

RESEARCH ARTICLE

EARTH SCIENCES

Fusing Physics and AI Enhances Accuracy and Diagnostic Capabilities in Air Quality Forecasting

Huangjian Wu^{1,2}, Zifa Wang^{*,1,2,3}, Yuanhang Zhang⁴, Wei Lin⁵, Xinwei Li¹, Meng Gao⁶, Hao Sun⁷, Dawei Zhang⁸, Wei Wang⁸, Xiao Tang^{1,2}, Jiang Zhu^{1,3}, Yu Song⁴, Chaoqun Ma⁹, Pingqing Fu¹⁰, Hang Su¹, Yafang Cheng⁹ & Gregory R. Carmichael¹¹

¹ State Key Laboratory of Atmospheric Environment and Extreme Meteorology, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

² Earth System Numerical Simulation Science Center, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

³ College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China

⁴ College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China

⁵ School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China

⁶ Department of Geography, Hong Kong Baptist University, Hong Kong SAR 999077, China

⁷ Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China

⁸ China National Environmental Monitoring Center, Beijing 100012, China

⁹ Aerosol Chemistry Department, Max Planck Institute for Chemistry, 55128 Mainz, Germany

¹⁰ Institute of Surface-Earth System Science, Tianjin University, Tianjin 300072, China

¹¹ Center for Global and Regional Environmental Research, University of Iowa, Iowa City, IA 52242, USA

* Corresponding author. E-mail: zifawang@mail.iap.ac.cn

26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

44
45
46

47 INTRODUCTION

Accurate and interpretable forecasting of airborne pollutants such as PM_{2.5} and O₃ is essential for protecting public health and informing environmental policies. Reliable predictions help individuals reduce exposures [1, 2], support timely warnings and emission-control actions by regulators [3], and inform the planning of intensive field campaigns for researchers [4]. Given the significant socioeconomic implications involved, it is crucial that forecasting systems deliver consistent performance across varied weather scenarios, reliably detect extreme pollution events, and offer transparent interpretations, allowing experts to trace pollutant origins and assess predictive uncertainty.

55 Process-based chemical transport models (CTMs), commonly employed in operational forecasting
56 systems worldwide [5], simulate transport, emissions, chemical reactions, and deposition of air
57 pollutants. Their explicit representation of atmospheric processes provides interpretability, facilitating
58 source apportionment to pinpoint the timing, location, and sectoral origins of pollutants [6] and
59 enabling attribution of air quality improvements to factors such as meteorological variability or
60 emission reductions [7]. CTMs also support diagnostic analysis of forecast failures by tracking
61 uncertainties in inputs and process parameterizations [8, 9], thereby guiding model improvements [10].
62 Nonetheless, CTM accuracy remains hampered by imperfect emission inventories [11], which alone
63 may account for over 90% of simulation bias in some cases [12], and incomplete knowledge of
64 physicochemical processes, leading to significant variations in predicted PM_{2.5} composition [13].

The rise of purely AI-based forecasting models has recently transformed numerical weather prediction (NWP) [14, 15], with several data-driven systems now rivaling or surpassing traditional physics-based alternatives [14]. In air quality forecasting, data-driven models have also shown promising skills across various applications [16], such as attention-based PM_{2.5} spatiotemporal

prediction [17] and transfer-learning-based O_3 prediction [18]. However, their operational and mechanism-oriented application remains more challenging than in NWP because air pollution is heavily influenced by highly uncertain anthropogenic emissions, complex nonlinear multiphase chemistry, the sparsity of observations for rare high-pollution events, and abrupt local control measures. Unlike weather systems driven largely by stable solar radiation, air pollution stems heavily from volatile and poorly quantified anthropogenic emissions [19, 20]. The combined challenges of emission uncertainties, complex chemical interactions, and limited observational data hinder the interpretability and generalizability of purely AI-based approaches, particularly for risk-sensitive applications or mechanism-focused studies [21]. Even advanced systems like Aurora [22], while showing encouraging skill for surface pollutants, remain constrained by limited training data, highly skewed pollutant-concentration distributions spanning clean background conditions to episodic high- $PM_{2.5}$ and high- O_3 events, and intricate precursor-emission patterns.

Hybrid approaches that integrate physical modeling with AI have thus attracted growing interest [23, 24]. Such integrations aim to enhance predictive skill while retaining interpretability and mechanistic insight [25]. Recent examples include deep learning enhanced with physical advection schemes (NowcastNet) [26] and differentiable solvers augmented with machine-learning modules (NeuralGCM) [27]. However, transitioning these emerging techniques to high-resolution atmospheric chemistry forecasting remains hindered by fundamental computational and mathematical bottlenecks. For instance, physics-informed ML relying on soft constraints [28] often struggles to converge when optimizing stiff chemical kinetics, while fully differentiable 3D CTMs [29] face prohibitive memory overheads for backpropagation. Consequently, in the domain of air quality forecasting, current frameworks mostly retreat to simpler integrations. AI-based postprocessing techniques do improve

predictive skill [30]; however, they provide little clarity about where, when, or through which processes prediction errors arise. Alternatively, replacing certain physical submodules with AI-based emulators can reduce computational expense [31], yet when these surrogates are trained exclusively on synthetic model outputs rather than real-world observations, they inevitably inherit the limits of their parent CTMs.

To overcome the persistent trade-off between accuracy and interpretability, we introduce ProNet: a tightly coupled forecasting framework that embeds observation-constrained neural correctors directly into the governing equations of CTM. Rather than treating AI as an external postprocessor or module replacement, ProNet employs trainable networks to systematically mitigate process-level uncertainties within the physical simulation, thereby curtailing error propagation, preserving physical consistency, and illuminating systemic model deficiencies. Evaluations across 338 Chinese cities confirm that ProNet exceeds existing physical and AI-only models in predicting $\text{PM}_{2.5}$ and O_3 generalizes robustly to extreme events, and surpasses forecaster-adjusted operational forecasts under the same lead-time and daily-mean evaluation protocol. This comparison provides a stringent human-in-the-loop benchmark because forecasters often improve upon raw CTM guidance during severe pollution episodes. More importantly, it identifies when and where conventional models fail, revealing critical gaps in current understandings of $\text{PM}_{2.5}$ formation mechanisms. By uniting physical modeling with trainable correctors in a unified architecture, ProNet provides both state-of-the-art forecasts and deeper insights into air quality forecasting, paving the way for a new generation of intelligible, high-fidelity Earth system modelling.

THE PRONET FRAMEWORK

The ProNet framework is designed based on the governing equation of air quality forecasting, which

could deduce that the CTM forecasting error can be decomposed into the error transported from the previous time step, and the newly generated error stemming from the uncertainties in the processes of CTM (see Materials and Methods). As shown in Fig. 1a, ProNet corrects the conventional CTM with an error transport model (ETM). The baseline CTM employed in this study is the Nested Air Quality Prediction Modeling System (NAQPMS), a widely adopted system in operational air quality forecasting [32]. Technical specifications and model configurations are detailed in Supplementary Information (SI) Section 2.1. The ETM dynamically corrects baseline predictions by numerically solving the governing equations for CTM errors. The ETM combines a process-based transport module that is methodologically aligned with the CTM (see SI Section 2.3) with two neural-network Correctors.

Specifically, ProNet features two neural correctors to address distinct error sources: the Emission-Deposition Corrector, which utilizes historical errors and meteorological states to correct general spatiotemporal uncertainties, and the Chemical-Reaction Corrector, which processes multi-species concentrations to correct $\text{PM}_{2.5}$ errors from chemical parameterizations (details in Supplementary Section 2.2). To train these correctors, the transport module isolates target uncertainty errors from overall CTM forecast errors by subtracting the computed transported error (Equation 2) from the total forecast error at each step (Fig. 1e).

The Ensemble Corrector detects and suppresses bias accumulation that would otherwise increase by orders of magnitude during forecasting. While structurally similar to the Emission-Deposition Corrector, it employs one spatial encoder and two distinct spatial decoders, one for ensemble weights and another for bias estimation (Fig. 1d). The final ProNet forecast is computed as the weighted average of ensemble forecasts plus the estimated ensemble bias. Further details regarding data

assimilation, network architectures, training techniques for the ProNet approach are elaborated in the SI Section 2-3.

Additionally, future meteorological forecast states (such as wind speed, wind direction, and boundary layer height) are incorporated as input features to help the correctors adjust to varying meteorological regimes. The framework's robustness against wind-field errors is primarily achieved by the 1-hour short-step cyclic data assimilation at the training stage, which resets advection displacement biases at each hourly step, and the hourly error decomposition that isolates transported errors from local uncertainties. Nevertheless, this physics-informed approach maintains the physical constraints of process-driven modules while offering new insights into these atmospheric processes.

