Machine learning data fusion for high spatio-temporal resolution $PM_{2.5}$ - Reply to referees

Andrea Porcheddu, Ville Kolehmainen, Timo Lähivaara, Antti Lipponen

1 Foreword

We would like to thank the reviewers for reading carefully the manuscript and giving their comments. Below we reply to each of the comments.

2 Answers to reviewer #1

2.1 The study aims to estimate 24-hourly $PM_{2.5}$ maps at 100 m resolution in urban areas. However, as shown in Table A1, most of the input data have resolutions coarser than 100 m, except for OpenStreetMap roads and DEM data, which are not directly related to $PM_{2.5}$. How do the authors justify that the estimated $PM_{2.5}$ resolution truly reaches 100 m?

While the primary predictors of $PM_{2.5}$ in our study are MERRA-2 variables (which have coarse resolution), the high-resolution features provide important supplementary information about potential pollution sources, sinks and transport. Since our target variable—NOODLESALAD $PM_{2.5}$ maps—is on a grid with cell size 100 m x 100 m, we train the model on the same grid, considering that some input features are coarser and others finer. With this approach the model represents spatial patterns at the target data scale when fusing information from multiple resolutions. Additionally, we emphasize that this study is a proof of concept, and the same model framework can operate on different grids with different pixel size, such as 500 m. Even considering 500 m resolution grid, this approach would still offer a significant improvement over the resolution of MERRA-2 $PM_{2.5}$ estimates. For the sake of clarity, we updated the article referring to a "grid with cell size 100 m x 100 m" instead of "100 m resolution".

2.2 The paper presents a deep learning-based estimation approach, but the description of the methodology remains unclear. First, Lines 148–149 mention that "The output is a 3-dimensional array containing 24 hourly $PM_{2.5}$ maps," but Lines 159–160 state that "the output layer is a 3D 1x1x1 convolution," which appears contradictory and should be clarified. Second, the construction of the loss function is confusing—it should ideally be constrained by $PM_{2.5}$ measurements from ground stations and NOODLESALAD $PM_{2.5}$, but its current formulation appears overly complex and difficult to understand.

When we refer to 3×3×3 or 1×1×1 convolutions in the context of 3D convolution, we describe the size of the convolutional kernel along the depth, height, and width of the input volume. These kernels slide across 3D space, processing small local regions of the data at each step. To calculate the number of parameters involved, we need to consider also the number of input channels and output channels for that convolutional layer. Let's consider the last convolutional layer in the our model (Fig. 1). We have a 4D input tensor of shape 24 × 960 × 960 × 16, where: 24 is the number time steps, 960 × 960 is the spatial dimension (height × width), 16 is the number of input channels. A 1×1×1 convolution in this case corresponds to kernels of shape 1×1×1×16 (depth × height × width × input channels). If our goal is to reduce the number of channels from 16 to 1, then we need a 1×1×1×16 kernel for each output channel. Since we want 1 output channel, we use just one such kernel. This results in an output tensor of shape 24 × 960 × 960 × 1 — the same temporal and spatial dimensions, but with the number of channels reduced from 16 to 1. If instead we wanted, say, 24 output channels, we would use 24 separate 1×1×1×16 kernels, resulting in an output shape of 24 × 960 × 960 × 24. A clarification about the model architecture has been added to the manuscript.

The loss function is structured this way to address the inherent imbalance in the number and type of $PM_{2.5}$ measurements. At satellite overpass times, we typically have orders of magnitude more valid pixel-level estimates compared to the relatively sparse ground station measurements. If we were to aggregate all errors directly, the satellite data would dominate the loss, potentially causing the model to neglect the ground station data (which offer more accuracy, and the only temporal information available far from the satellite overpass time). To balance for the different number of ground and satellite data, separate fidelity terms for ground and satellite data are utilized in the loss, and their contributions to the training are balanced by normalizing the fidelity terms by the number of measurements available from each source. Further, since temporal imbalance could happen also when all the ground stations data is available at certain hours, ground stations data are weighted (by the number of ground measurements available at the specific hour) before the ground data loss value is calculated. This accounts for variations in data availability throughout the day and ensures that all measurements are appropriately represented in the final loss.

