide targeted diversity, and a tree-based or neural post-processing layer can correct recurrent biases. If no CTM is available but dense observations, meteorology, land-use, and satellite predictors exist, Random Forest, gradient boosting, and stacked tree ensembles are strong low-to-moderate-cost baselines for tabular estimation. If dynamic spatial transport dominates and labelled data are sufficiently dense, graph neural networks, CNN-LSTM/GRU hybrids, attention models, or their snapshot/deep ensembles become more appropriate, but they require stricter leakage control and larger training sets. In data-poor or transfer settings, physics-informed, hierarchical, or domain-adapted hybrids are preferable to purely local black-box models because physical constraints and shared information reduce extrapolation risk.

The most effective combination is therefore context-dependent rather than universal. Under generous computational resources, the recommended hierarchy is as follows: diverse CTM members plus calibrated statistical/ML post-processing; source-aware CTM-ML hybrids where source apportionment is available; and deep spatio-temporal ensembles for high-resolution multi-sensor networks. Under limited resources, a carefully selected non-redundant CTM subset, a single CTM with MOS, or a boosted-tree model using CTM, meteorological, source, land-use, and satellite features is often preferable to a large but redundant ensemble. For interpretability-constrained applications, tree ensembles with SHAP, source-aware hybrids, and Bayesian or EMOS calibration are usually easier to justify than fully black-box deep ensembles. For edge or multi-institution deployments, federated ensembles are attractive only when centralized training is infeasible or when local personalization outweighs the engineering burden. Across all cases, diversity should be measured empirically: adding a member is useful only if it reduces independent validation error, improves calibration, or captures rare episodes not represented by existing members.

3. Multi-Model CTM Ensembles

3.1. CTM Configuration, Parameters, and ML-Assisted Calibration

Before discussing CTM ensembles, it is useful to make explicit what is being ensemble. A CTM forecast is determined not only by the governing transport and chemistry equations but also by a large set of configuration choices. Core decisions include the modeling domain and nesting strategy, horizontal and vertical resolution, spin-up duration, meteorological driver, boundary and initial conditions, chemical mechanism, aerosol module, dry and wet deposition schemes, biogenic and dust emissions, anthropogenic emission inventory, temporal disaggregation, chemical speciation, plume-rise treatment, and sub-grid parameterizations of boundary-layer mixing and surface exchange. Parameter values are usually selected from prior process studies, model documentation, sensitivity

analyses, and retrospective evaluation against observations. In operational systems, these choices are constrained by robustness, run time, and reproducibility as much as by local optimum skill.

Machine learning can support several parts of this configuration problem. First, sensitivity-screening and surrogate models can identify which CTM parameters or inputs most strongly control PM_{2.5}, PM₁₀, O₃, NO₂, or source-specific errors, reducing the number of expensive CTM experiments. Second, Bayesian optimization or emulation can tune uncertain parameters or emission scaling factors against observations while preserving physically plausible bounds. Third, source-aware ML post-processing can learn sector-specific correction factors when a CTM or source-apportionment module provides contributions from traffic, residential combustion, shipping, dust, fires, secondary aerosol, or boundary inflow. Fourth, neural emulators and PINN components can approximate computationally expensive chemistry or transport terms, although such components must be tested for numerical stability, conservation, and regime transfer before being embedded in operational CTMs. ML-assisted calibration should therefore be regarded as a complement to physical modeling, not as an unconstrained replacement for CTM parameter selection.

3.2. Background: The Theory of Ensemble Methods

The rationale for ensemble methods can be formalized through the classical bias–variance decomposition of the expected squared prediction error:

$$\mathbb{E}[(y - \hat{f}(x))^2] = \underbrace{\text{Bias}^2[\hat{f}(x)]}_{\text{systematic error}} + \underbrace{\text{Var}[\hat{f}(x)]}_{\text{random error}} + \sigma_\epsilon^2, \quad (1)$$

where y is the true value, $\hat{f}(x)$ is the model prediction, and σ_ϵ^2 is irreducible noise. Averaging over M uncorrelated models of identical skill collapses the variance component by a factor $1/M$, leaving the bias component unchanged. In practice, ensemble members share common inputs, training data, or model structure, so their outputs are correlated. If the pairwise correlation among members is ρ , the expected ensemble variance is as follows:

$$\text{Var}[\bar{f}] = \frac{1}{M}\sigma^2 + \frac{M-1}{M}\rho\sigma^2, \quad (2)$$

which reduces to $\rho\sigma^2$ as $M \rightarrow \infty$. This result, well established in the machine learning literature [64–66], shows why ensemble size alone is not enough: the benefit comes from members that make different errors. In Dietterich’s terminology, ensembles can help for statistical, computational, and representational reasons, but all three mechanisms depend on some degree of complementarity among the constituent predictors [65].

Multi-model CTM ensembles follow the same principle. Shared emission inventories, boundary conditions, or meteorological drivers create correlated errors [29], and the AQMEII analysis by Solazzo et al. [29] showed that CTMs also share substantial bias and variance beyond what common inputs alone can explain. This finding has a direct operational consequence: a reduced ensemble selected for diversity can outperform a larger but redundant one. The broader theory of ensemble learning reaches the same conclusion: accuracy, diversity, and complementarity must be considered together [67].

3.3. Bayesian Framework for Model Combination

A natural extension of simple averaging is the Bayesian paradigm, which assigns weights to models proportional to their posterior probability given observations. Bayesian Model Averaging (BMA) expresses the ensemble predictive distribution as a weighted mixture [68,69]:

$$p(y | \mathbf{x}, \mathcal{D}) = \sum_{k=1}^M w_k p(y | \mathbf{x}, \mathcal{M}_k, \mathcal{D}), \quad (3)$$

where w_k is the posterior model weight and $\mathcal{D}$ is the training dataset. BMA provides calibrated predictive distributions rather than point forecasts, which is particularly valuable when results are used to inform risk-based decisions.

Riccio et al. [49] provided an earlier Bayesian treatment of air quality data at the spatio-temporal level, demonstrating that a significant fraction of apparent CTM bias could be explained by subgrid spatial variability, with implications for model evaluation protocols, while Riccio et al. [50] provided the theoretical derivation showing that the widely used median-model combination [28] can be recovered as a special case of BMA under specific distributional assumptions, thereby establishing a rational basis for what had previously been a heuristic approach. Subsequently, Solazzo et al. [29] applied multi-model ensemble approaches to atmospheric dispersion, finding improvements in forecast skill of 5–40% relative to unweighted means, while also highlighting limitations related to weight portability across scenarios and the sensitivity to the length of the training dataset.

3.4. Operational Implementations

Multi-model CTM ensembles in air quality have their intellectual roots in numerical weather prediction and climate science, where statistical super-ensemble and multimodel weighting approaches were shown to improve forecast skill by learning model-specific weights from a training period [6–8]. In air quality, the first large-scale coordinated ensemble exercises were carried out under the ENSEMBLE platform developed at the European Commission's Joint Research Centre (JRC) for long-range transport and dispersion modeling [28]. These early experiments, which included operational models participating in the ETEX-I tracer experiment, demonstrated that median and BMA-based combinations reliably outperformed individual members and provided meaningful uncertainty bounds [50,70].

The AQMEII initiative subsequently scaled these concepts to regional-scale chemistry transport in Europe and North America. In its first phase, more than ten CTMs were run over a full year (2006) and their outputs compared against dense observation networks [10]. A key finding was that the most common multi-model approach—the unweighted ensemble mean—could be outperformed by optimally selected subsets of members, and that clustering methodologies could identify skillful reduced ensembles without requiring a priori knowledge of individual model skill. This insight was formalized by Solazzo et al. [29], who introduced information-theoretic diagnostics for redundancy and showed that reduced, non-redundant ensembles consistently outscored full ensembles in most performance metrics. Beyond the identification of skill-based ensemble subsets, information-theoretic and distance-based approaches have been applied to reduce the data complexity inherent in large multi-model dispersion ensembles. Riccio et al. [30] demonstrated that mutual information and correlation distance measures can be used to cluster the models of the ETEX-1 ensemble and select a compact, non-redundant subset that reproduces the performance of the full ensemble, providing a computationally efficient route to optimal ensemble design.

AQMEII is used in this review as a detailed and well-documented evidence base. Its distinctive value is the availability of harmonized, year-long simulations over Europe and North America, common evaluation protocols, and dense observational networks. Similar questions of shared inputs, correlated errors, inter-model dependence, and region-specific emission uncertainty also arise in CAMS, HTAP, and MICS-Asia intercomparison contexts. The AQMEII results are therefore interpreted as transferable design principles—especially

the need to separate skill from diversity—that must be re-evaluated in each emissions, meteorological, and monitoring regime.

The Copernicus Atmosphere Monitoring Service represents the current state of the art in operational multi-model CTM ensembles [9,31]. As of 2022, the CAMS European air quality ensemble comprises eleven regional models (CHIMERE, EMEP, EURAD-IM, LOTOS-EUROS, MATCH, MINNI, MOCAGE, MONARCH, SILAM, IFS-COMPO, DEHM), driven by common ECMWF meteorological fields and unified CAMS emission inventories, and providing daily forecasts at approximately 10 km resolution over a domain stretching from Iceland to the Mediterranean [71].

Products cover PM₁₀, PM_{2.5}, O₃, NO₂, SO₂, and CO for the next four days. Individual model outputs and the combined ENSEMBLE product are delivered operationally; for the regional production system, the combined product is a median of the available member forecasts, and the inter-member spread provides a practical indicator of forecast uncertainty [9,31,72].