In this framework, the diagnostic interpretability is distinctly defined across three levels. First, process-constrained correction is achieved by maintaining the physical consistency of mass transport. Second, interpretable error bookkeeping provides spatiotemporal source apportionment of forecast errors, highlighting specific regions and periods for priority model improvement. Third, for mechanistic inference, while ProNet cannot provide explicit chemical reaction pathways, it dissects statistical associations between component-specific uncertainties, including nitrate, sulfate, ammonia, SOA, and OC, and total $PM_{2.5}$ forecast errors at a sub-grid scale.

RESULTS

Interpretable analysis of the coupled model

The ProNet framework introduces an interpretable architecture that simultaneously addresses two critical challenges in air quality modeling, namely diagnosing the sources of forecasting errors and elucidating deficiencies in $PM_{2.5}$ formation mechanisms. Unlike conventional approaches that rely on costly sensitivity analyses or ambient $PM_{2.5}$ component observations [33], our coupled model

systematically quantifies both Uncertainty Errors and IC errors through an source apportionment technique [34] tagging error propagation, while also disentangling the complex physicochemical processes governing aerosol formation. This dual capability provides actionable insights for refining emission inventories and advancing mechanistic understanding.

Error apportionment analysis reveals the source of CTM errors across temporal and spatial scales, as illustrated by a pollution episode in Beijing's January 2023 (Fig. 2a-c). While the CTM successfully captured the pollution episode's temporal evolution with peak concentrations exceeding $100 \mu\text{g m}^{-3}$, it exhibited systematic overestimation (Fig. 2d). Temporal error apportionment reveals that the errors on 13 January mainly originated from accumulated Uncertainty Errors propagating from the prior four days (Fig. 2b). Spatial attribution analysis further attributes the 15 January $\text{PM}_{2.5}$ overestimation in Beijing to the comprehensive model forecast errors (confounding local emission uncertainties, meteorological transport biases, and boundary layer simulation errors) originating from the upwind Shanxi and Shandong provinces (Fig. 2c). By tracking regional source tags, we identified a transport trajectory where simulated upwind errors accumulated and propagated downwind (detailed in Supplementary Section 9 and Supplementary Video 1).

Concurrently, the model investigates $\text{PM}_{2.5}$ formation processes by isolating statistical correlations of component-specific uncertainties, including nitrate, sulfate, ammonia, secondary organic aerosols, and organic carbon, using observations of total $\text{PM}_{2.5}$ concentrations (technical details in SI Section 4). Experimental studies report uptake coefficients for NO_2 , nitrate radicals (NO_3), and dinitrogen pentoxide (N_2O_5) spanning orders of magnitude under varying conditions [35-37]. As demonstrated by Figs. 2d, f, elevated nitrate levels correlate with increased CTM forecast errors, suggesting a

potential systematic overestimation of nitrate formation. This aligns with observational evidence showing pronounced overestimation of nitrate levels in CTM simulations [38]. Crucially, this hypothesis is strongly supported by independent external validation from the MICS-Asia III multi-model comparison, which reveals that current CTMs systematically overestimate nitrate production (exhibiting artificially high nitric oxidation ratios) while underestimating sulfate concentrations across East Asia [39]. We hypothesize that this bias may arise from unrepresented mechanisms such as nonvolatile cations in fine-mode aerosols artificially elevating aerosol pH and enhancing nitrate partitioning [40], or the absence of nitryl chloride formation from dinitrogen pentoxide uptake in current CTMs [41]. Importantly, these specific chemical pathways are subjective hypotheses based on domain expertise and statistical correlations rather than direct causal discoveries of ProNet, as SHAP values indicate feature importance rather than definitive causation. Nevertheless, by providing spatiotemporally continuous correlations between precursors and simulation errors, ProNet serves as a practical diagnostic reference for screening new chemical hypotheses. Our analysis also reveals a systematic underestimation of sulfate formation, particularly at lower temperatures (Fig. 2e), alongside an overestimation of secondary organic aerosol production (Fig. 2d).

Analogous to conventional source attribution and chemical mechanism analysis based on CTM, ProNet's error attribution and chemical diagnostics provide valuable quantitative insights into process-level uncertainties under the standard assumption of prescribed meteorological inputs. A comprehensive uncertainty analysis (detailed in Supplementary Section 13, SI Fig. S14 and S15) evaluates the numerical stability of the diagnostic patterns identified by ProNet, providing a practical uncertainty reference for model refinement.

Mitigating bias accumulation through ensemble correction

Forecasting systems operate through recursive state predictions to extend forecast horizons. However, single-step model optimizations fail to ensure commensurate gains in multi-day forecasting accuracy. Even with a perfect model, small errors in the initial conditions can lead to widely diverging outcomes due to the chaotic behavior, also known as the butterfly effect [42]. While the ETM correctors reduce per-step uncertainties, they inadvertently induce small bias. Without proper handling, biases introduced at each time step accumulate in the total concentration, leading to substantial errors in long-lead-time forecasts. As shown in Fig. 3, the forecasts without Ensemble Corrector exhibit little bias at the beginning. However, the bias increases nearly linearly over the first few days, leading to an ensemble spread exceeding $30 \mu\text{g m}^{-3}$ for lead times beyond 3 days.

Deep learning-based forecasting models commonly face bias accumulation challenges stemming from two key sources. Firstly, intrinsic model errors emerge from data noise and optimization constraints. Secondly, the iterative state estimation process in multi-step forecasting progressively amplifies these single step errors. While bias mitigation is feasible, the bias-variance tradeoff [43] fundamentally constrains accuracy improvements. Furthermore, iterative effects during extended forecasting can magnify single step biases by orders of magnitude, leading to substantial forecast degradation. The Pangu model counters this by employing multi-timescale frameworks to reduce computational cycles [44], effectively limiting bias accumulation. However, these systems require training parallel models for different time steps, which introduces two limitations: computational overhead, and potential loss of high-frequency temporal patterns in coarse-grained models. To address this, we construct a 10-member ensemble of ETMs (using parameters from epochs 291–300) combined

relative diagnostics in Supplementary Section 8 to evaluate dynamical resemblance against the process-based CTM baseline: relative mass-variation resemblance, temporal smoothness, and transport realism (SI Fig. S4 and Supplementary Video 3). We clarify that these metrics evaluate relative physical resemblance to the CTM's dynamical core rather than absolute thermodynamic truth, as CTMs inherently incorporate discretization errors, sub-grid parameterizations, and continuous hourly mass fluxes from emissions and deposition. The evaluation reveals that the pure AI model exhibits a relative mass imbalance index of 0.15 (max 0.59) and a temporal variance deviation of 0.07 (max 0.33) relative to the CTM. In contrast, ProNet maintains a significantly closer dynamical resemblance to the CTM's physical core, reducing the relative mass imbalance to 0.11 (max 0.33) and temporal deviation to 0.03 (max 0.09). Moreover, ProNet achieves a transport realism correlation of 0.5 with the CTM-derived transport-only tendency fields, far exceeding the pure AI model (0.1). This stronger alignment with the CTM transport operator suggests physically coherent spatial transport, improving assimilation and forecasting under unseen scenarios.

Data assimilation techniques have been among the most important factors contributing to improved forecasting accuracy. With coupling, the ProNet approach assimilates observations into ICs and systematically reduces 1-day forecast RMSE across 97.3% of stations (Fig. 4a). As demonstrated in a sand and dust event case study (Fig. 4b-c and Supplementary Video 2), the pure-AI based model struggles to resolve short-term pollutant dynamics (≤ 24 hours), while the coupled system exhibits enhanced observational influence. The coupled system successfully predicts the upcoming high $\text{PM}_{2.5}$ concentrations in Beijing, achieving an RMSE 62% lower than that of the pure AI model.

The pure AI model optimizes an aggregated loss function across all training samples, performing well for clean days under common conditions. However, policymaking prioritizes accurate predictions

of polluted episodes and abrupt pollution changes, where our physics-AI coupled model demonstrates superior performance. It more accurately predicts daily concentration changes and achieves higher correlation coefficients across 91.4% of monitoring stations (Fig. 4d-f).

To test generalizability, we trained both models on data from three months (deliberately excluding winter). While performance was comparable, the pure AI model predicted national averaged $\text{PM}_{2.5}$ levels below $40 \mu\text{g m}^{-3}$ for all days, whereas the coupled model captured higher concentrations in months outside the training period (Fig. 4g-i). The superior generalizability of the coupled model is critical for air quality forecasting, because the compounding effects of dynamic policy adjustments (e.g., the Air Pollution Prevention and Control Action Plan [45]) and climate variability [46] render historical training data insufficient to cover future scenarios. More discussions about sensitivity experiments on training data scarcity are provided in Supplementary Section 11.

Furthermore, beyond data scarcity, we evaluated out-of-distribution generalization by applying models trained on 2022 data to predict 2023 conditions. The noticeable differences in pollution levels and distribution patterns between these two years provided a basis to assess model performance across various pollution levels, seasons, regions, and special events such as dust storms and festival emissions. ProNet showed better adaptability to unseen high-pollution periods compared to the Pure-AI model, although both models still faced challenges with sudden human-driven emission spikes like fireworks. Detailed stratification analyses are provided in Supplementary Section 14.