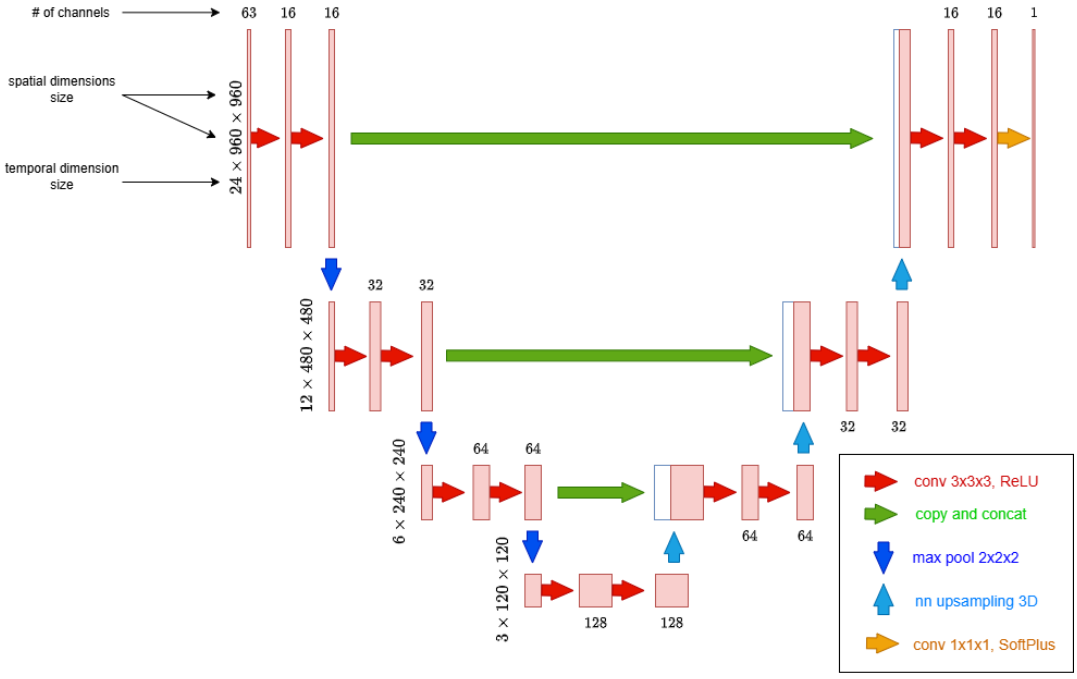

Figure 1. Visualization of the applied neural network architecture.

2.3 The study aims to estimate 24-hour, 100 m resolution $PM_{2.5}$ data, but most of the results presented are seasonal or monthly averages. We would like to see 24-hour $PM_{2.5}$ mapping results. Additionally, the comparison with MERRA2 focuses mainly on accuracy. Could the authors also better illustrate $PM_{2.5}$'s spatial distribution and gradient variations, or even capture specific pollution emissions?

While the goal of our work is to provide hourly $PM_{2.5}$ estimates at high spatial resolution, the main results focused on seasonal and monthly averages to better assess overall model performance. To address the reviewer's suggestion, we now include hourly $PM_{2.5}$ outputs in the revised manuscript.

Figure 2 presents an example of model performance at hourly resolution at Station 1 between 04.12.2019 and 13.12.2019, compared against MERRA-2 and OpenAQ observations. The comparison illustrates that our model captures the observed variability more accurately than MERRA-2.

Figure 3 shows hourly $PM_{2.5}$ concentration maps for the Paris region on 06.12.2019, with spatial patterns that agree well with OpenAQ station data. Figure 4 complements this by visualizing temporal gradients on the same day, highlighting a general pollution decrease across the area.

While we do not aim to interpret these results from a meteorological or atmospheric chemistry perspective, we note that some observed variations may be consistent with known events during this period (such as possible long-range sea salt transport,

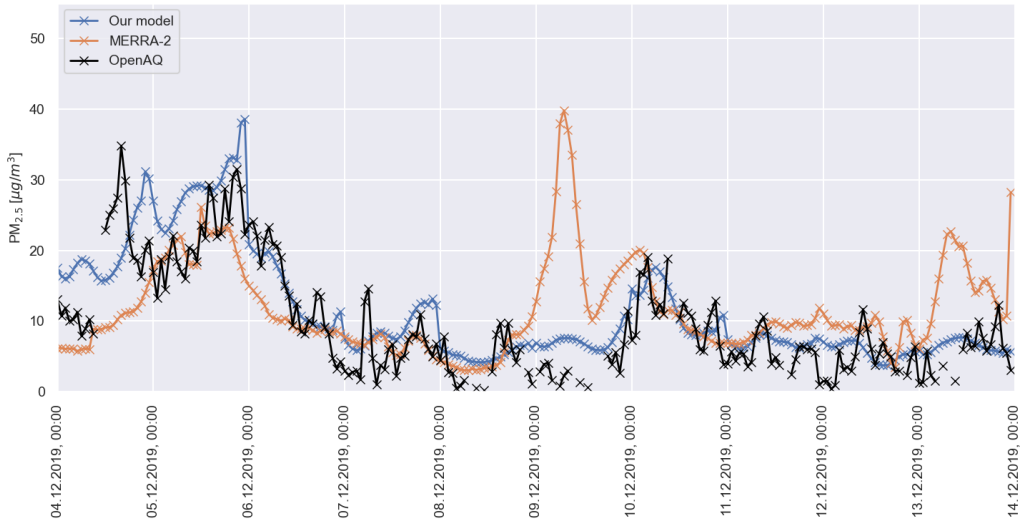

Figure 2. Comparison between hourly $PM_{2.5}$ estimates from our model (blue), MERRA-2 (orange) and OpenAQ ground stations measurements (black) at station 1. The period considered runs between 04.12.2019 and 13.12.2019.

organic matter peaks, temperature inversions, and rainfall events). These elements, as suggested by reanalysis data, could plausibly contribute to the patterns seen. However, our focus is on demonstrating the model's ability to reproduce such patterns at fine scale, rather than attributing them to specific processes.