Recent sensitivity analyses of the CAMS system, such as the evaluation of updated temporal emission profiles (CAMS-REG-TEMPO), illustrate the level of operational sophistication now achieved: NO₂ weekly-cycle correlations improved by up to +0.17 with updated road transport profiles, while PM₁₀ winter correlations improved by +0.13 [73]. Such studies highlight that ensemble performance depends not only on model diversity but also on the quality and representativeness of shared inputs.

3.5. Ensemble Diversity, Redundancy, and Optimal Combination

A key theoretical result in ensemble learning is that performance improvements depend not only on the accuracy of individual members but also on their diversity, since highly correlated predictors tend to share systematic errors and therefore provide limited ensemble benefit [74]. This is particularly important in air quality modeling, where a recurring finding is that CTMs share large fractions of both bias and variance, driven by common inputs and, to a larger extent, by structural similarities in parameterizations, chemical mechanisms, and numerical schemes Solazzo et al. [29]. This result implies that the effective number of independent models in a large ensemble is substantially smaller than the nominal count. Diversity and skill should therefore be considered disjointly when selecting ensemble members. Simply choosing the best-scoring individual models is sub-optimal because highly correlated top performers offer little additional variance reduction [29]. Conversely, a carefully selected subset of poorly correlated models—even if individually less skilled—can explain an equivalent or larger fraction of observed variance than the full ensemble.

Bayesian model averaging provides a unified probabilistic framework for weighting ensemble members based on observed performance. Potemski and Galmarini [70] found that, while BMA can improve upon the ensemble median in calibration, the portability of BMA weights across different release scenarios (i.e., event-specific combinations of source location, emission timing and magnitude, release height, meteorological regime, and boundary/initial conditions) is limited, an issue that emerges in the context of ML-based post-processing as well. The AQMEII-4 exercise extended ensemble evaluation to include dry deposition of ozone, revealing that land-use/land-cover (LULC) mask differences among models introduce an additional layer of structural diversity—or inconsistency—that substantially complicates cross-model evaluation [75].

3.6. Multi-Model Ensembles for Health and Agricultural Impact Assessment

Beyond air quality prediction *per se*, multi-model CTM ensembles have been used to propagate model uncertainty into health and agricultural impact assessments, an ap-

plication pioneered within the AQMEII framework. Solazzo et al. [76] used twelve PM_{2.5} and ozone fields from AQMEII-3 models, bias-adjusted to observations and combined in a multi-model ensemble, to estimate premature deaths and crop yield losses across Europe. The multi-model ensemble approach reduced uncertainty in ozone fields by 61% and in PM_{2.5} fields by 46% relative to individual models, translating into equivalent reductions in the uncertainty of premature mortality and crop loss estimates—a result that strengthens the evidence base for emission-control policy [76]. The analysis attributed approximately 370,000 premature deaths per year to PM_{2.5} and ozone exposure over Europe, with an inter-quartile range of 260,000–415,000 that reflects remaining model uncertainty.

4. Perturbed-Physics Ensembles

Perturbed-physics ensembles represent a distinct strategy for quantifying parametric uncertainty within a single modeling system, in contrast to the structural diversity exploited by multi-model CTM ensembles. Rather than assembling fully independent chemistry transport models, a perturbed-physics ensemble repeatedly executes one CTM or coupled NWP-chemistry system with systematically varied parameterizations—planetary boundary layer (PBL) schemes, aerosol microphysical modules, chemical mechanisms, or emission scaling factors—and derives an ensemble distribution from the resulting spread.

They are particularly relevant for coupled meteorology–chemistry systems because meteorological parameterizations propagate directly into dispersion, vertical mixing, photolysis, and deposition. The WRF sensitivity study over the Naples–Caserta coastal conurbation compared seven PBL schemes against aircraft profiles and found that TKE-closure schemes generally outperformed first-order closures, with MYNN2 closest to observed profiles in many cases; importantly, the study also showed that variability induced by the meteorological situation and sampling location was larger than variability attributable to PBL scheme choice [32].

4.1. Theoretical Motivation

From the bias–variance perspective (Section 3.2), perturbed-physics ensembles target parametric uncertainty rather than structural uncertainty. Because all members share the same dynamical core, their errors are more strongly correlated than those of a full multi-model ensemble, limiting the achievable variance reduction according to Equation (2). Nevertheless, the approach has important advantages when the ensemble is designed as a targeted set of perturbations rather than as a full factorial experiment. For a fixed number of members, it is often less costly than assembling independent CTMs because the same code base, pre-processing chain, domain configuration, and post-processing infrastructure are reused. This computational advantage can disappear if all combinations of PBL schemes, chemical mechanisms, aerosol modules, and emission perturbations are explored. In practice, perturbed-physics ensembles are therefore most useful as screening tools: they isolate specific uncertainty sources through ANOVA-style variance decomposition and help identify which parameterizations contribute most to spread before a larger multi-model exercise is undertaken.

4.2. Applications in Regional Air Quality Modeling

The most extensively studied application of perturbed-physics ensembles in air quality concerns PBL scheme uncertainty. An example of PBL uncertainty analysis is Tyagi et al. [32], where a seven-member WRF ensemble over the Naples–Caserta region was applied, comparing MYNN2 (TKE closure), MYJ, YSU, ACM2, and three additional non-local schemes. The study found that meteorology-induced variability—driven by the choice of PBL scheme—exceeded the variability attributable to other model parameters

for temperature, mixing depth, and near-surface winds, with direct consequences for modeled PM₁₀ and O₃ concentrations. TKE-based closures, particularly MYNN2, consistently outperformed first-order non-local schemes in reproducing the complex coastal boundary layer of the Naples area, a result with clear implications for ensemble member selection in operational systems.

At the chemical mechanism level, perturbed-physics ensembles have been constructed by running CMAQ with multiple gas-phase chemical mechanisms (CB6, SAPRC07, RACM2), revealing mechanism-dependent spreads in O₃ and PM_{2.5} that can exceed the spread from meteorological perturbations under low-NO_x conditions [5]. WRF-Chem and CAMx multi-physics suites, routinely used within AQMEII sensitivity experiments, provide further evidence that aerosol module choice (modal vs. sectional size distribution) introduces non-negligible spread in PM_{2.5} mass and composition [29].

Emission factor perturbations represent a third axis of the perturbed-physics strategy. Sensitivity runs within the CAMS framework have systematically perturbed road transport, residential heating, and ship emission totals, documenting how changes of ± 20 –30% propagate differently into NO₂ and PM_{2.5} depending on season and meteorological regime [73]. ANOVA decomposition of the resulting variance enables attribution of ensemble spread to individual emission sectors, providing direct guidance for prioritizing emission control strategies.

4.3. Comparison with Multi-Model Ensembles

The key distinction between perturbed-physics and multi-model ensembles is the nature of the diversity they exploit. Multi-model ensembles draw on structural differences across independently developed codes (different dynamical solvers, distinct chemical mechanisms, separate emission pre-processing chains), generating broader—and less correlated—spread. Perturbed-physics ensembles generate narrower spread, confined to the parameter space accessible within one model's design, but their variance decomposition capability makes them uniquely suited for sensitivity analysis and for identifying the dominant uncertainty drivers before a more expensive multi-model exercise is undertaken.

In practice, the two strategies are complementary: perturbed-physics runs to pre-screen model configurations and identify non-redundant members for the full multi-model ensemble [5,29]. The Italian case studies in Section 11 provide further examples of this complementarity.

5. Ensemble Machine Learning: Tree-Based Methods

Tree-Based Ensembles: Theoretical Basis and Performance

The development of modern ensemble methods is closely connected to broader advances in statistical learning theory and pattern recognition, which have provided the conceptual foundations for many of the predictive algorithms used in environmental modeling [77]. Tree-based ensemble methods—bagging (Random Forests), boosting (AdaBoost, gradient boosting, XGBoost, LightGBM, CatBoost), and stacking—have become the primary tools for air quality prediction from structured tabular data [36,41,78]. Their dominance reflects several favorable properties: robustness to missing data and outliers, implicit feature selection, natural handling of nonlinear interactions, and good calibration without extensive hyperparameter tuning.

Extremely randomized trees (Extra-Trees) extend this line of work by randomizing both the attribute and the cut-point at each node, further reducing variance at the cost of a small increase in bias, and providing substantial computational savings [79]. A related source of ensemble diversity is the random subspace method, in which individual trees

are trained on randomly selected subsets of the feature space, thereby reducing correlation among members and improving generalization [80].

Multiple systematic reviews confirm that, for $\text{PM}_{2.5}/\text{PM}_{10}$ estimation using meteorological and land-use predictors, Random Forests and gradient boosting outperform single learners (linear regression, SVMs, single decision trees) as well as many neural network architectures trained on tabular data [12,13,40]. The widespread adoption of Random Forests in air-quality applications has also been driven by meteorological normalization frameworks, which enable the separation of emission-driven trends from meteorological variability and have become standard tools for policy evaluation and trend analysis [81]. Recent head-to-head evaluations confirm this superiority: on the Taiwan Air Quality Dataset (2016–2024, 4.6 million hourly records from 74 stations), a weighted voting ensemble of Gradient Boosting, CatBoost, XGBoost, and LightGBM achieved a validation $R^2 = 0.9969$ and MSE of 0.6553, outperforming fifteen baselines including LSTM (MSE 45.4) [41]. SHAP-based interpretability confirmed that $\text{PM}_{2.5}$ and CO are dominant AQI predictors, demonstrating that ensemble accuracy need not come at the cost of complete opacity [41].