To justify the hybrid architecture, we evaluated an all-AI Corrector approach that corrects CTM outputs without a physical transport module. While incorporating CTM baseline simulation data achieves a significant performance improvement over Pure-AI models, its overall forecast performance remains inferior to ProNet (SI Fig. S21). Furthermore, the all-AI Corrector also

underperforms compared to ProNet in terms of assimilation data utilization, concentration change prediction, and performance under data-missing scenarios (SI Fig. S17, detailed in Supplementary Section 10).

Improved air quality forecasting

In ProNet, deep learning modules that correct uncertainties in CTMs were trained on 2022 forecasts and validated with 2023 data. These modules effectively capture spatial and temporal variations in Uncertainty Error (SI Fig. S3), reducing forecasting RMSE at 99.7% of cities and 99.2% days. While deep learning modules reduce Uncertainty Error by 8.2%, the coupled model achieves a 32.5% reduction in total error, indicating that the integration framework synergistically amplifies the improvements made by the deep learning modules.

An ablation study systematically evaluates the performance contributions of ProNet's key components (SI Fig. S21). Integrating data assimilation and physical transport reduces the 1–7 day forecast RMSE by 8.6% compared to raw CTM. Adding the neural correctors in the ETM yields an additional 13.1% error reduction. Finally, applying the Ensemble Corrector suppresses bias accumulation, contributing another 15.2% error reduction, with benefits becoming more pronounced at longer lead times.

To benchmark against SOTA global AI models, we evaluated Aurora over China for 2023 (SI Fig. S5 and Supplementary Section 12). Note that because Aurora does not output continuous hourly forecasts suitable for daily averages, this evaluation is based on hourly RMSE. Aurora's overall PM_{2.5} forecast error is comparable to the CTM, while our Pure-AI baseline outperforms both. ProNet further reduces the hourly forecast error to 25.6 $\mu\text{g m}^{-3}$, outperforming Aurora (32.6 $\mu\text{g m}^{-3}$) by 21%. While Aurora's use of global satellite data improves dust forecasting in northern regions, ProNet's physics-

coupled design and high-resolution station data yield superior accuracy across the vast majority of cities and lead times.

In China, human forecasters produce the official air quality forecasts by integrating CTM predictions with expert interpretation (see SI Section 6). This forecaster-adjusted product provides a stringent benchmark because it already improves upon raw CTM guidance during severe pollution episodes. To ensure a fair comparison, all methods are evaluated under the same D+1 to D+7 lead-time and daily-mean protocol, with hourly ProNet and CTM outputs aggregated to the same city-level daily means as the operational product.

As demonstrated in Fig. 5a, operational forecasts made by professional human forecasters across 338 cities achieve a $\text{PM}_{2.5}$ RMSE of $21.2 \mu\text{g m}^{-3}$ (14.1-18.9% lower than individual CTM). However, this expertise-driven approach presents inherent limitations: subjective interpretation challenges quantification, while manual processing struggles with high-dimensional datasets and intricate spatiotemporal relationships. With data-driven modules learning from historical datasets, ProNet outperforms human forecasts with 18.4% reduction in RMSE, achieving a 7-day RMSE lower than the 1-day RMSE by human forecasters. Accuracy rises nationally from 81.9% to 88.0% (Fig. 5c), with spatially heterogeneous gains (SI Fig. S7): 13.2% in BTHS, 5.4% in Yangtze River Delta Region (YRD), and 8.9% in Yangtze River Midstream City Cluster (YRM). The specific cities of the regions are listed in SI Table S2. More details about forecast accuracies in more regions are provided in SI Table S3.

For pollution regulation, human forecasters strategically prioritize high-concentration events. As shown in Figure 6b, the human predictions systematically produce higher concentrations, yielding a national average 16.1% higher than observations, which helps them achieve a higher hit rate for severe

interpretability of process-driven models while compensating for their deficiencies through deep neural networks. Compared with purely AI-based models, ProNet more effectively assimilates observations, predicts pollution variations, and adapts to extreme events beyond training data. Compared with human forecasting that similarly utilizes CTM outputs, ProNet demonstrates enhanced capacity to learn from large datasets and capture complex spatiotemporal and interspecies patterns. These advantages enable ProNet to address two critical challenges. First, limited training data availability: this study utilizes 1 year of training data because CTM undergo continuous mechanistic refinement and observational networks expand rapidly. This dynamic evolution renders long-term training datasets potentially suboptimal for identifying CTM systematic uncertainties. Second, substantial discrepancies between training and testing datasets: The training period (2022) coincided with China's COVID-19 lockdowns, whereas the testing period (post-January 8, 2023) followed lockdown lifting and experienced most sand storms in recent years. Notwithstanding these challenges, ProNet outperformed human operational forecasts, achieving RMSE reductions of 18.4% and accuracy increase of 6.1%. The framework also provided interpretable insights revealing uncertainties in current $\text{PM}_{2.5}$ formation mechanisms and identifying the spatial and temporal sources of CTM forecasting errors.

This study acknowledges three principal limitations that warrant further investigation. First, similar to deep-learning-based forecasting systems, ProNet's spatial resolution remains constrained by GPU memory limitations. Future research efforts could focus on developing a dynamic downscaling module to enhance urban-scale pollution simulations. Second, while the ensemble optimal interpolation method was selected for its computational efficiency in data assimilation, more sophisticated approaches such as the ensemble Kalman filter (EnKF) or four-dimensional variational (4D-Var)

assimilation could be implemented to optimize error covariance estimation and improve the physical consistency of assimilated datasets. Third, although our hybrid framework demonstrates superior relative mass-variation consistency compared to purely AI-driven models, residual imbalances persist. Subsequent developments could integrate rigorous mass-balance constraints through physics-informed neural network architectures to ensure strict mass balance for seasonal forecasting.

This study establishes a novel coupling framework between process-driven models and deep learning networks, specifically validated through its implementation in the CTM (NAQPMS) for air quality forecasting. Importantly, this framework is fundamentally model-agnostic and generalizable beyond NAQPMS; since it learns the error residuals of physical operators rather than relying on model-specific architectures, it can be seamlessly migrated to other Eulerian grid-based CTMs (such as CMAQ or CAMx) simply by training the correctors on their respective baseline forecast and assimilation fields. While successful implementation of this framework demands both process-driven modeling expertise and rigorous analysis, key methodological innovations exhibit broader applicability across Earth system components. First, integrating data assimilation with process-driven modules generates enriched training datasets for deep learning components, facilitating submodule pattern recognition even under sparse observational constraints. Second, our ensemble forecasting architecture incorporates dynamic weight adjustment mechanisms that effectively mitigate bias accumulation during iterative cycles. Third, coupling architectures that learn process-driven model uncertainties through data-driven networks, rather than attempting full substitution, create interpretable frameworks for the diagnosis of mechanistic deficiencies. Collectively, these methodological advances offer transformative potential for next-generation Earth system modeling, significantly enhancing both predictive accuracy and diagnostic capabilities of critical processes. Our

current understanding of the Earth system remains very limited, primarily due to the accuracy of existing models. By applying our framework, it is possible to both ensure model accuracy and deepen our understanding of the Earth system and its internal feedback processes. This approach can also improve our knowledge of future disaster evolution, thereby greatly enhancing humanity's capacity to respond to and adapt to the challenges posed by global change.

MATERIALS AND METHODS

Datasets

The training dataset contains 371 daily CTM forecasts from 26 December 2021 to 31 December 2022, each with dimensions of 464 (longitude) $\times$ 300 (latitude) $\times$ 169 (hours) and 30 variables (SI Table S1). The testing dataset similarly contains 371 forecasts from 26 December 2022 to 31 December 2023. Data before 1 January 2023 from overlapping forecasts were used for training, while 2023 data were held for evaluation.

Surface observations were obtained from the China National Environmental Monitoring Center. After quality control, data from 7,036 sites (SI Fig. S2) were used for model training, with a subset of 1,577 national-control sites utilized for evaluation.

Governing equations of the ProNet

ProNet builds upon CTM [39], whose governing equation is given by:

$$c_{t+1}^b = c_t^b + \left(\nabla \cdot (K^b \nabla c_t^b) - \nabla \cdot (\mathbf{v}_t^b c_t^b) + E_t^b + R^b(c_t^b, \mathbb{C}_t^b) + D^b(c_t^b) \right) \Delta t \quad (1)$$

where c^b and $\mathbb{C}^b$ denote the concentrations of the target and other relevant chemical species, respectively. K^b represents the eddy diffusivity, $\mathbf{v}^b$ is the wind vector, and E^b , R^b , and D^b refer to the rates of emission, reaction, and deposition processes, respectively. The superscript “b” indicates baseline forecasts generated by the CTMs.

parameterizations. Input variables for Emission-Deposition Corrector include temperature, relative humidity, wind speed and divergence, u- and v-wind components, terrain elevation, land type, planetary boundary layer height, target species concentrations from CTM, and changes of concentration beside transportation. Additionally, the historical space encoder incorporates CTM forecast errors and target uncertainty errors from the preceding seven days. Variables associated with chemical reactions, described in Table S1, are selected as predictors. These include temperature, relative humidity, and changes of concentrations beside transportation for 18 chemical species.