2.4 The study applies explainable AI techniques to explore the importance of different features, showing that SHAP values identify 2-meter air temperature as the most important feature. However, this analysis could be further improved. First, the underlying reasons for why certain variables are important (or not) are not sufficiently explored. Second, a broader perspective could be considered—how much of the variability in $PM_{2.5}$ can be explained by meteorological variables overall?

Since our methodology is purely data-driven, a clear interpretation of how the input data affect the output is not straightforward, considering that many input features could act as proxy for other variables. Fig. 5 shows the feature importances determined by summing the normalized absolute SHAP values for predictions at station 1. In the manuscript we stated that T2M influences the temporal variability of $PM_{2.5}$ through boundary layer dynamics and contains information about seasonal emission changes. Considering that no global temporal information is present in the input features, T2M could act as proxy in this sense, and removing it could affect significantly time series trends predictions. QLML (surface specific humidity) and windspeed are again two variables that could be linked to aerosol deposition and transport. Specific aerosol variables such as BCCMASS (Black Carbon Column Mass Density), could give the model an idea of how much important black carbon concentration is for the final $PM_{2.5}$ estimate, but at the same time act as proxy for other species related to black carbon emission sources. Among the

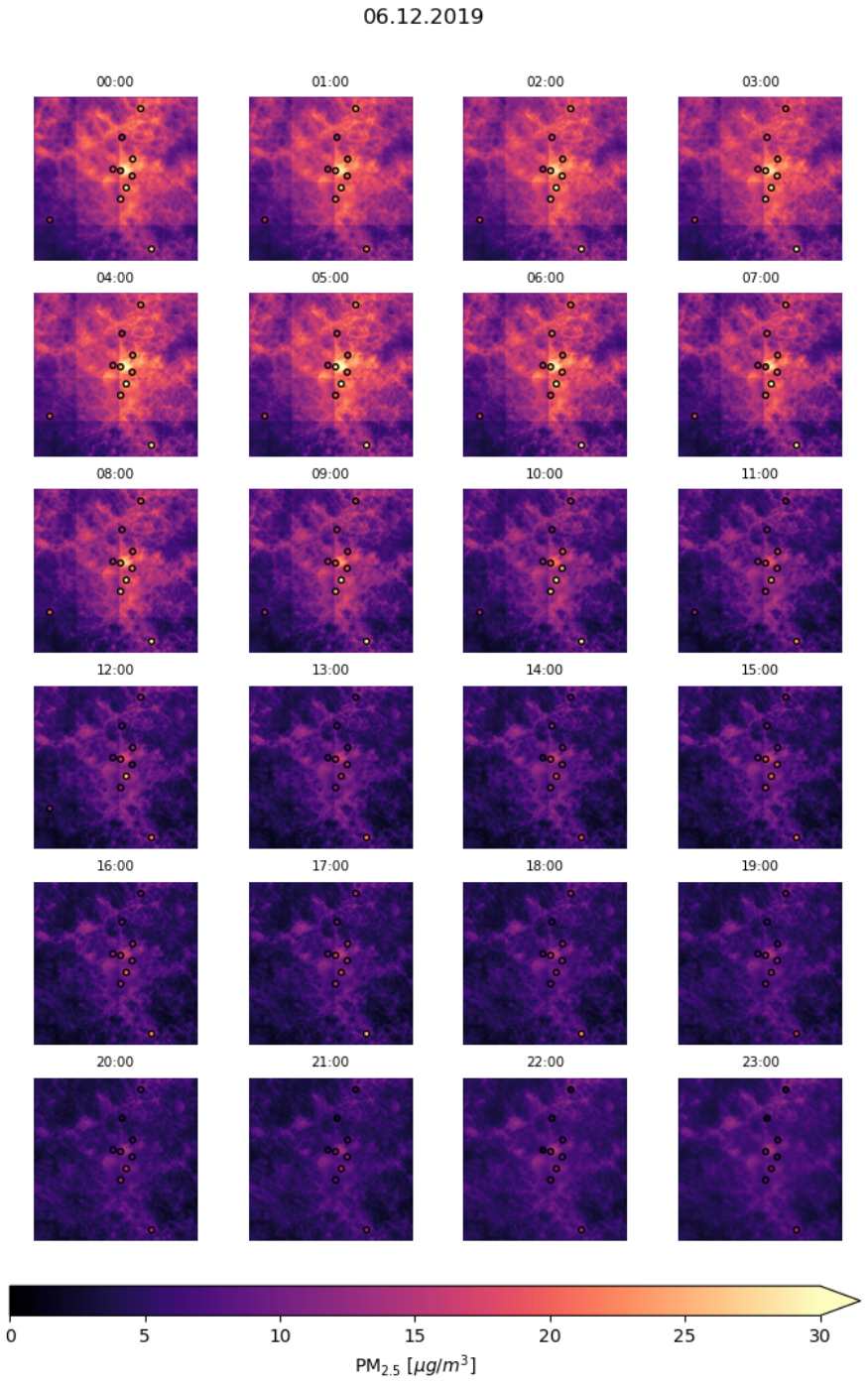

Figure 3. $PM_{2.5}$ map on 06.12.2019. The dots reprent $PM_{2.5}$ measurements from OpenAQ ground stations.

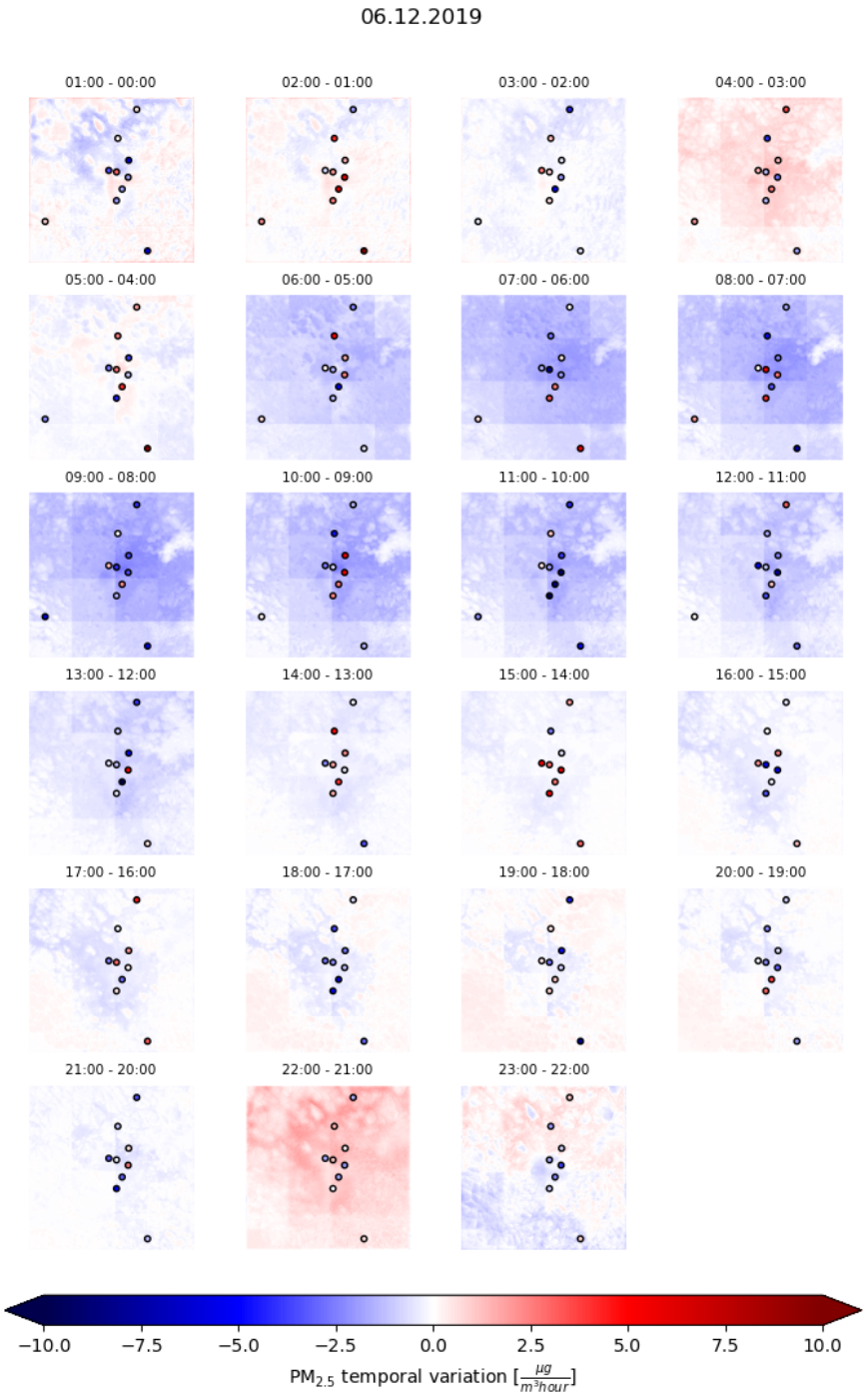

Figure 4. Temporal $PM_{2.5}$ gradients map on 06.12.2019. The dots reprent temporal gradients of the $PM_{2.5}$ measurements from OpenAQ ground stations.

most important high resolution input features, ASTERDEM (ASTER Digital Elevation Model) and BlackMarble (NASA Black Marble Night Lights) offer information about terrain topology and human activities location. While the former could provide useful information about aerosol transport, the latter could act as a proxy for aerosol sources spatial distribution. They both clearly contribute to the spatial distribution of $PM_{2.5}$ model output maps. More generally, aggregating the feature importances in Fig. 5, one can estimate the importance of athmosperic variables (35%), aerosol variables (25%) and high resolution indicators (40%). The discussion about SHAP explainability has been updated in the manuscript.