These high validation scores require explicit qualification. Random record-wise cross-validation primarily tests interpolation within a densely sampled feature space; temporal block validation tests robustness across unseen seasons or emission regimes; spatial leave-site-out validation tests geographical portability; and external validation tests transfer to a different city, year, or monitoring network. The latter two settings are usually more demanding and are especially important when reported R^2 values exceed 0.99. Such values should be interpreted together with the pollutant concentration range, number and spatial density of monitoring stations, station-type balance, frequency of exceedance episodes, and the proportion of low- versus high-concentration records. A dataset dominated by background or moderate-concentration hours may yield a very high global R^2 while still underpredicting rare pollution peaks; a dense urban network may also support interpolation that would not transfer to sparse rural stations or a different city. For this reason, the present review treats such values as task-specific upper benchmarks rather than as general estimates of operational forecasting skill. When the original study does not report the pollutant concentration range, the number of exceedance cases, the station-type distribution, or the degree of spatial sparsity, those omissions should be considered part of the uncertainty surrounding the reported score. Future high-performing ML studies should therefore report at minimum the pollutant range and distribution, the number and density of monitoring stations, the fraction of high-pollution or AQI-threshold exceedance records, and the exact temporal, spatial, or external validation protocol.

6. Hybrid Statistical–Time-Series ML Ensembles

Hybrid statistical–time-series ML ensembles are useful when pollutant series combine regular cycles with abrupt, nonlinear deviations. In these pipelines, classical time-series or decomposition methods capture persistence, seasonality, and trend, while ML or DL components learn the residual structure that remains. Autoregressive models (ARIMA, ARIMAX, seasonal decomposition) are therefore often paired with SVR, neural networks, or gradient boosting [42,82]. Reviews consistently report that such hybrid pipelines reduce forecast errors by 10–30% relative to pure ARIMA or pure ML models, particularly for multi-step horizons where error accumulation degrades purely recurrent models [55,82].

Signal decomposition strategies—empirical mode decomposition (EMD), variational mode decomposition, seasonal-trend decomposition by LOESS (STL)—can be combined with gradient boosting to capture both low-frequency trend and high-frequency variability. Naizabayeva et al. [46] applied this hybrid approach in Almaty, Kazakhstan, achieving

$R^2 > 0.98$ for $\text{PM}_{2.5}$ and $R^2 > 0.83$ for CO on hourly data, outperforming individual XGBoost and LightGBM models.

7. Hybrid CTM–ML Ensembles (MOS and Data Assimilation)

Hybrid CTM–ML ensembles do not replace physical modeling. Instead, they use CTM outputs as physically informed predictors and apply statistical or ML layers to correct, calibrate, and downscale the resulting fields against observations. This family includes model output statistics (MOSs), neural bias correction, variational and ensemble data assimilation, and Bayesian hierarchical fusion. The approach preserves the spatial and chemical consistency of CTMs while using observations to remove systematic errors, without requiring new CTM runs at inference time. The CAMS-MOS product, launched in 2024, exemplifies this approach: it applies ensemble model output statistics (EMOSs) combined with a spatio-temporal SPDE-INLA interpolation to the CAMS multi-model output, achieving near-zero bias, correlations routinely exceeding 0.9 for all regulated pollutants, and significant improvement in exceedance prediction even for the stringent WHO 2021 thresholds [83]. An earlier demonstration of this paradigm for Italy used CAMS ensemble outputs as predictors fed to a multi-layer perceptron, which corrected the strong seasonal bias of raw CTM outputs for daily PM_{10} and improved forecasts up to four days ahead at 413 monitoring stations [47].

Source-Aware Hybrid Statistical–ML Approaches

A recent extension of CTM–ML post-processing uses source contributions as predictors rather than relying only on total CTM-simulated pollutant concentrations. In these source-aware hybrids, the CTM or source-apportionment module first decomposes $\text{PM}_{2.5}$, PM_{10} , NO_2 , SO_2 , or other pollutants into sectoral or process contributions (for example biomass burning, road transport, residential combustion, shipping, dust, secondary aerosol, or boundary inflow). A statistical or ML model then uses these contributions, together with meteorology and observations, as features for bias correction or concentration prediction. This design is more informative than a black-box correction of the total CTM field because it allows the ML layer to learn source-specific error patterns and provides a direct link to emission-control measures.

The approach is consistent with recent source-decoupled hybrid studies in which CMAQ- or CTM-derived source contributions are combined with LightGBM or related learners to improve $\text{PM}_{2.5}$ prediction and interpretability [84,85]. Source-aware hybrids are especially useful when emission inventories are uncertain, when different sectors dominate different seasons, or when high-concentration events are driven by episodic sources such as biomass burning or dust. They also provide a natural bridge between operational forecasting and source apportionment because the same model can output both corrected concentrations and feature-level evidence about likely drivers.

8. Deep Learning Ensembles

8.1. Recurrent and Convolutional Architectures

Long short-term memory (LSTM) and gated recurrent unit (GRU) networks have been extensively applied to air quality time series, exploiting their ability to capture long-range temporal dependencies [27]. Convolutional neural networks (CNNs) extract local spatial features from gridded fields or multi-station arrays, and their combination with LSTMs in CNN–LSTM architectures jointly models spatial and temporal dynamics [86]. The widespread use of LSTM models in air-quality forecasting is rooted in their ability to mitigate vanishing-gradient effects and preserve information over long temporal horizons, as originally introduced by Hochreiter and Schmidhuber [87]. A comprehensive review of

deep learning for air pollutant concentration prediction [27] documents that deep models consistently outperform shallow ML baselines for short-term forecasting in dense urban networks, with R^2 values above 0.9 being common for single-station next-hour predictions.

8.2. Attention Mechanisms and Multi-Modal Fusion

Graph neural networks (GNNs) [88] explicitly model spatial correlations between monitoring stations through learnable adjacency matrices, and their combination with LSTM or transformer components in spatial–temporal attention architectures represents the current frontier of data-driven air quality prediction. A recent study [54] built a hybrid LSTM–CNN–GNN model with stacking and boosting ensemble fusion, achieving $R^2 > 0.92$, MAE = 11.30, and RMSE = 15.60 on the Beijing dataset, outperforming standalone LSTM baselines. The multi-modal attention-based spatio-temporal network MAST-Net, combining LSTM, CNN, satellite imagery from Sentinel-5P, MODIS, and Landsat-8 with ground observations, achieved RMSE improvements of 23–31% and correlations of 0.91–0.94 across PM_{2.5}, PM₁₀, NO₂, and O₃ [56].

8.3. Snapshot and Optimization-Trajectory Deep Ensembles

Deep ensembles can also be generated from a single optimization trajectory. Snapshot Ensembling saves multiple network states at different local minima along cyclical learning-rate training and averages their predictions at inference time [89]. Related weight-space or trajectory-based methods try to obtain ensemble diversity without training many independent networks from scratch. In air-quality applications, this family is attractive when GPU time is limited but a deep spatio-temporal model still requires uncertainty-aware or more robust predictions.

These methods remain compatible with the bias–variance–diversity interpretation developed in Section 3.2. Snapshot members are not independent in the same sense as separately trained networks or different CTMs, so their effective diversity may be smaller. Nevertheless, if the cyclical optimization path reaches functionally different solutions, averaging predictions can reduce variance and improve calibration. The theoretical caveat is the same as for CTM ensembles: gains depend on the correlation of member errors, not on the nominal number of saved snapshots. Thus, snapshot ensembles are best viewed as low-cost diversity mechanisms whose value must be verified with independent temporal, spatial, or external validation.

9. Physics-Informed Neural Network (PINN) Ensembles

Physics-informed neural network (PINN) ensembles differ from the other categories in Table 2. They embed known physical laws, such as the partial differential equations governing advection, diffusion, and chemical reaction, directly into the loss function or network architecture. The general PINN framework was formalized by Raissi et al. [61], who showed how governing differential equations can be included in neural-network training for both forward and inverse problems. In air quality, this idea is attractive because it can regularize learning, improve extrapolation to unmonitored regions, and reduce the risk that a black-box model fits patterns that are statistically convenient but physically implausible [57]. CTENet, which embeds the advection–diffusion–reaction equation into a PINN framework using an Eulerian representation, achieved RMSE improvements of 45.8% on the USA dataset and 21.0% on the China dataset relative to state-of-the-art baselines [58]. Air-DualODE integrates dual branches of neural ordinary differential equations to model open-system dynamics, addressing the mismatch between closed-system assumptions in classical CTMs and the open boundary conditions of real atmospheric domains [59]. PINNs can also play a deeper role than regularized regression. They may be embedded inside

a CTM as differentiable solvers, surrogate chemistry modules, or learned parameterizations of computationally expensive PDE-governed processes. A recent physics-informed deep neural network coupled within WRF-Chem, for example, targeted stiffness-driven secondary organic aerosol chemistry and demonstrated how a neural component can accelerate or replace part of a CTM solver while retaining physical constraints [90]. In this configuration, the PINN is not merely post-processing CTM output; it becomes part of the numerical modeling chain and must be evaluated for stability, conservation, generalization across chemical regimes, and computational speed-up.

These physics-guided ensembles point toward models that are not only accurate, but also easier to interpret and more credible when they are applied outside the conditions represented in the training data.

10. Federated and Distributed Ensembles

As urban IoT sensor networks proliferate, federated learning (FL) has become relevant to air quality mainly as a distributed forecasting architecture for multi-operator, bandwidth-limited, or edge-computing monitoring systems. In FL, local models are trained at monitoring stations, city clients, or sensor gateways; only model updates (gradients or parameters) are transmitted to an aggregator, which constructs a global model combining information from all participants [21,22]. The concept of distributed computing for coupled environmental modeling predates modern FL: early grid-computing architectures already demonstrated how atmospheric, air quality, and ocean circulation models could be integrated across geographically distributed resources into a single operational forecasting chain. Giunta et al. [91] described such a system for the Naples area, coupling MM5, PNAM, POM, and WW3 over a Globus Toolkit grid infrastructure to deliver continuous weather and air quality forecasts—an early precursor of the distributed ensemble architectures now explored through FL.