These correctors are trained simultaneously using the target uncertainty errors isolated by the physical transport module, with an additional L1 penalty of 0.05 applied to the chemical module to distinguish reaction-specific errors. Detailed specifications of network architectures, layers, parameter counts, and training settings are provided in Supplementary Section 2.

Training of AI Correctors in ProNet

During data preprocessing, baseline forecasts were generated using conventional CTM simulations without data assimilation; observed data were then assimilated to produce analyzed concentrations. The correctors took as input the hourly concentration changes (excluding transport) for each species, with the baseline-analysis differences defining the target Uncertainty errors (Fig. 1e).

We jointly trained the Emission-Deposition and Chemical-Reaction Correctors inside the ETM using CTM outputs, transport-free concentration changes, and corresponding errors. The correctors are optimized using a spatially weighted loss function that evaluates errors exclusively in grid cells with valid station observations. To suppress error accumulation during continuous forecasting, we construct a 10-member ensemble using corrector parameters saved at training epochs 291 to 300. The minor parameter variations across these epochs act as perturbations that counterbalance individual

biases. The final Ensemble Corrector, containing about 18 million parameters across convolutional space and temporal modules, is trained to merge these forecasts using a standard MSE loss for the base model, and a pollution-weighted loss for the high-pollution variant Couple^H. In the operational phase, the full framework utilizes an ensemble approach totaling approximately 240 million parameters. All models ran on an A800 GPU and were optimized with Adam, using learning rates of 2×10^{-3} , 5×10^{-5} , and 2×10^{-4} for the Emission-Deposition, Chemical-Reaction, and Ensemble Correctors, respectively. Each 7-day forecast is treated as a batch for the Chemical-Reaction Corrector and a sample for other Correctors. While conventional chemical transport models and data assimilation steps consume vast computational resources, our AI correctors are highly efficient. The entire multi-epoch training requires 92 hours utilizing one A800 GPU and ten 56-core CPU nodes, whereas executing a single multi-day operational forecast takes merely 92 seconds beside running CTM and data assimilation. More details are provided in Supplementary Sections 2 and 7.

DATA AVAILABILITY

The complete dataset generated and analyzed in this study totals approximately 13.9 TB (1.2 TB training data, 1.2 TB testing data, and 11.5 TB model outputs). Due to storage constraints, the full datasets are available from the corresponding author upon reasonable request. However, to facilitate reproducibility, a 34 GB demonstration subset is publicly available (<https://ug.link/couplegroup/filemgr/share-download/?id=f4feb4f9d68745dea50cf4898a4156b0>). This subset includes all trained model parameters (~2.6 GB) and complete input data for a single day, which is sufficient to independently run the model and reproduce the forecast case study in Figure 3.

CODE AVAILABILITY

All source codes associated with this study were written in the Julia programming language and publicly accessible through the GitHub repository at <https://github.com/wuhuangjianCN/ProNet>.

ACKNOWLEDGMENTS

We thank S.X. Chen for directional advice, H.R. Sun and H.Y Wang for useful discussion. We also appreciate the constructive feedback provided by B. Liu at the China National Environmental Monitoring Center for helpful feedback. Additionally, we acknowledge the technical support of the National Large Scientific and Technological Infrastructure “Earth System Numerical Simulation Facility” (<https://cstr.cn/31134.02.EL>).

FUNDING

This work was supported by the Jing-Jin-Ji Regional Integrated Environmental Improvement-National Science and Technology Major Project (Grant No. 2025ZD1201702), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB0760100) and the National Natural Science Foundation of China (Grant No. 42407148).

AUTHOR CONTRIBUTIONS

Z.W. and Y.Z. conceptualized the work and advised the research direction. Z.W managed the research project. H.W. explored the methodology and wrote the code. X.T studied the difference between ProNet and CTMs. D.Z. and W.W. collected and processed the observation and operational forecast data. J.Z., C.M., and H.S. contributed to the model architecture, while P.F., H.S., and Y.S. contributed to the interpretable chemical analysis. H.W. and W.L. wrote the original article. Y.C., X.L., G.C. and M.G. revised the final article.

505

506 ***Conflict of interest statement.*** None declared.

507

ORIGINAL UNEDITED MANUSCRIPT

508 **REFERENCES**

- 509 1. Feng Y, Castro E, Wei Y *et al.* Long-term exposure to ambient PM_{2.5}, particulate constituents and hospital
510 admissions from non-respiratory infection. *Nat Commun.* 2024; **15**(1): 1518.
- 511 2. Landrigan PJ. Air pollution and health. *The Lancet Public Health.* 2017; **2**(1): e4–e5.
- 512 3. Jin Y, Andersson H, Zhang S. Air pollution control policies in China: a retrospective and prospects. *International*
513 *journal of environmental research and public health.* 2016; **13**(12): 1219.
- 514 4. Lawrence MG, Rasch PJ, Von Kuhlmann R *et al.* Global chemical weather forecasts for field campaign planning:
515 predictions and observations of large-scale features during MINOS, CONTRACE, and INDOEX. *Atmos Chem*
516 *Phys.* 2003; **3**(1): 267–289.
- 517 5. Zhang Y, Baklanov A, Appel KW *et al.* Training Materials and Best Practices for Chemical Weather/Air Quality
518 Forecasting. 2020.
- 519 6. Zhu YH, Huang L, Li JY *et al.* Sources of particulate matter in China: Insights from source apportionment studies
520 published in 1987–2017. *Environ Int.* 2018; **115**: 343–357. doi: 10.1016/j.envint.2018.03.037
- 521 7. Cheng J, Su J, Cui T *et al.* Dominant role of emission reduction in PM_{2.5} air quality improvement in Beijing
522 during 2013–2017: a model-based decomposition analysis. *Atmos Chem Phys.* 2019; **19**(9): 6125–6146.
- 523 8. Aleksankina K, Reis S, Vieno M *et al.* Advanced methods for uncertainty assessment and global sensitivity
524 analysis of an Eulerian atmospheric chemistry transport model. *Atmos Chem Phys.* 2019; **19**(5): 2881–2898.
525 doi: 10.5194/acp-19-2881-2019
- 526 9. Mallet V, Sportisse B. Uncertainty in a chemistry-transport model due to physical parameterizations and
527 numerical approximations: An ensemble approach applied to ozone modeling. *J Geophys Res: Atmos.* 2006;
528 **111**(D1). doi: 10.1029/2005jd006149
- 529 10. Kumar R, Peuch V-H, Crawford JH *et al.* Five steps to improve air-quality forecasts. *Nature.* 2018; **561**(7721):
530 27–29. doi: 10.1038/d41586-018-06150-5
- 531 11. Zhang Y, Bocquet M, Mallet V *et al.* Real-time air quality forecasting, part II: State of the science, current
532 research needs, and future prospects. *Atmos Environ.* 2012; **60**: 656–676. doi: 10.1016/j.atmosenv.2012.02.041
- 533 12. Wu H, Tang X, Wang Z *et al.* High-spatiotemporal-resolution inverse estimation of CO and NO_x emission
534 reductions during emission control periods with a modified ensemble Kalman filter. *Atmos Environ.* 2020; **236**.
535 doi: 10.1016/j.atmosenv.2020.117631
- 536 13. Chen L, Gao Y, Zhang MG *et al.* MICS-Asia III: multi-model comparison and evaluation of aerosol over East
537 Asia. *Atmos Chem Phys.* 2019; **19**(18): 11911–11937. doi: 10.5194/acp-19-11911-2019
- 538 14. Bi K, Xie L, Zhang H *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature.*
539 2023. doi: 10.1038/s41586-023-06185-3
- 540 15. Sun X, Zhong X, Xu X *et al.* A data-to-forecast machine learning system for global weather. *Nat Commun.*
541 2025; **16**(1): 6658.
- 542 16. Liao Q, Zhu MM, Wu L *et al.* Deep Learning for Air Quality Forecasts: a Review. *Current Pollution Reports.* 2020;
543 **6**(4): 399–409. doi: 10.1007/s40726-020-00159-z
- 544 17. Zhu JQ, Deng F, Zhao JC *et al.* Attention-based parallel networks (APNet) for PM_{2.5} spatiotemporal prediction.
545 *Sci Total Environ.* 2021; **769**. doi: 10.1016/j.scitotenv.2021.145082
- 546 18. Ma W, Yuan Z, Lau AKH *et al.* Optimized neural network for daily-scale ozone prediction based on transfer
547 learning. *Sci Total Environ.* 2022; **827**: 154279. doi: 10.1016/j.scitotenv.2022.154279
- 548 19. Li M, Zhang Q, Kurokawa J-i *et al.* MIX: a mosaic Asian anthropogenic emission inventory under the
549 international collaboration framework of the MICS-Asia and HTAP. *Atmos Chem Phys.* 2017; **17**(2). doi:
550 10.5194/acp-17-935-2017