2.5 The description of NOODLESALAD $PM_{2.5}$ and its role in this study is unclear. The authors should provide a more detailed explanation rather than merely citing previous studies.

The description of NOODLESALAD $PM_{2.5}$ has been updated, the updated subsection 2.1 is as follows:

"NOODLESALAD $PM_{2.5}$ (Porcheddu et al., 2024) retrievals are obtained applying a deep learning based post-process correction approach to the MERRA-2 AOD-to-$PM_{2.5}$ conversion ratio. The post-process corrected AOD-to-$PM_{2.5}$ conversion ratio is utilized to map high resolution POPCORN SENTINEL-3 SYNERGY AOD estimate (Lipponen et al., 2022) to high resolution $PM_{2.5}$ estimate. The post-process correction of MERRA-2 AOD-to-$PM_{2.5}$ conversion ratio is carried out deploying an ensemble of fully-connected feed-forward neural networks and a fusion of surface in-situ $PM_{2.5}$ observations, MERRA-2 reanalysis model AOD and $PM_{2.5}$ data, spectral AERONET AOD, satellite-observed spectral top-of-atmosphere reflectances, meteorology data, and various high-resolution geographical indicators. The ensemble technique leads to a distribution of predictions for a single $PM_{2.5}$ estimate. The median of the ensemble is considered as the $PM_{2.5}$ estimate and the width of the distribution is regarded as an uncertainty related to the machine learning model training (model uncertainty). NOODLESALAD $PM_{2.5}$ offers high resolution on a grid with cell size 100 m x 100 m and is currently available for Sentinel-3A and 3B overpasses, covering Central Europe for the year 2019. The two Sentinel-3 satellites currently flying provide revisit times of less than two days for OLCI and less than one day for the SLSTR instrument at equator. Swath width of the OLCI instrument is 1270 km. SLSTR swath width is 1420 km for the nadir view and 750 km for the oblique view.

Evaluation metrics for $PM_{2.5}$ at satellite overpass (R^2=0.55, RMSE=6.2 $\mu g/m^3$) and $PM_{2.5}$ monthly averages (R^2=0.72, RMSE=3.7 $\mu g/m^3$) show good agreement between NOODLESALAD $PM_{2.5}$ and OpenAQ ground stations data (Porcheddu et al., 2024). Given the better spatial coverage compared to ground stations and the high spatial resolution at satellite overpass, we utilize NOODLESALAD $PM_{2.5}$ to inform the model about $PM_{2.5}$ fine spatial distribution. In this work, we consider NOODLESALAD $PM_{2.5}$ retrievals in Paris, France, in 2019, and utilize them as part of the target data to train our model."