More recent workflow-based environmental cyberinfrastructures extended this idea from grid-enabled forecasting chains to portal-based model coupling and data management. For example, Montella et al. [92] described the FACE-IT Galaxy–Globus workflow engine for operational food-quality prediction in mussel farms in the Bay of Naples, showing how heterogeneous environmental models, pre/post-processing steps, and large spatially referenced outputs can be orchestrated through cloud-enabled workflow services. Recent studies have demonstrated the feasibility of federated AQI forecasting in multi-site urban environments, showing that federated models can approach the accuracy of centralized baselines while reducing the need to move high-volume raw sensor streams [24]. A spatially distributed FL framework using convolutional recurrent neural networks across the Kanto region of Japan showed that geographically partitioned local models, aggregated through a federated protocol, improved prediction accuracy relative to non-cooperative local models [63]. The added engineering complexity is justified only when centralized learning is impractical, for example because data are produced by several institutions, communication bandwidth is limited, real-time edge inference is required, or local personalization is important. When all data can be reliably centralized and validation conditions are stable, a conventional centralized or hybrid CTM–ML model is usually simpler and may remain preferable. The AQMEII experience provides a useful conceptual analogy: just as CTM member selection in AQMEII sought to maximize diversity while minimizing redundancy, FL frameworks need to balance global model coherence with local model diversity to avoid the known failure mode of FL aggregation under heterogeneous (non-IID) data distributions [93].

10.1. Federated Learning Algorithms and Aggregation Strategies

The canonical aggregation algorithm, FedAvg [62], averages client model weights proportionally to local dataset size and was originally introduced as a communication-efficient approach for training deep networks on decentralized, non-IID, and unbalanced data without transferring complete local datasets. Extensions addressing the non-IID data problem include FedProx, which adds a proximal regularization term to client objectives to bound deviation from the global model, and loss-adaptive boosting schemes that up-weight clients with high local error. For air quality specifically, Lu et al. [93] demonstrated that geographically heterogeneous local data distributions—typical of monitoring networks spanning urban, suburban, and rural environments—require careful regularization to avoid aggregation instability.

For operational air quality, the most relevant engineering issues are communication efficiency, robustness to non-IID data, resilience against faulty or adversarial nodes, and graceful degradation when sensors disconnect. Blockchain-secured FL architectures and secure aggregation protocols have been explored mainly as reliability and auditability mechanisms for industrial multi-pollutant monitoring [23]. UAV-enabled FL deployments further extend the paradigm to mobile sensing platforms in regions with sparse ground-station coverage [21].

10.2. Performance Relative to Centralized Systems

Empirical comparisons consistently show that FL models approach but do not fully match centralized baselines when data distributions are heterogeneous [22,24]. Fed-BiLSTM achieved RMSE of 6.78 on multi-city PM_{2.5} forecasting, comparable to centralized LSTM while avoiding central transfer of complete local sensor records [21]. The FL-CRNN framework deployed across the Kanto region of Japan improved upon non-cooperative local models and demonstrated graceful degradation as the number of participating nodes decreased, a robustness property of particular value for IoT deployments subject to intermittent connectivity. Multi-pollutant benchmarks (covering O₃, NO₂, SO₂, VOC, and CO simultaneously) remain scarce, representing one of the most significant gaps in the FL-for-AQ literature identified in this review. The distribution of ensemble forecast products to end users and downstream services is an equally important dimension of operational ensemble systems. Montella et al. [94] described a grid-oriented component architecture for on-demand delivery of environmental datasets—including meteorological fields and modeled concentration outputs—by integrating OPeNDAP-based data servers into GT4 web services and exploiting a resource broker to optimize replication and load balancing, anticipating the service-oriented paradigms now used in CAMS and similar operational platforms.

11. Case Studies in Complex Italian Environments

Italy offers two particularly instructive case study regions for multi-model ensemble methods: the Po Valley, one of the most polluted areas in Europe, and the metropolitan area of Naples (Campania region), a topographically and meteorologically complex coastal urban-industrial environment. These regions share the characteristic of PM concentrations that frequently exceed EU thresholds, driven by a combination of intense local emissions and strong synoptic controls on transport and dispersion.

The case studies are included as illustrative examples of methodological transfer rather than as a country-specific evidence base. The Po Valley highlights basin-scale stagnation, winter particulate accumulation, and operational CTM ensemble post-processing. Naples and Campania highlight coastal recirculation, complex topography, port and road-transport emissions, Saharan-dust influence, trajectory ensembles, and high-resolution source attri-

bution. Together, they provide compact examples of situations in which ensemble diversity, post-processing, and physical interpretability matter for operational decisions.

11.1. Po Valley: Data Fusion and Ensemble CTM Post-Processing

The Po Valley in Northern Italy is enclosed by the Alps to the north and west and the Apennines to the south, creating an atmospheric basin highly susceptible to pollutant accumulation during anticyclonic conditions. Multi-model CTM ensembles are a natural tool for this region due to the large spread among models for winter PM₁₀, attributable primarily to uncertainties in residential heating emissions and boundary-layer parameterizations [95].

The extreme PM pollution episode over the Po Valley in February 2024 illustrates the operational value of multi-model ensembles: the CAMS 11-model ensemble forecast PM₁₀ daily maxima above 150 $\mu\text{g m}^{-3}$ and PM_{2.5} above 100 $\mu\text{g m}^{-3}$ in the Milan area [96], forecasts that informed public health advisories across Northern Italy and the Balkans. The ensemble spread provided regulators with a direct visualization of forecast uncertainty, a capability that single-model deterministic forecasts cannot provide.

11.2. Naples and the Campania Region: Ensemble Trajectory Classification and Source Apportionment

An early contribution to understanding this complexity was the application of an ensemble approach to backward-trajectory classification. Riccio et al. [97] used an ensemble of back-trajectories computed with HYSPLIT for different arrival heights and positions around Naples (1995–2004), combined with *k*-means and PCA clustering following trajectory-classification methodologies widely adopted in atmospheric transport studies [98], to classify eight dominant airflow regimes and their associations with ozone and PM₁₀ levels. This ensemble trajectory approach demonstrated that stagnation and recirculation under anti-cyclonic summer conditions simultaneously elevated both ozone and PM, and identified North African contributions as drivers of peak PM₁₀ episodes.

A complementary modeling framework based on high-resolution meteorological and chemical transport simulations was later developed for the Naples–Caserta region. Using WRF at 1 km resolution, a multi-physics ensemble of seven PBL schemes revealed that TKE closure schemes (especially MYNN2) consistently outperformed first-order closures, consistent with previous WRF intercomparison studies [32,99].

More recently, Giarra et al. [100] extended the trajectory-based approach over a six-year period (2018–2023) using a residence time analysis based on HYSPLIT back-trajectories and source-receptor concepts closely related to Potential Source Contribution Function (PSCF) methodologies [101], identifying seven airflow clusters and documenting seasonal dynamics in PM₁₀. Elevated winter levels were attributed to residential heating and temperature inversions, while summer peaks reflected Saharan dust transport and secondary aerosol formation [100].

The state-of-the-art source apportionment study for Naples [102] used a high-resolution WRF–CAMx–PSAT modeling framework over the Campania region to attribute NO₂, PM_{2.5}, PM₁₀, and SO₂ to major emission sectors. Road transport and residential heating dominated NO₂, PM₁₀, and PM_{2.5} in urban and suburban settings, while maritime emissions significantly impacted SO₂ and NO₂ at coastal sites near the Port of Naples. Secondary organic aerosols and sea salt contributed non-negligibly to PM levels, especially at inland locations. Such multi-source attribution, only possible with high-resolution coupled modeling frameworks, provides actionable information for targeted emission controls [102,103].

The related mobility-scenario analysis for Naples [103] used CAMx to test sector-specific interventions against a 2017 baseline. Shore-side electrification of ships at berth reduced annual port emissions by about 18% for NO_x, 7% for SO₂, and 12% for PM₁₀, with

local NO₂ reductions near the docks exceeding 10%. For road traffic, targeted retirement of the highest-emitting vehicles was more effective than proportional fleet downsizing, reducing NO_x, PM₁₀, and PM_{2.5} emissions by roughly 37%, 43%, and 49%, respectively, and lowering modeled NO₂ by more than 2.3 ppbV in dense traffic areas.

11.3. A Synthesis: Lessons from Italian Case Studies

The Italian case studies point to several transferable lessons for ensemble air quality modeling in complex environments:

- Ensemble diversity is not only a theoretical construct but has measurable consequences: the CAMS ensemble spread is a reliable predictor of forecast uncertainty in the Po Valley, and reduced non-redundant ensembles achieve performance comparable to the full system [29].
- Hierarchical Bayesian data fusion can effectively reconcile the scale mismatch between grid-averaged CTM outputs and point observations, and dramatically improve gridded analysis products for exposure assessment [49].
- Ensemble trajectory approaches, used in combination with statistical clustering, provide an interpretable bridge between meteorological variability and observed pollutant distributions, informing both source apportionment and the design of ensemble experiments [97,100].
- The Italian examples illustrate two complementary uses of CTM-based modeling. High-resolution WRF–CAMx–PSAT simulations support source attribution and scenario analysis in coastal urban environments, whereas the national CAMS–MLP PM₁₀ application demonstrates that neural post-processing can correct seasonal CTM bias and improve short-range forecasts without rerunning the CTM. Thus, the evidence for neural post-processing comes from the CAMS–MLP post-processing study, while the Naples/Campania studies illustrate source attribution and control-scenario analysis [47,102,103].

