Review for egusphere-2026-1109

This manuscript presents a relevant and technically solid framework for annual high-resolution air quality mapping over Europe based on the integration of deterministic model outputs, in-situ observations, satellite data, and geographic/meteorological predictors in an ensemble ML/DL framework. The paper is promising and publishable after major revision.

Also, it is a good read and a coherent addition to the existing AI/ML approaches to downscale regional-scale air quality fields in different regions or for different pollutants. The consistent framing and approach as an "annual downscaling tool for entire Europe for annual means with 500m resolution" and the strong use of mechanistic models for air pollution and meteorology as inputs, make it suitable for submission to GMD.

The main improvements needed are not about changing the core method, but about strengthening interpretation and validation framing with respect to:

- a better separation of "enhancement over CAMS" vs. "independent validation of CHROMAP system",
- a more explicit handling of spatial validation limitations,
- stronger regional diagnostics of errors and gains,
- precise wording in abstract and conclusions.

Strengths of the publication are:

- a clear objective and scope: 500 m annual maps for NO2, O3, PM2.5, PM10, and SOMO35 over 2013-2023.
- a straightforward, coherent and promising workflow and implementation transparency.
- the ensemble approach (Ridge, RF, XGBoost, MLP).
- the evaluation goes beyond one split (independent 2021 test + 5-fold CV over multiple years.
- SHAP values to support interpretability.
- (static) exposure sensitivity analysis is policy-relevant as a first step.
- a solid basement of mechanistic models for air pollutants and meteorology.

Major Comments

1) CAMS is both key input and benchmark: claims should reflect this more clearly

The manuscript clearly states that CHEM_CTM_* features come from CAMS reanalysis fields in section 2.2.1 and that CHROMAP maps are compared against CAMS_ens to assess enhancement in section 3.1. This is valid for a downscaling paper, but interpretive claims should consistently reflect that design. Part of the gain is expected to be "CAMS-informed refinement + bias correction", not fully independent information gain. And this is especially important because feature importance confirms the dominant role of CHEM_CTM_* for all pollutants (Manuscript and Supplement).

So, my suggestion for a revision here would be to explicitly separate two messages throughout Results/Discussion and to align Abstract and Conclusion wording with this

distinction: “relative enhancement over CAMS” vs. “absolute agreement with held-out observations”.

Also to take into account here are the expected urban performance gains by design (in comparison to CAMS): Please explicitly discuss that CAMS reanalysis assimilation is primarily based on background stations, while CHROMAP training includes background, industrial, and traffic stations. So, this setup likely explains part of the observed improvement in urban/traffic contexts and should be framed as expected added value rather than fully independent skill gain.

2) Independent validation is useful, but spatially strict validation is still missing

The 2021 independent test in section 3.3.1 and the supplement diagnostics are definitely valuable. However, the holdout appears stratified by station type (background/industrial/traffic) without explicit spatial de-correlation controls. But in spatially autocorrelated settings, random station holdout can overestimate transferability. I think, this is especially relevant when maps are presented for sparsely monitored regions. For a revision, it would be necessary to add one spatially stricter sensitivity test (e.g., leave-region-out or block CV) for at least 1-2 representative years. And report side-by-side performance change relative to the current random-stratified approach.

3) Regional diagnostics should be expanded (where and why errors occur)

The current evaluation is strong on aggregate metrics, but very limited in diagnosing regional behaviour. Europe-wide annual means and pooled scores do not fully reveal where errors are concentrated, particularly for short-lived/heterogeneous pollutants (NO2, regime-dependent O3. It leaves the reader a bit with the question where (in Europe) and why are some metrics still so poor (RRMSE, in table 2). And the claimed practical use of CHROMAP depends on regional reliability, not only continental averages. So, the high-resolution added value is easiest to support with local XOR regional evidence.

For the revision, region-stratified performance summaries (e.g., macro-regions or climate regions), residual/bias diagnostics by region and station type, and a brief interpretation of a few high-error regions need to be added.

4) SHAP is included but not fully leveraged

Interpretability is a clear strength of the manuscript, but SHAP analysis remains mostly global and descriptive in 3.2, although SHAP can directly support mechanistic interpretation of gains and failures. But, without stratification, it is hard to connect feature effects to regional or station-type behaviour.
Therefore, please provide compact stratified SHAP summaries (region and/or station type), potentially add a small set of dependence plots for key predictors per pollutant and link SHAP patterns to residual patterns in high-error situations.

5) Uncertainty and future-frequency claims should be narrowed

The paper states multiple times uncertainty estimation from ensemble variability. That is useful, but this should be framed as ensemble spread rather than a full uncertainty budget. During the first read, I was expecting a full uncertainty estimate (or at least a Gaussian Processes like possibility to identify different types of error and estimate uncertainty).
Similarly, the statement that the framework could be extended to higher temporal frequencies (in Abstract and Conclusion) is plausible, but currently not demonstrated.

Please clarify that uncertainty = inter-model spread within the chosen ensemble and somehow rephrase high-frequency extension as future work, ideally with a short feasibility note (data completeness, compute burden, expected uncertainty implications due to new models to be trained).

6) GMD requirements on data and code
Given the journal's emphasis on reproducibility and open science, I strongly encourage the authors to make CHROMAP input and output datasets openly and directly accessible (not only upon request), ideally through a persistent repository with versioning, metadata, and a clear data dictionary.

Minor comments (technical clarity)

- In methods/results text, explicitly reiterate that models are trained separately for each year and pollutant, to avoid (my) confusion with temporal forecasting.
- In section 3.3.1, explicitly state that the split unit is station observations for year 2021 (not a temporal forecast split).
- Consider adding a concise table in the main text that maps each validation setup to its intended claim:
    - independent holdout for station generalization within year,
    - 5-fold CV over years for robustness over annual model runs,
    - CAMS comparison for enhancement over coarse res. baseline.
- The manuscript should explicitly state that validation targets spatial generalization within each year, not temporal forecasting skill across years.
- Maye adding MQI/MQO could be interesting for policy purposes but then, if MQI is added, present it as a complementary policy-facing metric, not as the primary evidence of generalization performance (given RMSE-like training objectives of CHROMAP models; just thinking loud here).

Minor revisions: potential language/format issues to check

- check for consistent use of "Krigging" or "Kriging"
- line 105 "CHROMAP is a model...".
- ensure consistent naming/casing for feature identifiers (e.g., CHEM_Sat_no2 vs. CHEM_Sat_NO2; Geo_* vs. GEO_*) across main text, figures, and supplement.

- verify consistency of "at 500 m resolution" vs "at a resolution of 500m" formatting and spacing.
- Chen & Guestrin, 206 → likely 2016 (section 2.4, XGBoost citation).
- Geo_RoadDens appears in feature-importance text while GEO_RoadNet is used in feature table/methods (harmonize naming).
- reference entry Guion, A. CHROMAPv1.0 … 2016 → should be 2026?

One last comment on exposure analysis

The static exposure analysis is appropriate as a first Europe-wide step (section 3.4). Dynamic exposure would be methodologically stronger but is currently difficult to operationalize at continental scale. The manuscript should keep this section but should frame health-impact implications as “potentially significant and motivating follow-up HIA work” rather than as a direct health impact quantification.