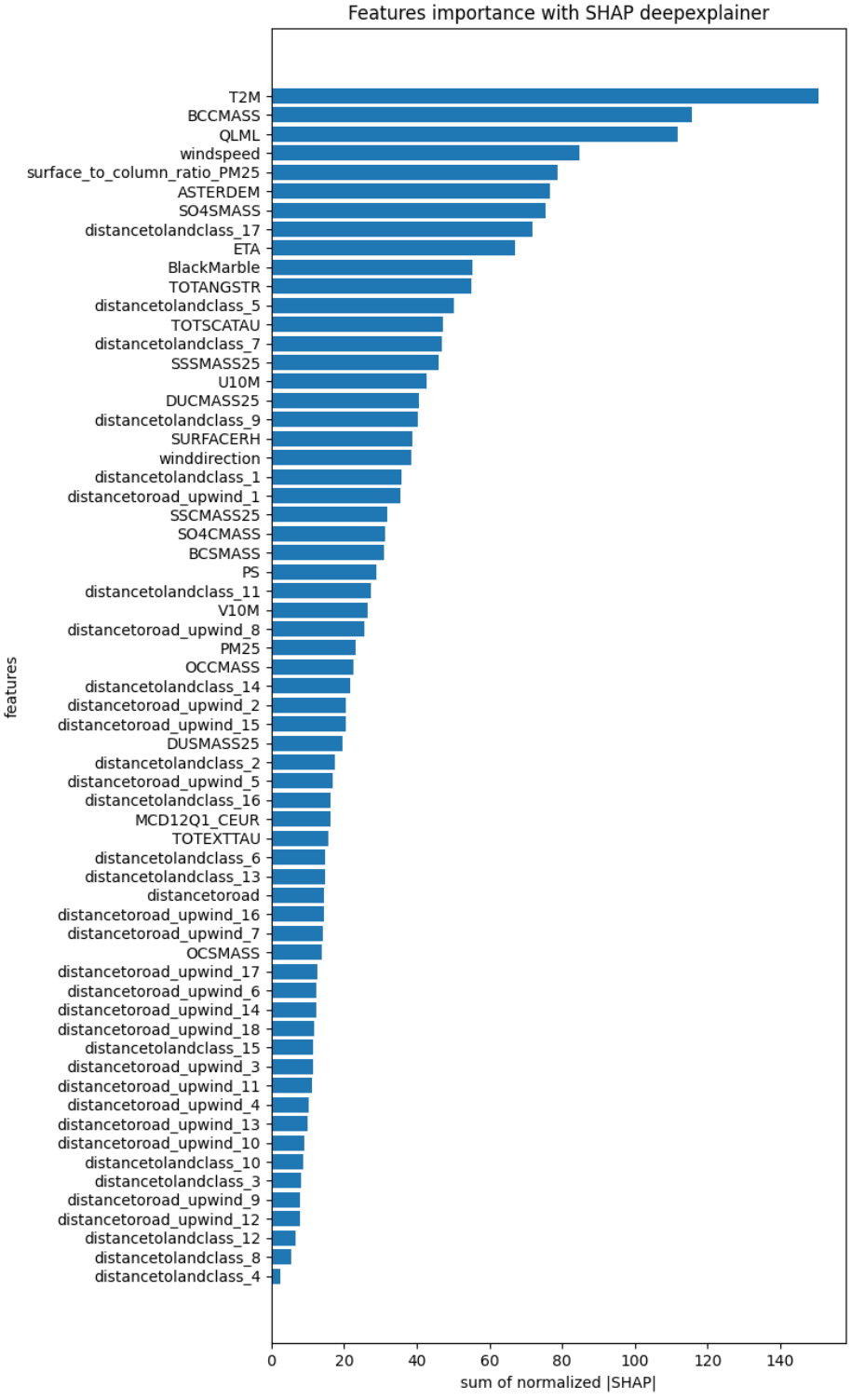

Figure 5. Feature importance calculated as sum of the normalized absolute SHAP values for predictions at station 1.

2.6 The results and analysis section could be further improved. First, it is recommended to structure the results into separate subsections rather than mixing everything together. Second, the quality of Figures 3–6 should be improved—currently, the font size is too small, and the figure titles could be removed (since the descriptions are already included in the captions). Lastly, additional results, such as 24-hour high-resolution $PM_{2.5}$ maps, could enhance the persuasiveness of the study.

Subsections have been added to improve the structure of the results section. The figures have been improved and updated in the manuscript. Figures showing high temporal resolution have been added to the manuscript as discussed in our reply in Section 2.3.

2.7 The references in the paper are somewhat outdated, with few studies from the recent three years included. It is recommended to update and supplement them.

To complement the short discussion related to fine particulate matter health risks, we added a reference to a recent article (Thangavel et al., 2022). A short discussion of machine learning methods exploiting spatio-temporal correlations has been added to the manuscript, referencing two recent papers (Koo et al., 2024; Muthukumar et al., 2022). One of these two papers has been referenced also when discussing $PM_{2.5}$ interpolation using ground monitoring stations, as a useful example of how the kriging method is utilized in the literature for this task (Koo et al., 2024). A recent book has been added as complementary reference for deep learning architectures and techniques (Bishop and Bishop, 2024).

2.8 Some minor issues: (1) Figure 1: Does the figure represent the road network? Please clarify. (2) Line 134: "3D $PM_{2.5}$ maps" could be misinterpreted as three-dimensional spatial maps (including altitude). Is this the correct terminology? (3) Figure 2: The representation is somewhat abstract. It would be better if the inputs and outputs were explicitly illustrated. (4) Line 279: "consistent with prior findings" should be supported with references.

1) Figure 1 represents a map of the region of interest, where the position of the ground monitoring stations is represented relatively to the road network. 2) To clarify, we updated the manuscript writing "time series of surface $PM_{2.5}$ maps (3D $PM_{2.5}$ arrays, two spatial dimensions and one time dimension)". 3) The architecture visualization has been improved for clarity (Fig. 1 in this document). A new figure (Fig. 6 in this document) has been added to complement the architecture visualization and highlight the data flow in our study. 4) Here "consistent with prior findings" is an auto-reference: the link between variations of $PM_{2.5}$ levels and variations of the boundary layer height has been discussed when referencing the maps showing $PM_{2.5}$ distributions (Figure 5 and 6).