12. Explainability and Interpretability in Ensemble Air Quality Models

Interpretability is a critical requirement for regulatory adoption of model predictions. Deep learning ensembles, while achieving the highest accuracy, often operate as black boxes, limiting stakeholder trust and obscuring the physical drivers of predicted pollutant events. SHAP (SHapley Additive exPlanations) [104] has emerged as the dominant interpretability tool for tree-based ensembles in air quality applications: it assigns to each feature a contribution to the prediction consistent with game-theoretic fairness axioms, and can reveal interactions between meteorological and emission variables that drive AQI fluctuations [41,105]. In the Delhi SERM study, SHAP analysis identified PM_{2.5} and CO as the dominant drivers of AQI fluctuations, supporting the use of explainable stacked ensembles as policy-facing forecasting tools [105].

Physics-informed neural networks (PINNs) represent a structurally different route to interpretability, by embedding known physical laws directly into the model architecture rather than appending post hoc explanations [57]. Hybrid physics–explainable AI frameworks that combine PINNs with SHAP or attention-based attribution have been proposed for scalable AQI prediction, simultaneously achieving high accuracy and physical consistency [60].

The CAMS National Collaboration Programme Italy, launched in its second phase in 2024, illustrates how operational ensemble systems are evolving towards greater interpretability and downscaling capability: Italian regional models nested within the CAMS ensemble are being used to provide local-scale forecasts with explicit uncertainty quantification, including the detection and attribution of Saharan dust episodes [106].

13. Benefits and Challenges

13.1. Documented Benefits

CTM ensemble studies, systematic reviews of ML/DL applications, and applications in complex environments document the following recurring benefits, with the magnitude of each benefit depending on validation design, pollutant, spatial scale, and data availability:

- Improved accuracy and robustness: multi-model ensembles consistently reduce RMSE and bias relative to individual models across pollutants and geographic regions [5,10,12,13]. The improvement is most pronounced when base models are diverse and when sufficient observational data are available for training or post-processing.
- Uncertainty quantification: ensemble spread and probabilistic formulations (BMA, conformal prediction, deep ensemble variance) provide calibrated uncertainty estimates that inform risk-based decisions [50,70,72,107].
- Resilience to model failure: combining diverse models reduces the risk that a single member's failure (due to parametric error, unusual meteorology, or missing data) will catastrophically degrade overall service quality [5,10].
- Improved impact assessment: for neural-network-based systems, deep ensembles provide a simple and scalable approach to predictive uncertainty estimation, often producing well-calibrated uncertainty without requiring fully Bayesian training procedures [108]. In air quality modeling, multi-model ensembles reduce uncertainty in derived quantities—health impacts, crop losses, exposure indices—to a much greater degree than could be achieved by single-model improvements, because ensemble combination addresses structural model uncertainty rather than merely parametric uncertainty [76].

13.2. Key Challenges and Open Problems

Despite these advantages, several challenges limit the routine operational deployment of advanced ensemble methods:

- Computational cost: multi-model CTM ensembles require running and archiving outputs from multiple independent modeling systems at high resolution; deep learning ensembles require training multiple large models. Efficient ensemble design, model pruning, and knowledge distillation are therefore essential [29].
- Interpretability: deep and hybrid ensembles often operate as black boxes. Physics-informed architectures and SHAP-based post hoc attribution partially address this, but remain incomplete solutions [41,57].
- Data quality and representativeness: ensemble performance depends critically on the quality of training and validation data, including the spatial representativeness of monitoring stations, the accuracy of emission inventories, and the completeness of meteorological inputs [12,49,73].
- Transferability and external validity: ensembles trained in one region or period may generalize poorly to other climates, emission mixtures, regulatory networks, and monitoring densities. This limitation is particularly important for low- and middle-income regions and data-poor environments, where sparse observations make purely local training unstable. Domain adaptation, meta-learning, hierarchical pooling, and physics-informed transfer learning are therefore active areas of investigation [12,13].
- Dependence structures: models sharing inputs or training data have correlated errors; ignoring this leads to overconfident uncertainty estimates and suboptimal weighting [29,70].

- Extreme-event skill: severe pollution episodes are disproportionately important for public-health interventions, yet many ML models, especially autoregressive or average-loss-optimized models, underpredict peaks because extreme events are rare in the training distribution. Event-weighted losses, quantile or conformal prediction, rare-event resampling, meteorological-regime conditioning, and CTM-guided predictors are promising strategies for improving high-concentration forecasts.
- Standardization: there is no universally accepted benchmark protocol for evaluating air quality ensemble models, making cross-study comparisons difficult [12,109].

14. Research Gaps and Future Directions

Table 3 summarizes the most important research gaps identified in this review, mapped against the type of ensemble approach and application domain. The gap labels are qualitative: they combine the frequency with which limitations were reported in systematic reviews and case studies, the presence or absence of validated applications in each method–domain cell, and expert judgment on operational maturity. They should therefore be read as a transparent synthesis framework rather than as a formal risk-of-bias score, formal study-quality grade, or bibliometric count.

Table 3. Research gaps matrix: ensemble method categories vs. application domains. Cov.: covered (studies exist); Part.: partially covered; Gap: no or minimal literature.

Method /Domain	Urban AQ	Regional	Indoor AQ	Federated	Health/Agri	Explainability
Multi-model CTM	Cov.	Cov.	Gap	Gap	Cov.	Part.
Tree ensembles	Cov.	Part.	Part.	Part.	Part.	Cov.
DL hybrid ensembles	Cov.	Part.	Gap	Part.	Gap	Part.
Physics-informed	Part.	Part.	Gap	Gap	Gap	Part.
Transferability	Part.	Part.	Gap	Gap	Gap	Gap
Extreme-event forecasting	Part.	Part.	Gap	Part.	Part.	Part.
Std. benchmarks	Part.	Part.	Gap	Gap	Gap	Gap

Priority research directions include:

Interpretable ensembles Developing frameworks that achieve state-of-the-art accuracy while providing mechanistic insight—through hybrid physics–ML architectures, attention-based attribution, or causal inference—is essential for regulatory adoption and scientific understanding.

Transferable and adaptive ensembles Most current ensembles are trained on data from specific regions or time periods. Robust cross-region transfer learning, meta-learning, and online adaptation strategies would dramatically extend applicability, particularly in regions with sparse monitoring networks.

Standardised evaluation protocols Establishing shared benchmark datasets and evaluation metrics comparable to what AQMEII achieved for CTMs would accelerate progress in ML-based ensemble development. Such protocols should explicitly distinguish random, temporal, spatial, and external validation, and should report pollutant, forecast horizon, station density, training period, and severe-event skill.

Forecasting severe pollution episodes Improving the prediction of high-impact episodes is a priority because exceedance warnings support immediate public-health interventions. Future ensemble systems should evaluate peak-specific metrics, combine CTM regime information with ML flexibility, and use training objectives that do not sacrifice rare high-concentration events for average error reduction.

Federated ensemble scaling Scaling federated ensemble architectures to national and continental sensor networks, handling non-IID data distributions, communication efficiency, node failures, and local personalization, is a key challenge for next-generation distributed air quality services. The added complexity is justified mainly when multi-

operator data access, edge inference, or bandwidth constraints make centralized learning impractical.

Integration with health and climate models Embedding ensemble air quality predictions into integrated assessment chains—from emissions to concentrations to health outcomes—with explicit uncertainty propagation, is needed for credible policy support [76].

Next-generation operational systems Seamless coupling of CTMs with Artificial Intelligence of Things (AIoT) architectures, in which AI-as-a-Service supports network self-organization and the orchestration of environmental micro-services [110].

15. Conclusions

Multi-model ensemble approaches—spanning coordinated CTM systems, tree-based and deep learning ensembles, and hybrid data-assimilation frameworks—have become central to modern air quality prediction. Across the evidence reviewed here, they often outperform single models in accuracy, robustness, and uncertainty quantification when evaluated within comparable settings; however, the magnitude of the gain is context-specific and should not be interpreted as a formal pooled effect size or as the result of a formal risk-of-bias assessment. The theoretical basis for this behavior is well established in the bias–variance framework and in the Bayesian theory of model combination, and is supported by large-scale coordinated intercomparisons such as AQMEII and by operational systems such as CAMS.

Diversity and skill must be considered jointly but not conflated when designing ensembles—a principle formalized within the AQMEII framework and consistent with broader ensemble-learning theory. Structural dependencies among CTMs and redundant ML base learners alike limit the achievable performance gain; principled ensemble design, based on information-theoretic diversity measures, Bayesian weighting, and data-driven post-processing, is essential to realize the full potential of multi-model systems.

Operational case studies in the Po Valley and the Naples/Campania region illustrate how these principles translate into practice: Bayesian hierarchical data fusion of CAMS ensemble outputs reduces PM₁₀ bias by two-thirds and achieves out-of-sample R^2 above 0.83; multi-physics WRF ensembles reveal the dominant role of meteorological uncertainty in boundary layer parameterization; and high-resolution CTM source apportionment provides actionable guidance on dominant pollutant contributors.

Reviews of deep learning stress the need to represent temporal, spatial, and spatio-temporal correlations explicitly; WRF sensitivity experiments confirm that meteorological uncertainty is a first-order control on air-quality modeling; and explainable stacked ensembles show that predictive skill becomes most valuable when it can be translated into feature-level evidence for decision-makers.

The convergence of physics-informed neural networks, explainable AI, federated learning, and next-generation dynamical CTMs opens new directions for ensemble design. Combining physically consistent, interpretable, and computationally efficient predictions with rigorous uncertainty quantification—including reliable forecasts of severe pollution episodes—will be central to translating air quality science into actionable public health and climate policy.