- 551 20. Wu H, Kong L, Tang X *et al.* Air Quality Forecasting with Inversely Updated Emissions for China. *Environ Sci*
552 *Technol Lett.* 2023; **10**(8): 655–661. doi: 10.1021/acs.estlett.3c00266
- 553 21. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable
554 models instead. *Nat Mach Intell.* 2019; **1**(5): 206–215. doi: 10.1038/s42256-019-0048-x
- 555 22. Bodnar C, Bruinsma WP, Lucic A *et al.* A foundation model for the Earth system. *Nature.* 2025. doi:
556 10.1038/s41586-025-09005-y
- 557 23. Zhu Y, Zhang R-H, Moum JN *et al.* Physics-informed deep-learning parameterization of ocean vertical mixing
558 improves climate simulations. *Nat Sci Rev.* 2022; **9**(8). doi: 10.1093/nsr/nwac044
- 559 24. Irrgang C, Boers N, Sonnewald M *et al.* Towards neural Earth system modelling by integrating artificial
560 intelligence in Earth system science. *Nat Mach Intell.* 2021; **3**(8): 667–674. doi: 10.1038/s42256-021-00374-3
- 561 25. Reichstein M, Camps-Valls G, Stevens B *et al.* Deep learning and process understanding for data-driven Earth
562 system science. *Nature.* 2019; **566**(7743): 195.
- 563 26. Zhang Y, Long M, Chen K *et al.* Skilful nowcasting of extreme precipitation with NowcastNet. *Nature.* 2023.
564 doi: 10.1038/s41586-023-06184-4
- 565 27. Kochkov D, Yuval J, Langmore I *et al.* Neural general circulation models for weather and climate. *Nature.* 2024;
566 **632**(8027): 1060–1066. doi: 10.1038/s41586-024-07744-y
- 567 28. Kashinath K, Mustafa M, Albert A *et al.* Physics-informed machine learning: case studies for weather and
568 climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*
569 *Sciences.* 2021; **379**(2194). doi: 10.1098/rsta.2020.0093
- 570 29. Shen C, Appling AP, Gentine P *et al.* Differentiable modelling to unify machine learning and physical models
571 for geosciences. *Nature Reviews Earth & Environment.* 2023; **4**(8): 552–567. doi: 10.1038/s43017-023-00450-
572 9
- 573 30. Sayeed A, Lops Y, Choi Y *et al.* Bias correcting and extending the PM forecast by CMAQ up to 7 days using
574 deep convolutional neural networks. *Atmos Environ.* 2021; **253**. doi: 10.1016/j.atmosenv.2021.118376
- 575 31. Keller CA, Evans MJ. Application of random forest regression to the calculation of gas-phase chemistry within
576 the GEOS-Chem chemistry model v10. *Geosci Model Dev.* 2019; **12**(3): 1209–1225. doi: 10.5194/gmd-12-
577 1209-2019
- 578 32. Wang Z, Maeda T, Hayashi M *et al.* A nested air quality prediction modeling system for urban and regional
579 scales: Application for high-ozone episode in Taiwan. *Water, Air, Soil Pollut.* 2001; **130**(1/4): 391–396. doi:
580 10.1023/a:1013833217916
- 581 33. Zhang RY, Wang GH, Guo S *et al.* Formation of Urban Fine Particulate Matter. *Chem Rev.* 2015; **115**(10): 3803–
582 3855. doi: 10.1021/acs.chemrev.5b00067
- 583 34. Bove MC, Brotto P, Cassola F *et al.* An integrated PM_{2.5} source apportionment study: Positive Matrix
584 Factorisation vs. the chemical transport model CAMx. *Atmos Environ.* 2014; **94**: 274–286. doi:
585 10.1016/j.atmosenv.2014.05.039
- 586 35. Bertram TH, Thornton JA. Toward a general parameterization of N₂O₅ reactivity on aqueous particles: the
587 competing effects of particle liquid water, nitrate and chloride. *Atmos Chem Phys.* 2009; **9**(21): 8351–8363. doi:
588 10.5194/acp-9-8351-2009
- 589 36. Brown SS, Stutz J. Nighttime radical observations and chemistry. *Chemical Society Reviews.* 2012; **41**(19): 6405–
590 6447. doi: 10.1039/c2cs35181a
- 591 37. McDuffie EE, Fibiger DL, Dube WP *et al.* Heterogeneous N₂O₅ Uptake During Winter: Aircraft Measurements
592 During the 2015 WINTER Campaign and Critical Evaluation of Current Parameterizations. *J Geophys Res:*
593 *Atmos.* 2018; **123**(8): 4345–4372. doi: 10.1002/2018jd028336
- 594 38. Miao RQ, Chen Q, Zheng Y *et al.* Model bias in simulating major chemical components of PM_{2.5} in China. *Atmos*

- 595 *Chem Phys.* 2020; **20**(20): 12265–12284. doi: 10.5194/acp-20-12265-2020
- 596 39. Chen L, Gao Y, Zhang M *et al.* MICS-Asia III: multi-model comparison and evaluation of aerosol over East Asia.
- 597 *Atmos Chem Phys.* 2019; **19**(18): 11911–11937. doi: 10.5194/acp-19-11911-2019
- 598 40. Vasilakos P, Russell A, Weber R *et al.* Understanding nitrate formation in a world with less sulfate. *Atmos Chem*
- 599 *Phys.* 2018; **18**(17): 12765–12775. doi: 10.5194/acp-18-12765-2018
- 600 41. Sarwar G, Simon H, Xing J *et al.* Importance of tropospheric ClNO₂ chemistry across the Northern Hemisphere.
- 601 *Geophys Res Lett.* 2014; **41**(11): 4050–4058. doi: 10.1002/2014gl059962
- 602 42. Lorenz EN. DETERMINISTIC NONPERIODIC FLOW. *J Atmos Sci.* 1963; **20**(2): 130–141. doi: 10.1175/1520-
- 603 0469(1963)020<0130:Dnf>2.0.Co;2
- 604 43. Kohavi R, Wolpert DH. Bias plus variance decomposition for zero-one loss functions. In: *International*
- 605 *Conference on Machine Learning, 1996*, p. 275–283. ICML.
- 606 44. Bi KF, Xie LX, Zhang HH *et al.* Accurate medium-range global weather forecasting with 3D neural networks.
- 607 *Nature.* 2023; **619**(7970): 533–+. doi: 10.1038/s41586-023-06185-3
- 608 45. Feng Y, Ning M, Lei Y *et al.* Defending blue sky in China: Effectiveness of the “Air Pollution Prevention and
- 609 Control Action Plan” on air quality improvements from 2013 to 2017. *Journal of environmental management.*
- 610 2019; **252**: 109603.
- 611 46. Hong C, Zhang Q, Zhang Y *et al.* Impacts of climate change on future air quality and human health in China.
- 612 *Proceedings of the national academy of sciences.* 2019; **116**(35): 17193–17200.
- 613 47. Yin Z, Huo Q, Ma X *et al.* Mechanisms of dust source accumulation and synoptic disturbance triggering the
- 614 2023 spring sandstorm in northern China. *Transactions of Atmospheric Sciences.* 2023; **46**(3): 321–331.
- 615 48. Chen Y, Chen SY, Bi HR *et al.* Where is the Dust Source of 2023 Several Severe Dust Events in China? *Bull Am*
- 616 *Meteorol Soc.* 2024; **105**(11): E2085–E2096. doi: 10.1175/bams-d-23-01211
- 617