3 Answers to reviewer #2

3.1 The data accuracy of NOODLESALAD PM2.5 should be described in section 2.1. Moreover, what are essential roles of this unique product in the proposed deep learning framework, needs to clarify

We used NOODLESALAD $PM_{2.5}$ primarily because it was immediately available as a result of our earlier work and offers high spatial resolution along with demonstrated accuracy. In that sense, this study can be seen as a follow-up to that earlier effort, where a satellite-derived $PM_{2.5}$ product like NOODLESALAD serves as the input to the second-stage data fusion model.

The methodology we propose is not dependent on NOODLESALAD specifically: any comparable satellite-derived $PM_{2.5}$ product could be used in its place, depending on availability and regional suitability. We do not claim NOODLESALAD is the only or best option, but rather one example suitable for our study region.

Regarding the accuracy of NOODLESALAD $PM_{2.5}$, we've updated Section 2.1 to include validation metrics that demonstrate its performance, as you suggested.

3.2 Since the authors only used 11 stations for reference, is this adequate to depict PM2.5 variability across space in the study area?

It's true that the number of ground stations (11 in total) is relatively small, and that's actually one of the key reasons behind this work. The goal was to explore how satellite and geospatial data can help fill in the gaps where monitoring stations are sparse or completely missing.

That said, we're aware that more stations would provide a stronger basis for both training and validation. To make the most of the available data, we used a leave-one-out cross-validation strategy, which allowed us to evaluate how well the model performs across the different locations. The results suggest that the model is able to generalize reasonably well, combining multiple data sources to estimate $PM_{2.5}$ patterns that align with the ground observations.

We agree that having a denser network of stations would open up possibilities for further analysis—for instance, studying how sensitive the model is to the spatial distribution of the training data. This is something we'd like to look at in future work. But overall, we believe this study shows that even with limited in-situ data, it's possible to make meaningful improvements in air quality estimation using a data fusion approach.

3.3 MERRA-2 PM2.5 estimates: since no nitrates are provided in MERRA-2 aerosol diagnostics, the corresponding PM2.5 estimates are prone to large uncertainty. The data accuracy of this PM2.5 product should be validated as well.

We're aware of the limitations in MERRA-2 $PM_{2.5}$ estimates, including the absence of nitrate aerosol components, which can lead to uncertainty. That said, our study does not rely on MERRA-2 as a definitive data source: we use it as one example of low-resolution geophysical model output to demonstrate how our data fusion approach can improve upon such sources.

In principle, any similar model product could be used in place of MERRA-2: the core of the study is the methodology for combining multiple data sources, not the evaluation of a specific model dataset. MERRA-2 was selected because it is widely used in air quality research and has been previously validated in multiple studies (Buchard et al., 2017; Jin et al., 2022)

We believe the conclusions of the study are not tied to this particular dataset, and future applications of the method could incorporate other chemical transport models depending on availability and regional relevance.

3.4 The authors used a set of geographic variables with varying spatial resolution, how did the authors collocate them in the deep learning framework, no such descriptions.

To ensure consistency across inputs, we first regridded all geographic variables with the original grid size larger than 100 meters to a common 100-meter resolution grid using the Universal Transverse Mercator (UTM) projection. Linear interpolation method was used for continuous variables and nearest neighbors for categorical ones. This preprocessing step ensured accurate spatial collocation of all features prior to input into the deep learning model.

To clarify this detail, we have added the following text into the manuscript: "All geographic variables with the original resolution larger than 100 m were regridded to a common spatial grid with a resolution of 100 m using the Universal Transverse Mercator (UTM) projection. Linear interpolation method was used for continuous features and nearest neighbor interpolation for categorical variables. This preprocessing ensured that all features were spatially collocated prior to input into the deep learning model."

3.5 A flow chart depicting the deep learning architecture, particularly the data flow, is essential for understanding and reproducibility.

A new figure (Fig. 6 in this document) highlighting the data flow and complementing the architecture visualization has been added to the manuscript. The architecture visualization has also been improved for clarity (Fig. 1 in this document).

3.6 Equations should be numbered.

We revised the manuscript and numbered the equations.

3.7 Methodology: the authors mentioned that both satellite- and ground-based PM2.5 data were used as the learning target. Since these datasets have distinct data accuracy, would this undermine the learning capacity of the deep learned model?

This is a valid point: satellite-derived and ground-based $PM_{2.5}$ estimates do indeed differ in their accuracy, and ideally, this would be accounted for in the training process (e.g., through a weighted loss function based on uncertainty). Unfortunately, explicit uncertainty estimates for the satellite-derived $PM_{2.5}$ were not available, so we treated both data sources equally in the training phase.