Author Contributions: Conceptualization, A.R.; methodology, A.R.; software, A.R.; validation, A.R.; formal analysis, A.R.; investigation, A.R.; writing—original draft preparation, A.R. and E.C.; writing—review and editing, A.R. and E.C.; visualization, A.R.; supervision, A.R. and E.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors acknowledge the use of CAMS ensemble products and monitoring data from Regional Agencies for the Protection of Environment, as well as data and model outputs from the AQMEII initiative coordinated by the European Commission Joint Research Centre. The work of A.R. on multi-model ensembles was partly made within the AQMEII framework.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and health impacts of air pollution: A review. *Front. Public Health* **2020**, *8*, 14. <https://doi.org/10.3389/fpubh.2020.00014>.
- Ghorani-Azam, A.; Riahi-Zanjani, B.; Balali-Mood, M. Effects of air pollution on human health and practical measures for prevention in Iran. *J. Res. Med. Sci.* **2016**, *21*, 65. <https://doi.org/10.4103/1735-1995.189646>.
- Kumar, V.; S, H.; Huligowda, L.K.D.; Umesh, M.; Chakraborty, P.; Thazeem, B.; Singh, A.P. Environmental pollutants as emerging concerns for cardiac diseases: A review on their impacts on cardiac health. *Biomedicines* **2025**, *13*, 241. <https://doi.org/10.3390/biomedicines13010241>.
- Zhang, Y. Online-coupled meteorology and chemistry models: History, current status, and outlook. *Atmos. Chem. Phys.* **2008**, *8*, 2895–2932. <https://doi.org/10.5194/acp-8-2895-2008>.
- Baklanov, A.; Schlünzen, K.H.; Suppan, P.; Baldasano, J.M.; Brunner, D.; Aksoyoglu, S.; Carmichael, G.; Douros, J.; Flemming, J.; Forkel, R.; et al. Online coupled regional meteorology chemistry models in Europe: current status and prospects. *Atmos. Chem. Phys.* **2014**, *14*, 317–398. <https://doi.org/10.5194/acp-14-317-2014>.
- Krishnamurti, T.N.; Kishtawal, C.M.; Zhang, Z.; LaRow, T.; Bachiochi, D.; Williford, E.; Gadgil, S.; Surendran, S. Multimodel ensemble forecasts for weather and seasonal climate. *J. Clim.* **2000**, *13*, 4196–4216. [https://doi.org/10.1175/1520-0442\(2000\)013<4196:MEFFWA>2.0.CO;2](https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2).
- Knutti, R.; Sedláček, J.; Sanderson, B.M.; Lorenz, R.; Fischer, E.M.; Eyring, V. A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.* **2017**, *44*, 1909–1918. <https://doi.org/10.1002/2016GL072012>.
- Abramowitz, G.; Herger, N.; Gutmann, E.; Hammerling, D.; Knutti, R.; Leduc, M.; Lorenz, R.; Pincus, R.; Schmidt, G.A. ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth Syst. Dyn.* **2019**, *10*, 91–105. <https://doi.org/10.5194/esd-10-91-2019>.
- Peuch, V.; Engelen, R.; Rixen, M.; Dee, D.; Flemming, J.; Suttie, M.; Ades, M.; Agustí-Panareda, A.; Ananasso, C.; Andersson, E.; et al. The Copernicus Atmosphere Monitoring Service: From research to operations. *Bull. Am. Meteorol. Soc.* **2022**, *103*, E2650–E2668. <https://doi.org/10.1175/bams-d-21-0314.1>.
- Solazzo, E.; Bianconi, R.; Vautard, R.; Appel, K.W.; Moran, M.D.; Hogrefe, C.; Bessagnet, B.; Brandt, J.; Christensen, J.H.; Chemel, C.; et al. Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII. *Atmos. Environ.* **2012**, *53*, 60–74. <https://doi.org/10.1016/j.atmosenv.2012.01.003>.
- LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. <https://doi.org/10.1038/nature14539>.
- Agbehadj, I.E.; Obagbuwa, I.C. Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. *Atmosphere* **2024**, *15*, 1352. <https://doi.org/10.3390/atmos15111352>.
- Rybarczyk, Y.; Zalakeviciute, R. Machine learning approaches for outdoor air quality modelling: A systematic review. *Appl. Sci.* **2018**, *8*, 2570. <https://doi.org/10.3390/app8122570>.
- Garbagna, L.; Saheer, L.B.; Oghaz, M.M.D. AI-driven approaches for air pollution modelling: A comprehensive systematic review. *Environ. Pollut.* **2025**, *373*, 125937. <https://doi.org/10.1016/j.envpol.2025.125937>.
- Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
- Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. <https://doi.org/10.1006/jcss.1997.1504>.

19. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
20. Ansari, A.; Quaff, A.R. Comparative benchmarking of eleven machine learning regressors for daily air quality index (AQI) forecasting: Sensitivity analysis and SHapley Additive exPlanations (SHAP)-Based interpretability framework. *Theor. Appl. Climatol.* **2025**, *156*, 620. <https://doi.org/10.1007/s00704-025-05832-1>.
21. Sarkar, N.; Keserwani, P.K.; Ranjan, R.; Govil, M.C. Deep Learning and Federated Learning in Air Quality Forecasting: Trends, Insights, Challenges, and Future Perspectives. *Arch. Comput. Methods Eng.* **2026**, 1–40. <https://doi.org/10.1007/s11831-026-10555-1>.
22. Jiang, J.C.; Kantarci, B.; Oktug, S.; Soyata, T. Federated learning in smart city sensing: Challenges and opportunities. *Sensors* **2020**, *20*, 6230. <https://doi.org/10.3390/s20216230>.
23. Ramadan, M.N.; Ali, M.A.; Jaber, H.; Alkhedher, M. Blockchain-secured IoT-federated learning for industrial air pollution monitoring: A mechanistic approach to exposure prediction and environmental safety. *Ecotoxicol. Environ. Saf.* **2025**, *300*, 118442. <https://doi.org/10.1016/j.ecoenv.2025.118442>.
24. Alwabli, A. Federated learning for privacy-preserving air quality forecasting using IOT sensors. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 16069–16076. <https://doi.org/10.48084/etasr.7820>.
25. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. <https://doi.org/10.1007/BF00058655>.
26. Gryech, I.; Boulahya, M.; Ghogho, M. Machine learning and Internet of Things for air quality monitoring: A review. *arXiv* **2024**, arXiv:2401.01788.
27. Zhang, B.; Rong, Y.; Yong, R.; Qin, D.; Li, M.; Zou, G.; Pan, J. Deep learning for air pollutant concentration prediction: A review. *Atmos. Environ.* **2022**, *290*, 119347. <https://doi.org/10.1016/j.atmosenv.2022.119347>.
28. Galmarini, S.; Bianconi, R.; Klug, W.; Mikkelsen, T.; Addis, R.; Andronopoulos, S.; Astrup, P.; Baklanov, A.; Bartniki, J.; Bartnicki, J.; et al. Ensemble dispersion forecasting, Part I: Concept, approach and indicators. *Atmos. Environ.* **2004**, *38*, 4607–4617. <https://doi.org/10.1016/j.atmosenv.2004.05.030>.
29. Solazzo, E.; Riccio, A.; Kioutsioukis, I.; Galmarini, S. Pauci ex tanto numero: Reduce redundancy in multi-model ensembles. *Atmos. Chem. Phys.* **2013**, *13*, 8315–8333. <https://doi.org/10.5194/acp-13-8315-2013>.
30. Riccio, A.; Ciamarella, A.; Giunta, G.; Galmarini, S.; Solazzo, E.; Potemski, S. On the systematic reduction of data complexity in multimodel atmospheric dispersion ensemble modeling. *J. Geophys. Res. Atmos.* **2012**, *117*, D05314. <https://doi.org/10.1029/2011JD016503>.
31. Colette, A.; Collin, G.; Besson, F.; Blot, E.; Guidard, V.; Meleux, F.; Royer, A.; Petiot, V.; Miller, C.; Fermond, O.; et al. Copernicus Atmosphere Monitoring Service–Regional Air Quality Production System v1.0. *Geosci. Model Dev.* **2025**, *18*, 6835–6883. <https://doi.org/10.5194/gmd-18-6835-2025>.
32. Tyagi, B.; Magliulo, V.; Finardi, S.; Gasbarra, D.; Carlucci, P.; Toscano, P.; Zaldei, A.; Riccio, A.; Calori, G.; D’Allura, A.; et al. Performance analysis of planetary boundary layer parameterization schemes in WRF modeling set up over Southern Italy. *Atmosphere* **2018**, *9*, 272. <https://doi.org/10.3390/atmos9070272>.
33. Luecken, D.; Phillips, S.; Sarwar, G.; Jang, C. Effects of using the CB05 vs. SAPRC99 vs. CB4 chemical mechanism on model predictions: Ozone and gas-phase photochemical precursor concentrations. *Atmos. Environ.* **2008**, *42*, 5805–5820. <https://doi.org/10.1016/j.atmosenv.2007.08.056>.
34. Kim, Y.; Sarwar, G.; Appel, K.W.; Napelenok, S.L.; Bash, J.O.; Fahey, K.M.; Pouliot, G.; Gilliam, R.C.; Hogrefe, C. Evaluation of different chemical mechanisms on O₃ and PM_{2.5} in the CMAQ model. *Appl. Sci.* **2022**, *12*, 8576. <https://doi.org/10.3390/app12178576>.
35. Zhang, B.; Wang, Y.; Hao, J. Performance evaluation of the WRF-Chem model with different physical parameterization schemes during an extremely high PM_{2.5} pollution episode in Beijing. *Aerosol Air Qual. Res.* **2016**, *16*, 2177–2192. <https://doi.org/10.4209/aaqr.2015.10.0610>.
36. Singh, K.P.; Gupta, S.; Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmos. Environ.* **2013**, *80*, 426–437. <https://doi.org/10.1016/j.atmosenv.2013.08.023>.
37. Hu, X.; Belle, J.H.; Meng, X.; Wildani, A.; Waller, L.A.; Strickland, M.J.; Liu, Y. Estimating PM_{2.5} concentrations in the conterminous United States using the random forest approach. *Environ. Sci. Technol.* **2017**, *51*, 6936–6944. <https://doi.org/10.1021/acs.est.7b01210>.
38. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. An ensemble-based model of PM_{2.5} concentration across the contiguous United States with high spatiotemporal resolution. *Environ. Int.* **2019**, *130*, 104909. <https://doi.org/10.1016/j.envint.2019.104909>.
39. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.S.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate PM_{2.5} concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ.* **2018**, *636*, 52–60. <https://doi.org/10.1016/j.scitotenv.2018.04.251>.