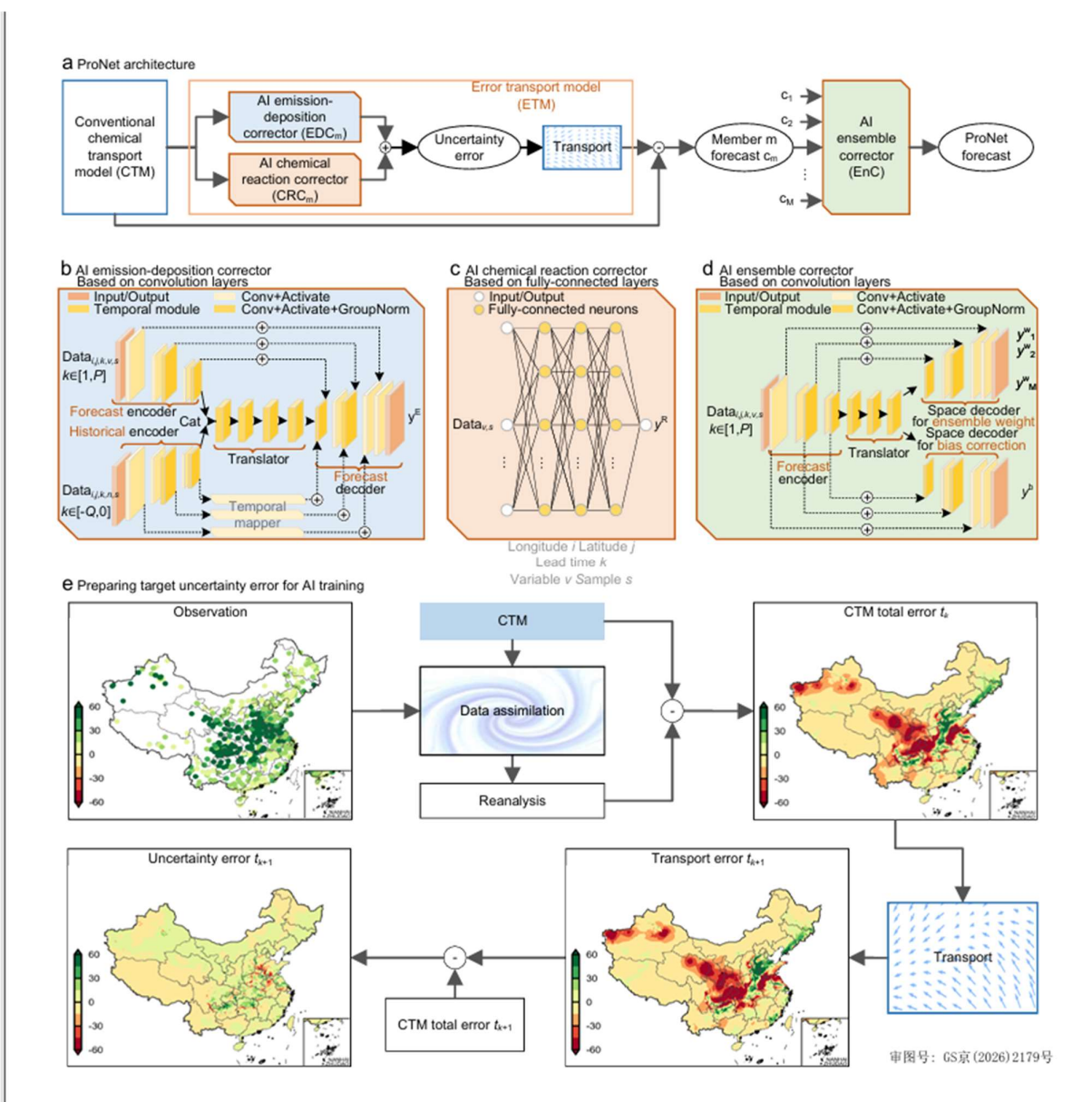


Figure. 1 | Architecture and training framework of ProNet. a, Schematic of the ProNet forecasting framework. ProNet integrates a CTM for baseline predictions with an ETM to refine forecasts. The ETM combines a physical process module (transportation) with two neural networks: (i) the Emission-Deposition Corrector, which learns spatiotemporal patterns to compensate for errors arising from emission and deposition processes in the CTM; and (ii) the Chemical Reaction Corrector, which captures interactions among chemical species to correct inaccuracies in the chemical reaction processes. The outputs from the CTM and multiple ETMs are merged to generate member forecasts,

which are subsequently processed by the Ensemble Corrector to mitigate the accumulation of biases stemming from AI Correctors in ETM. **b-d**, Architectural schematics of the Emission-Deposition Corrector (**b**), Chemical-Reaction Corrector (**c**), and Ensemble Corrector (**d**). **e**, Techniques to generate target Uncertainty Error for the training of AI Correctors in ETM.

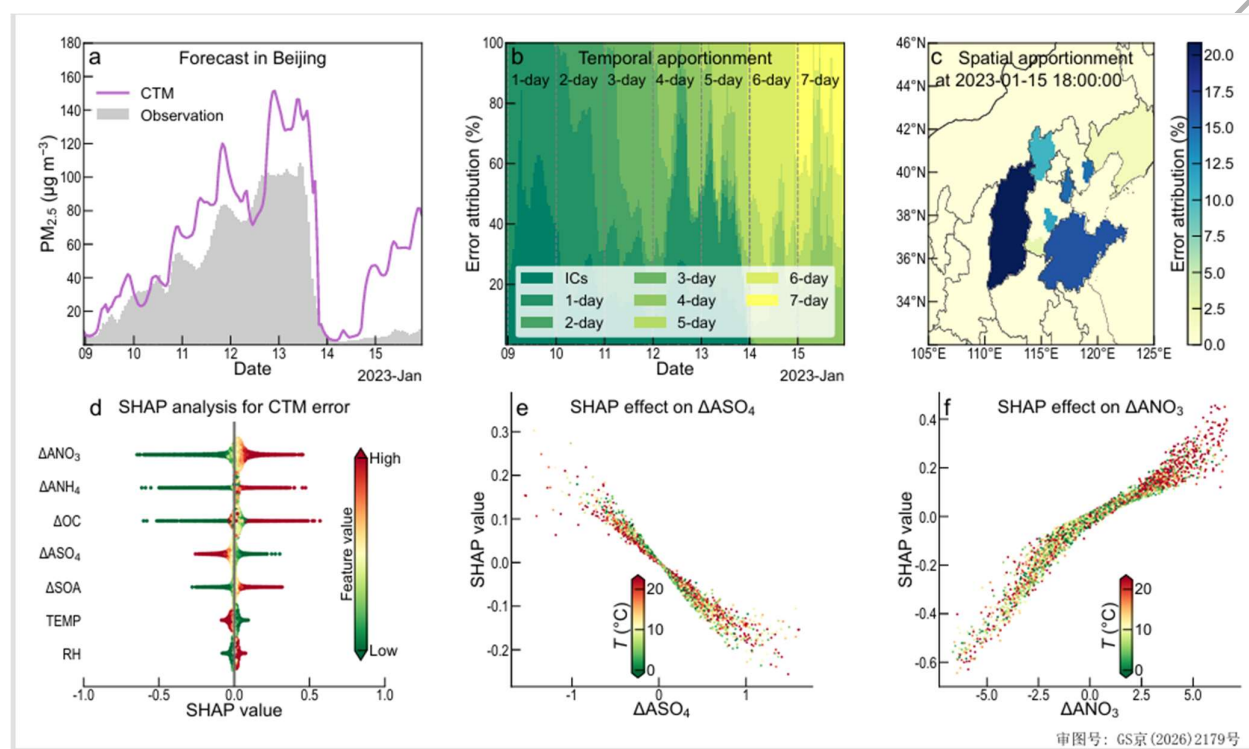


Figure. 2 | Interpretable analysis of PM_{2.5} forecasting. **a-c**, Error apportionment for a forecast in Beijing. An animated version of the error apportionment is provided in Supplementary Video 1. **a** is the forecast to analysis, **b** is the temporal error apportionment identifying when the total error is stemming from, including the ICs (before the forecast), and uncertainties of CTM in 1 to 7 day forecasts. **c** is the spatial error apportionment identifying where the total forecast error in Beijing at 2023-01-15 18:00:00 is stemming from. **d-f**, uncertainty analysis of PM_{2.5} formation mechanisms. **d**, Beeswarm plots of key drivers influencing PM_{2.5} concentration errors. Each point represents an

individual sample at a specific grid and time step, where its x-axis, i.e. the SHAP values represent the impact of predictors on CTM forecast error for $\text{PM}_{2.5}$. Point density is visualized through vertical stacking. An increase in SHAP value alongside the concentration of a $\text{PM}_{2.5}$ component indicates an overestimation in the formation of that component, while a decrease suggests an underestimation. **e-f**, SHAP dependence plots reveal parameter-specific patterns for sulfate (**e**) and nitrate (**f**). The SHAP value increases with the rise in ΔANO_3 and decreases with ΔASO_4 , indicating an overestimation of nitrate formation and an underestimation of sulfate formation. A pronounced temperature effect is evident in **e**, revealing a greater underestimation of sulfate formation during winter.

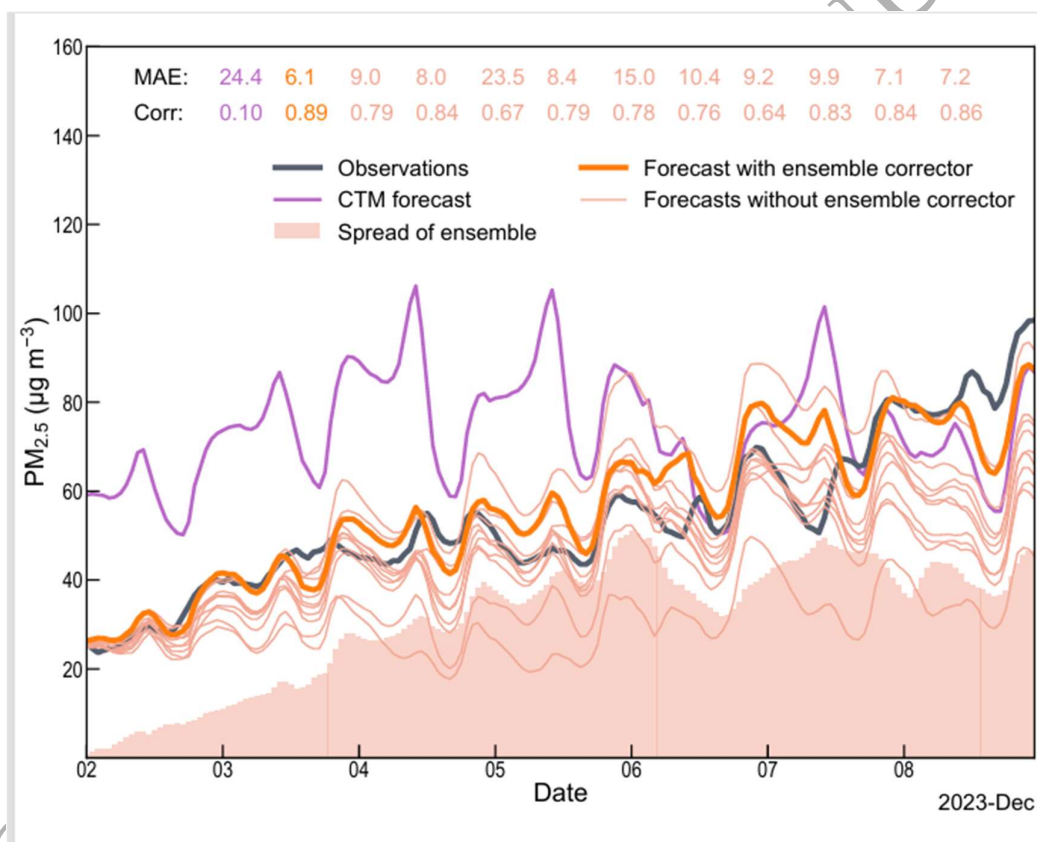


Figure. 3 | Suppressing bias accumulation via the Ensemble Corrector. The ChengYu Region in the Sichuan Basin, being relatively isolated from external pollution sources, functions as a semi-closed

651 system ideal for assessing bias accumulation. Without ensemble correction, systematic biases from
652 data-driven modules accumulate progressively, leading to an increase in ensemble spread by orders of
653 magnitude. The proposed approach constructs a 10-member ensemble using neural network
654 parameters sampled from different training epochs; these are subsequently processed through the
655 Ensemble Corrector to detect and mitigate bias accumulation, thereby improving long-term forecast
656 accuracy.

ORIGINAL UNEDITED MANUSCRIPT

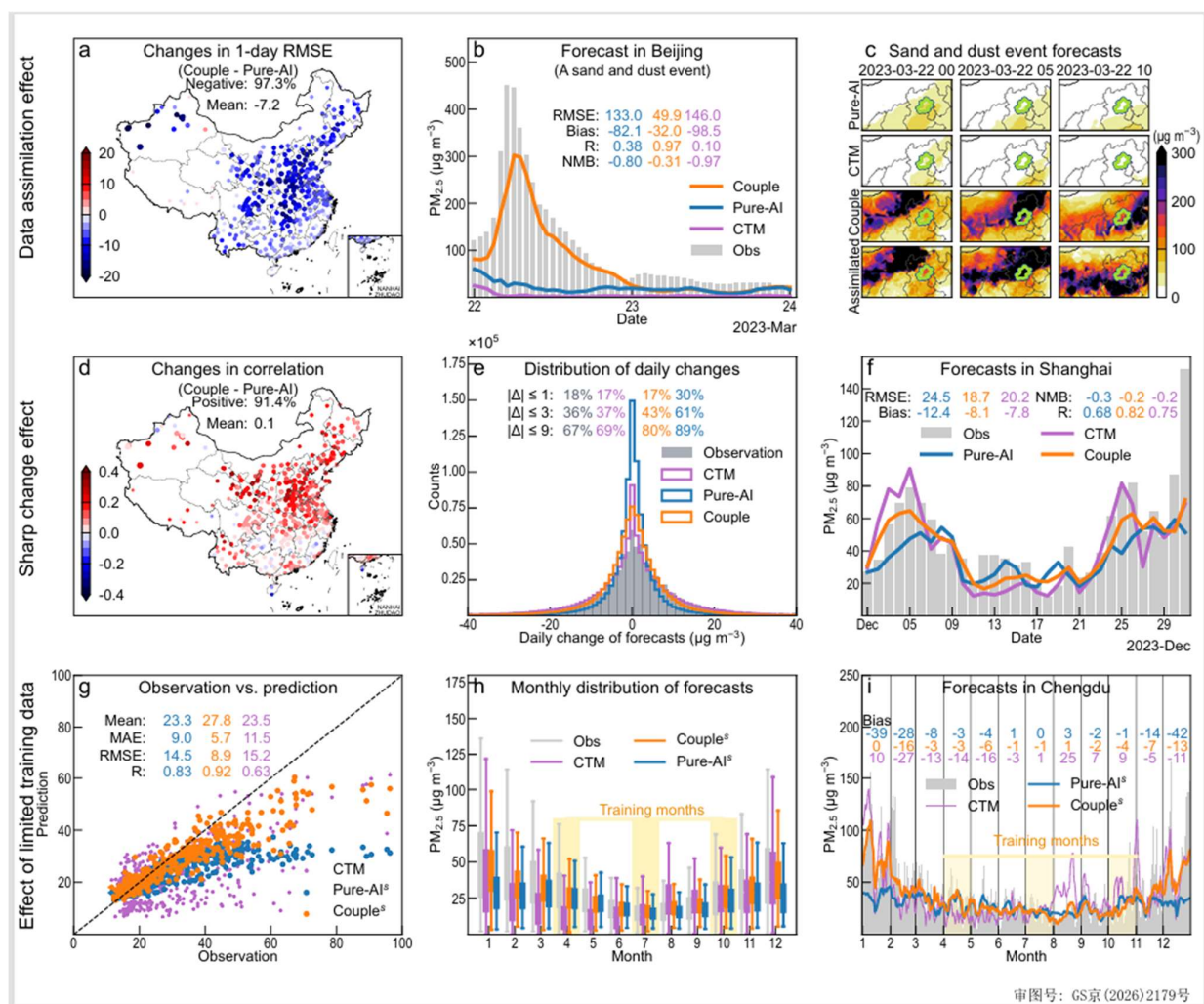


Figure 4 | Performance advantages of the coupled model over pure AI-based approaches. a-c,

The coupled model improves short-term forecasts by effectively assimilates observations into ICs. **a,**

Distribution of 1-day RMSE changes by coupling. **b,** A Forecast case in Beijing. **c,** PM_{2.5} distributions

of forecasts during a sand and dust event. **d-f,** The coupled model outperforms pure AI approaches in

sharp pollution changes. **d,** Distribution of correlation coefficient changes induced by coupling. **e,**

Distribution of daily PM_{2.5} concentration changes in the forecasts. **f,** Forecasts for Guangzhou in

January. **g-i,** The coupled model exhibits better generalization with limited training data (April, July,

and October 2022). **g**, Scatter plot of national daily mean $PM_{2.5}$ concentrations in 2023. **h**, Monthly distribution of forecasts. **i**, Forecasts in Chengdu.