40. Özüpak, Y.; Alpsalaz, F.; Aslan, E. Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction. *Water Air Soil Pollut.* **2025**, *236*, 464. <https://doi.org/10.1007/s11270-025-08122-8>.
41. Singh, S.; Kumar, M.; Sengar, V.; Kumar, A.; Abhishek, K.; Shafeeq, B.A. Ensemble learning for air quality index prediction: Integrating gradient boosting, XGBoost, and stacking with SHAP-based interpretability. *Sci. Rep.* **2026**, *16*, 8544. <https://doi.org/10.1038/s41598-026-39232-w>.
42. Shahriar, S.A.; Kayes, I.; Hasan, K.; Hasan, M.; Islam, R.; Awang, N.R.; Hamzah, Z.; Rak, A.E.; Salam, M.A. Potential of ARIMA-ANN, ARIMA-SVM, DT and CatBoost for atmospheric PM_{2.5} forecasting in Bangladesh. *Atmosphere* **2021**, *12*, 100. <https://doi.org/10.3390/atmos12010100>.
43. Duan, W.; Wang, L.; Cheng, S.; Wang, X.; Zhang, H. A novel hybrid-GARCH model based on ARIMA and SVM for PM_{2.5} concentrations forecasting. *Atmos. Pollut. Res.* **2017**, *8*, 850–860. <https://doi.org/10.1016/j.apr.2017.01.003>.
44. Liu, H.; Mi, X.; Li, Y. A hybrid deep learning technology for PM_{2.5} air quality forecasting. *Environ. Sci. Pollut. Res.* **2021**, *28*, 39409–39422. <https://doi.org/10.1007/s11356-021-12657-8>.
45. Qi, C.; Huang, J.; Wang, B.; Zhou, J.; Wang, W. A hybrid algorithm for short-term forecasting of PM_{2.5} in China. *Atmos. Environ.* **2019**, *200*, 264–279. <https://doi.org/10.1016/j.atmosenv.2018.11.049>.
46. Naizabayeva, L.; Sembina, G.; Aliman, A.; Satymbekov, M.; Barlykbay, N.; Seilova, N. Air pollution forecasting in Almaty using ensemble machine learning models. *J. Appl. Data Sci.* **2025**, *6*, 2461–2476. <https://doi.org/10.47738/jads.v6i4.821>.
47. Chianese, E.; Camastra, F.; Ciaramella, A.; Landi, T.C.; Staiano, A.; Riccio, A. Spatio-temporal learning in predicting ambient particulate matter concentration by multi-layer perceptron. *Ecol. Inform.* **2019**, *49*, 54–61. <https://doi.org/10.1016/j.ecoinf.2018.12.001>.
48. Riccio, A.; Chianese, E. Accurate, reliable, and high-resolution air quality predictions by improving the Copernicus Atmosphere Monitoring Service using a novel statistical post-processing method. *Atmos. Chem. Phys.* **2024**, *24*, 1673–1689. <https://doi.org/10.5194/acp-24-1673-2024>.
49. Riccio, A.; Barone, G.; Chianese, E.; Giunta, G. A hierarchical Bayesian approach to the spatio-temporal modeling of air quality data. *Atmos. Environ.* **2006**, *40*, 554–566. <https://doi.org/10.1016/j.atmosenv.2005.09.070>.
50. Riccio, A.; Giunta, G.; Galmarini, S. Seeking for the rational basis of the Median Model: The optimal combination of multi-model ensemble results. *Atmos. Chem. Phys.* **2007**, *7*, 6085–6098. <https://doi.org/10.5194/acp-7-6085-2007>.
51. Casciaro, G.; Cavaola, M.; Mazzino, A. Calibrating the CAMS European multi-model air quality forecasts for regional air pollution monitoring. *Atmos. Environ.* **2022**, *287*, 119259. <https://doi.org/10.1016/j.atmosenv.2022.119259>.
52. Fioravanti, G.; Cameletti, M.; Martino, S.; Cattani, G. Spatio-temporal modelling of PM₁₀ daily concentrations in Italy using the SPDE approach. *Atmos. Environ.* **2021**, *248*, 118192. <https://doi.org/10.1016/j.atmosenv.2021.118192>.
53. Wang, S.; Li, Y.; Zhang, J.; Meng, Q.; Meng, L.; Gao, F. PM_{2.5}-GNN: A domain knowledge enhanced graph neural network for PM_{2.5} forecasting. In Proceedings of the 28th International Conference on Advances in Geographic Information Systems, Virtual, 3–6 November 2020; pp. 163–166. <https://doi.org/10.1145/3397536.3422208>.
54. Raja, S.; Damodaran, A.; Manohar, G. Enhanced forecasting of air quality index through an integrated deep learning framework. *Environ. Sci. Pollut. Res. Int.* **2025**, *32*, 30503–30516. <https://doi.org/10.1007/s11356-025-37269-4>.
55. Madan, T.; Sagar, S.; Singh, Y.; Jain, A.; Krejcar, O.; Frischer, R.; Namazi, H.; Panwar, A. Hybrid deep learning model for air quality prediction and its impact on healthcare. *Sci. Rep.* **2026**, *16*, 6036. <https://doi.org/10.1038/s41598-026-36564-5>.
56. Kalaiselvi, S.; Anitha, V.; Manimaran, V.; Lawrence, T.S. Air quality prediction using multi-source remote sensing data integration with hybrid deep learning framework. *Sci. Rep.* **2025**, *16*, 2688. <https://doi.org/10.1038/s41598-025-32466-0>.
57. Hettige, K.H.; Ji, J.; Xiang, S.; Long, C.; Cong, G.; Wang, J. AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Quality Prediction. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
58. Zhang, X.; Wang, S.; Jin, G.; Yuan, Z.; Yuan, H.; Ruan, S. Eulerian neural network informed by chemical transport for air quality forecasting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2025), San Diego, CA, USA, 2–7 December 2025.
59. Liang, Y.; Xu, R.; Chen, P.; Guo, C.; Zhou, A.; Pan, L.; Rao, Z.; Yang, B. Air quality prediction with physics-guided dual neural odes in open systems. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; Volume 2025, pp. 55132–55154.
60. Chandrashekar, S.V.; Khan, F.; Sridhar, S.; Dhanaraj, R.K.; Balusamy, B.; Pamucar, D.; Sathyamoorthy, M. A hybrid physics-informed neural and explainable AI approach for scalable and interpretable AQI predictions. *MethodsX* **2025**, *15*, 103597. <https://doi.org/10.1016/j.mex.2025.103597>.
61. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *J. Comput. Phys.* **2019**, *378*, 686–707. <https://doi.org/10.1016/j.jcp.2018.10.045>.

62. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the Artificial Intelligence and Statistics. PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
63. Nguyen, D.V.; Zettsu, K. Spatially-distributed federated learning of convolutional recurrent neural networks for air pollution prediction. In Proceedings of the 2021 IEEE International Conference on Big Data, Virtual, 15–18 December 2021; pp. 3601–3608. <https://doi.org/10.1109/BigData52589.2021.9671336>.
64. Krogh, A.; Vedelsby, J. Neural Network Ensembles, Cross Validation, and Active Learning. In *Proceedings of the Advances in Neural Information Processing Systems*; Tesauro, G., Touretzky, D., Leen, T., Eds.; MIT Press: Cambridge, MA, USA, 1994; Volume 7.
65. Dietterich, T.G. Ensemble Methods in Machine Learning. *Lect. Notes Comput. Sci.* **2000**, 1857, 1–15. https://doi.org/10.1007/3-540-45014-9_1.
66. Gong, Z.; Zhong, P.; Hu, W. Diversity in machine learning. *IEEE Access* **2019**, 7, 64323–64350. <https://doi.org/10.1109/ACCESS.2019.2917620>.
67. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; CRC Press: Boca Raton, FL, USA, 2012.
68. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and El George, and a rejoinder by the authors). *Stat. Sci.* **1999**, 14, 382–417. <https://doi.org/10.1214/ss/1009212519>.
69. Raftery, A.E.; Gneiting, T.; Balabdaoui, F.; Polakowski, M. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **2005**, 133, 1155–1174. <https://doi.org/10.1175/MWR2906.1>.
70. Potemski, S.; Galmarini, S. Est modus in rebus: Analytical properties of multi-model ensembles. *Atmos. Chem. Phys.* **2009**, 9, 9471–9489. <https://doi.org/10.5194/acp-9-9471-2009>.
71. Copernicus Atmosphere Monitoring Service. European Air Quality Forecast Plots. 2022. Available online: <https://atmosphere.copernicus.eu/european-air-quality-forecast-plots> (accessed on 25 May 2026).
72. Copernicus Atmosphere Monitoring Service. The CAMS European Air Quality Ensemble Forecasts Welcomes Two New State-of-the-Art Models, 2022. Available online <https://atmosphere.copernicus.eu/cams-european-air-quality-ensemble-forecasts-welcomes-two-new-state-art-models> (accessed on 25 May 2026).
73. Guevara, M.; Colette, A.; Guion, A.; Petiot, V.; Adani, M.; Arteta, J.; Benedictow, A.; Bergström, R.; Bolignano, A.; Camps, P.; et al. Sensitivity of the CAMS regional air quality modelling system to anthropogenic emission temporal variability. *Atmos. Chem. Phys.* **2025**, 25, 13245–13278. <https://doi.org/10.5194/acp-25-13245-2025>.
74. Kuncheva, L.I.; Whitaker, C.J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* **2003**, 51, 181–207. <https://doi.org/10.1023/A:1022859003006>.
75. Kioutsioukis, I.; Hogrefe, C.; Makar, P.A.; Alyuz, U.; Bash, J.O.; Bellasio, R.; Bianconi, R.; Butler, T.; Clifton, O.E.; Cheung, P.; et al. Operational, diagnostic, and probabilistic evaluation of AQMEII-4 regional-scale ozone dry deposition: Time to harmonize our LULC masks. *Atmos. Chem. Phys.* **2025**, 25, 12923–12953. <https://doi.org/10.5194/acp-25-12923-2025>.
76. Solazzo, E.; Riccio, A.; Van Dingenen, R.; Valentini, L.; Galmarini, S. Evaluation and uncertainty estimation of the impact of air quality modelling on crop yields and premature deaths using a multi-model ensemble. *Sci. Total Environ.* **2018**, 633, 1437–1452. <https://doi.org/10.1016/j.scitotenv.2018.03.317>.
77. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
78. Zhu, X.; Li, B.; Cao, Y.; Zhang, Q. Applicability analysis of tree-based ensemble learning for air pollutant prediction models. *Sci. Rep.* **2026**, 1, 9602. <https://doi.org/10.1038/s41598-025-32652-0>.
79. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, 63, 3–42. <https://doi.org/10.1007/s10994-006-6226-1>.
80. Ho, T.K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, 20, 832–844. <https://doi.org/10.1109/34.709601>.
81. Grange, S.K.; Carslaw, D.C. Using Meteorological Normalisation to Detect Interventions in Air Quality Time Series. *Sci. Total Environ.* **2019**, 653, 578–588. <https://doi.org/10.1016/j.scitotenv.2018.10.344>.
82. Sivhugwana, K.S.; Ranganai, E. An ensemble approach to short-term wind speed predictions using stochastic methods, wavelets and gradient boosting decision trees. *Wind* **2024**, 4, 3. <https://doi.org/10.3390/wind4010003>.
83. WHO. *WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; WHO: Geneva, Switzerland, 2021.
84. Sukprasert, S.; Thanasutives, P.; Araki, S.; Uranishi, K.; Matsuo, T.; Shimadera, H. A comparative study on CMAQ-based predictors and reanalysis AOD for improving PM_{2.5} estimation in Thailand using machine learning. *Atmos. Environ.* **2026**, 375, 122006. <https://doi.org/10.1016/j.atmosenv.2026.122006>.
85. Sukprasert, S.; Thanasutives, P.; Araki, S.; Uranishi, K.; Matsuo, T.; Shimadera, H. Machine learning based bias correction of CMAQ using emission source decoupled features in Thailand. *Geomate J.* **2026**, 30, 116–125. <https://doi.org/10.21660/2026.139.g15298>.
86. Du, S.; Li, T.; Yang, Y.; Horng, S.J. Deep air quality forecasting using hybrid deep learning framework. *IEEE Trans. Knowl. Data Eng.* **2019**, 33, 2412–2424. <https://doi.org/10.1109/TKDE.2019.2954510>.

87. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
88. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017. Available online: <https://arxiv.org/abs/1609.02907> (accessed on 30 May 2026).
89. Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J.E.; Weinberger, K.Q. Snapshot Ensembles: Train 1, Get M for Free. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017. Available online: <https://openreview.net/pdf?id=BJYwwY9ll> (accessed on 30 May 2026).
90. Sharma, H.; Shrivastava, M.; Singh, B. Physics-informed deep neural network embedded and coupled within WRF-Chem for atmospheric chemistry. *Npj Clim. Atmos. Sci.* **2023**, *6*, 28. <https://doi.org/10.1038/s41612-023-00353-y>.
91. Giunta, G.; Montella, R.; Mariani, P.; Riccio, A. Modeling and computational issues for air/water quality problems: A grid computing approach. *Il Nuovo Cimento C* **2005**, *28*, 215–224. <https://doi.org/10.1393/ncc/i2005-10184-3>.
92. Montella, R.; Brizius, A.; Di Luccio, D.; Porter, C.; Elliot, J.; Madduri, R.; Kelly, D.; Riccio, A.; Foster, I. Using the FACE-IT portal and workflow engine for operational food quality prediction and assessment: An application to mussel farms monitoring in the Bay of Napoli, Italy. *Future Gener. Comput. Syst.* **2020**, *110*, 453–467. <https://doi.org/10.1016/j.future.2018.03.002>.
93. Lu, Z.; Pan, H.; Dai, Y.; Si, X.; Zhang, Y. Federated learning with non-IID data: A survey. *IEEE Internet Things J.* **2024**, *11*, 19188–19209. <https://doi.org/10.1109/JIOT.2024.3376548>.
94. Montella, R.; Giunta, G.; Riccio, A. Using Grid Computing Based Components in On Demand Environmental Data Delivery. In Proceedings of the 2nd Workshop on Use of P2P, GRID and Agents for the Development of Content Networks (UPGRADE-CN'07), Monterey, CA, USA, 25 June 2007; pp. 81–86. <https://doi.org/10.1145/1272980.1272995>.
95. Stortini, M.; Arvani, B.; Deserti, M. Operational forecast and daily assessment of the air quality in Italy: A Copernicus-CAMS downstream service. *Atmosphere* **2020**, *11*, 447. <https://doi.org/10.3390/atmos11050447>.
96. Copernicus Atmosphere Monitoring Service. Extreme Episode of Particulate Matter Air Pollution Across Italy's Po Valley. Available online: <https://atmosphere.copernicus.eu/extreme-episode-particulate-matter-air-pollution-across-italys-po-valley>, 2024. (accessed on 23 May 2026).
97. Riccio, A.; Giunta, G.; Chianese, E. The application of a trajectory classification procedure to interpret air pollution measurements in the urban area of Naples (Southern Italy). *Sci. Total Environ.* **2007**, *376*, 198–214. <https://doi.org/10.1016/j.scitotenv.2007.01.068>.
98. Dorling, S.; Davies, T.; Pierce, C. Cluster analysis: A technique for estimating the synoptic meteorological controls on air and precipitation chemistry—Method and applications. *Atmos. Environment. Part A Gen. Top.* **1992**, *26*, 2575–2581. [https://doi.org/10.1016/0960-1686\(92\)90110-7](https://doi.org/10.1016/0960-1686(92)90110-7).
99. Hu, X.M.; Nielsen-Gammon, J.W.; Zhang, F. Evaluation of three planetary boundary layer schemes in the WRF model. *J. Appl. Meteorol. Climatol.* **2010**, *49*, 1831–1844. <https://doi.org/10.1175/2010JAMC2432.1>.
100. Giarra, A.; Riccio, A.; Chianese, E.; Annetta, M.; Toscanesi, M.; Trifuoggi, M. Transport mechanisms and pollutant dynamics influencing PM₁₀ levels in a densely urbanized and industrialized region near Naples: A residence time analysis. *Atmosphere* **2025**, *16*, 393. <https://doi.org/10.3390/atmos16040393>.
101. Hopke, P.; Barrie, L.; Li, S.M.; Cheng, M.D.; Li, C.; Xie, Y. Possible sources and preferred pathways for biogenic and non-sea-salt sulfur for the high Arctic. *J. Geophys. Res. Atmos.* **1995**, *100*, 16595–16603. <https://doi.org/10.1029/95JD01712>.
102. Lino, S.; Toscano, D.; Piccoli, A.; Riccio, A.; De Angelis, E.; Pirovano, G.; Agresti, V. Source apportionment of key air pollutants in Naples using a high-resolution WRF-CAMx-PSAT modeling framework. *Atmos. Environ.* **2026**, *364*, 121638. <https://doi.org/10.1016/j.atmosenv.2025.121638>.
103. Lino, S.; Toscano, D.; Riccio, A.; Pirovano, G.; Agresti, V. Mobility scenarios and their impact on air quality in the city of Naples. *City Environ. Interact.* **2026**, *30*, 100340. <https://doi.org/10.1016/j.cacint.2026.100340>.
104. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
105. Sriya, N.S.; Sowmya, M.; Lakshmi, A.R.H.; Amirtharajan, R. Explainable AI for urban air quality: SHAP interpretation of stacked ensemble AQI forecast. *Theor. Appl. Climatol.* **2025**, *156*, 533. <https://doi.org/10.1007/s00704-025-05741-3>.
106. Copernicus Atmosphere Monitoring Service. CAMS National Collaboration Programme Italy. 2022. Available online: <https://atmosphere.copernicus.eu/italy>. (accessed on 23 May 2026).
107. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. <https://doi.org/10.1198/016214506000001437>.
108. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.

109. Chang, J.C.; Hanna, S.R. Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **2004**, *87*, 167–196. <https://doi.org/10.1007/s00703-003-0070-7>.
110. Chochliouros, I.P.; Pages-Montanera, E.; Alcázar-Fernández, A.; Zahariadis, T.; Velivassaki, T.H.; Skianis, C.; Rossini, R.; Belesioti, M.; Drosos, N.; Bakiris, E.; et al. NEMO: Building the next generation meta operating system. In Proceedings of the 3rd Eclipse Security, AI, Architecture and Modelling Conference on Cloud to Edge Continuum, Ludwigsburg, Germany, 17 October 2023; pp. 1–9. <https://doi.org/10.1145/3624486.3624504>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.