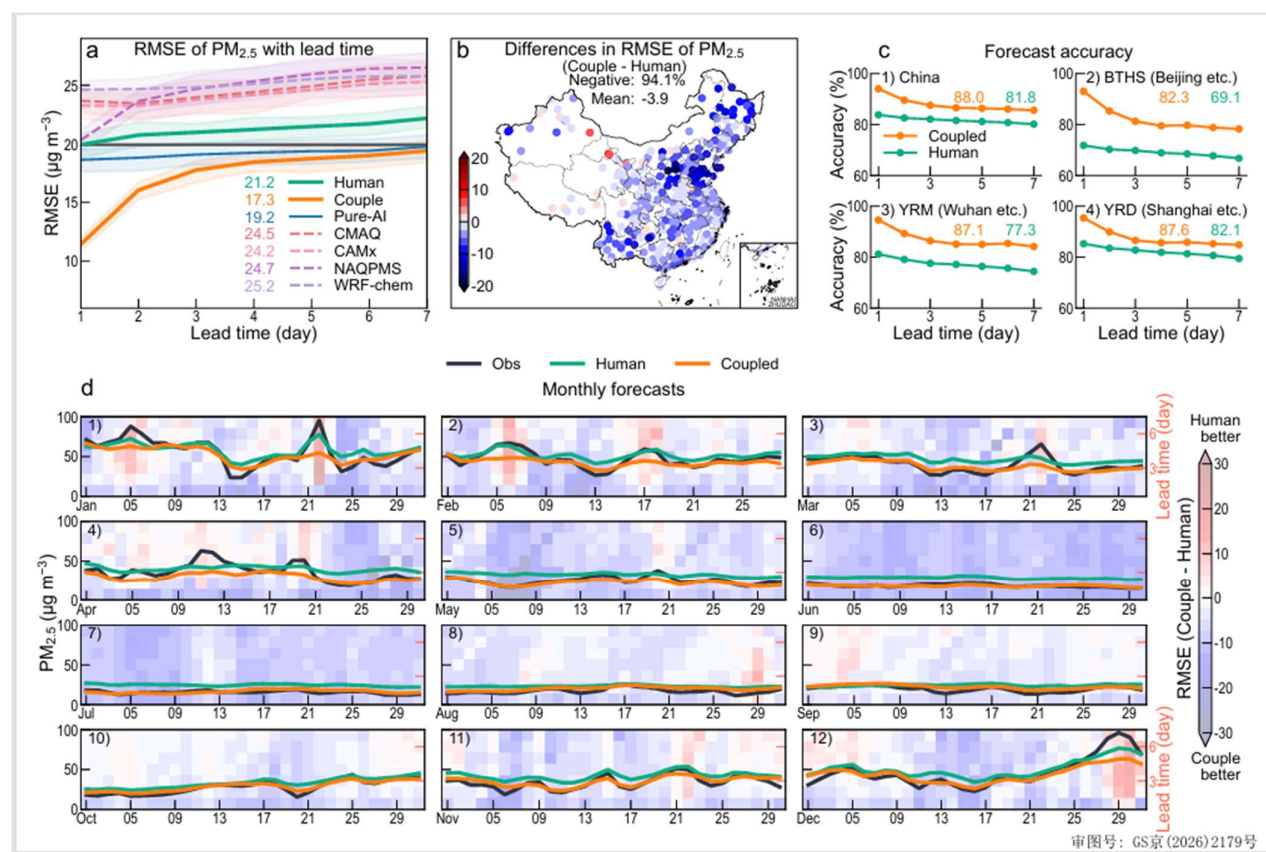


Figure. 5 | Performance comparison of the coupled model against operational forecasts over China. **a**, RMSE curves as functions of lead time for city-level daily-mean $PM_{2.5}$ across all monitored cities, with shaded regions denoting 95% confidence intervals. Human forecasts and model outputs are evaluated using the same daily-mean metric and lead-time definition. Numbers beside the legend indicate mean RMSE averaged across D+1 to D+7. **b**, Spatial distributions of RMSE reduction achieved by the coupled model for $PM_{2.5}$. **c**, Forecasting accuracy of $PM_{2.5}$ for China as a whole (1) and three regions (2-4). The specific cities of the regions are listed in SI Table S2, and the procedures for accuracy calculation is detailed in SI Section 5. **d**, Monthly evaluation of $PM_{2.5}$ forecasts during

678 2023: solid lines show observed and forecasted concentrations (left y-axis), while background colors
679 quantify RMSE differences between human and couple forecasts (right y-axis). Collectively, these
680 results demonstrate ProNet's superior performance over both human forecasts and conventional CTMs
681 in spatiotemporal coverage and prediction fidelity.

682

ORIGINAL UNEDITED MANUSCRIPT

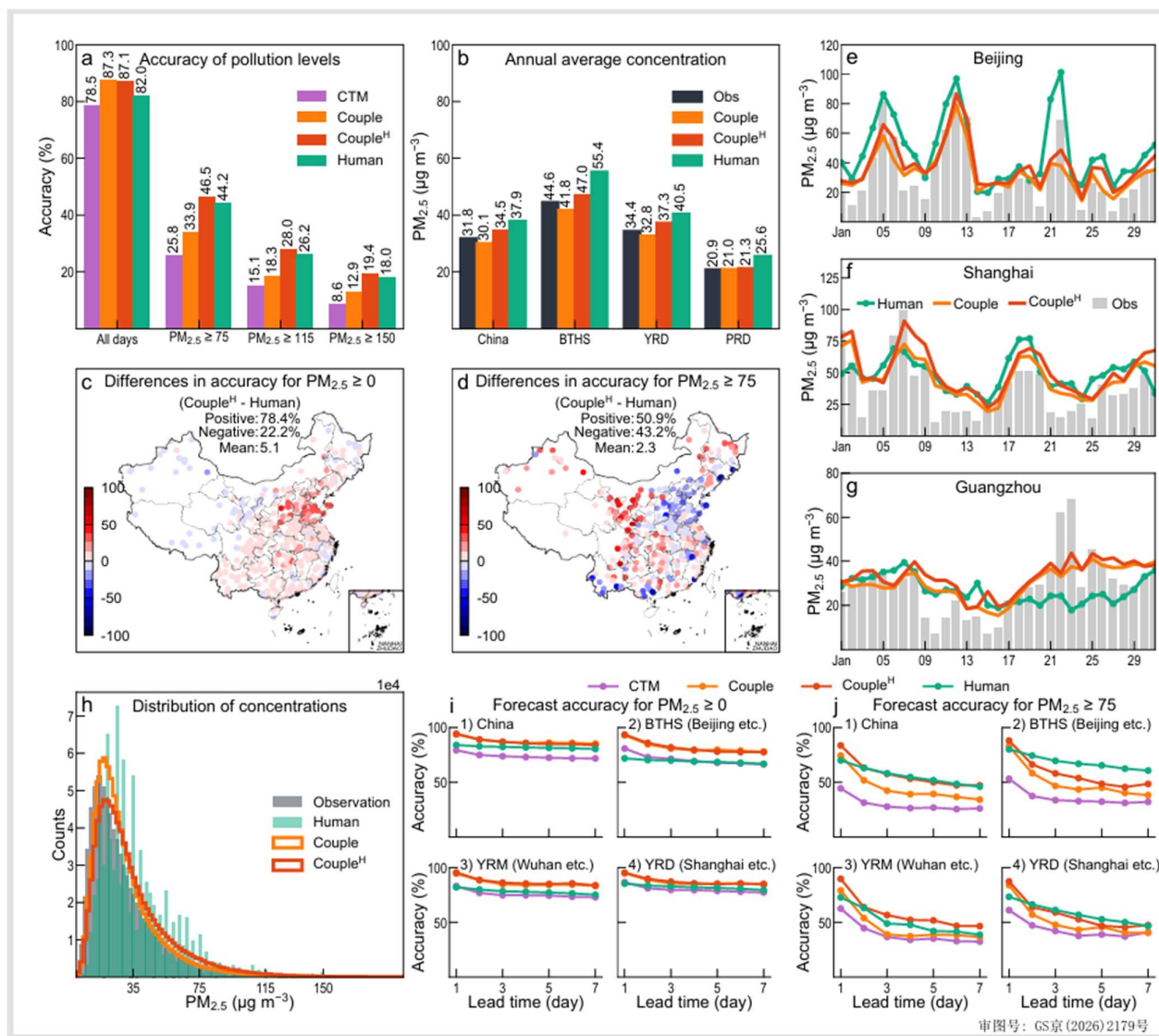


Figure 6 | Model fine-tuning to improve PM_{2.5} forecast accuracy for polluted days. **a**, Forecast accuracy stratified by PM_{2.5} concentration thresholds. “Couple” denotes baseline ProNet model, while “Couple^H” represents the fine-tuned variant, which slightly sacrifices overall accuracy to improve performance for high-pollution events. Couple^H denotes the coupled model trained with a pollution-aware loss that upweights polluted days, whereas ProNet uses the standard loss. Severe pollution episodes are defined as days with daily-mean PM_{2.5} > 75 $\mu\text{g m}^{-3}$. This ablation quantifies the polluted-

691 day trade-off and shows how the pollution-aware loss improves forecast accuracy under polluted
692 conditions. **b**, Annual average concentrations of observed and forecasted $\text{PM}_{2.5}$, showing that human
693 forecasts are significantly overestimated, particularly in the BTHS region. **c-d**, Accuracy comparisons
694 for all days (**c**, $\text{PM}_{2.5} > 0 \mu\text{g m}^{-3}$) and polluted days (**d**, $\text{PM}_{2.5} > 75 \mu\text{g m}^{-3}$). **e-g**, Forecasts for Beijing
695 (**e**), Shanghai (**f**), and Guangzhou (**g**) in January 2023. **h**, Density distributions of forecasted versus
696 observed $\text{PM}_{2.5}$ concentrations, showing that both human and Couple^H forecasts are more frequently
697 distributed at higher concentrations. **i-j**, Forecast accuracy for all days (**i**, $\text{PM}_{2.5} > 0 \mu\text{g m}^{-3}$) and polluted
698 days (**j**, $\text{PM}_{2.5} > 75 \mu\text{g m}^{-3}$) by station.

699