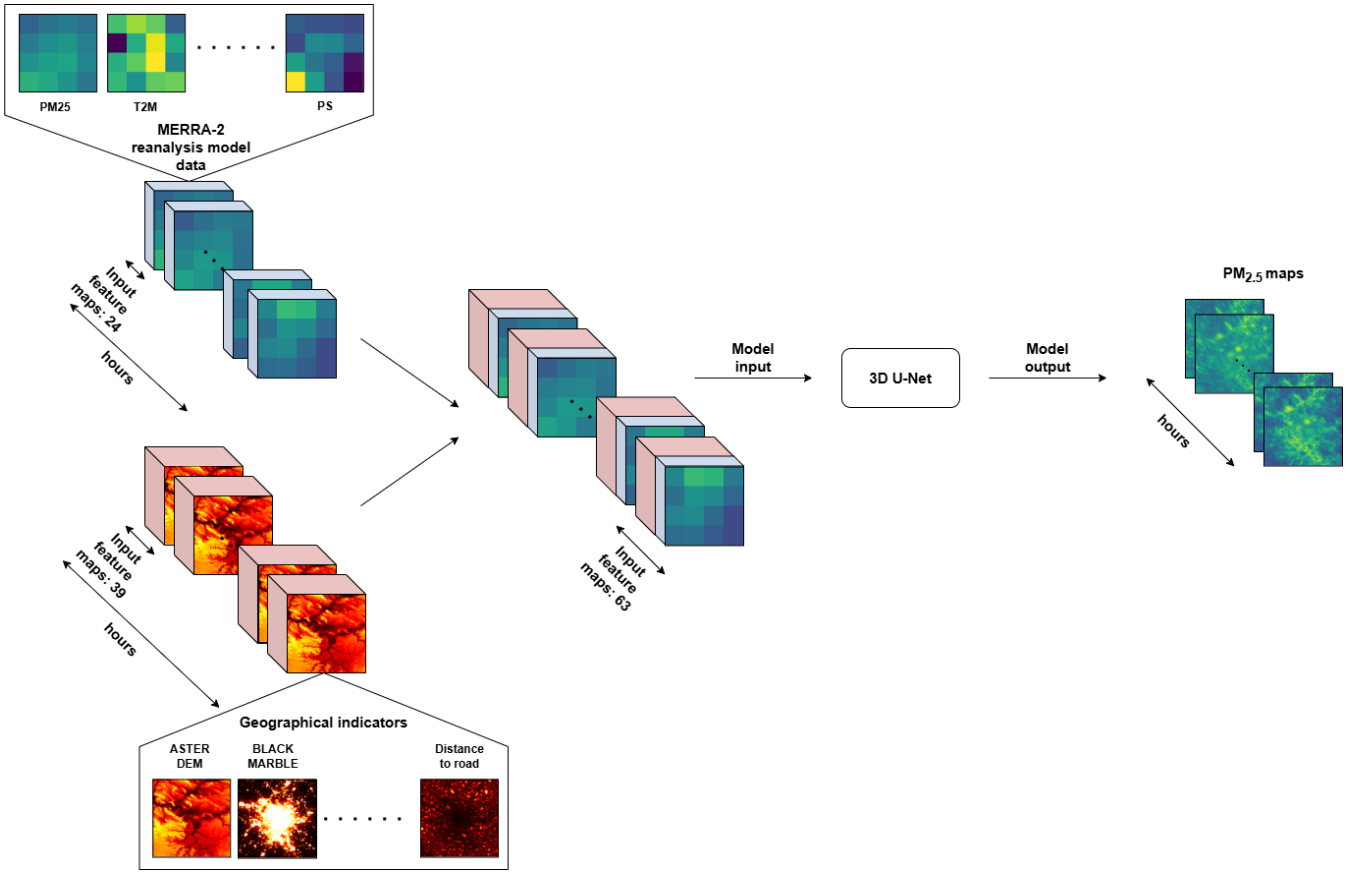

Figure 6. Visualization of the data flow in our method. Low spatial resolution (MERRA-2) data and high spatial resolution geographical indicators are projected on a common grid, joined and utilized as model input. The model output consists of hourly $PM_{2.5}$ maps.

That said, we chose to include both because they offer complementary strengths: ground-based measurements provide accurate point-wise information, while satellite estimates improve spatial coverage, especially in areas with few or no ground stations.

While incorporating uncertainty information would likely improve model performance, our results suggest that the model learning capability is not undermined: the model is still able to learn meaningful patterns from the combined data. The consistent improvement over baseline estimates, especially in cross-validation, indicates that the learning process is robust—even without explicitly modeling the uncertainty in the targets.

We agree that this is an interesting direction for future work and could lead to better integration of heterogeneous data sources in deep learning models.

3.8 Line 207-209: this would result in imbalanced training sets at different hours, which could also influence the learning accuracy, as the learned model is more likely to predict PM2.5 during the satellite overpasses.

We agree that the temporal imbalance introduced by satellite overpasses (where more data is available at specific times of day) could affect model learning, potentially biasing predictions toward those hours. This is a valid concern, especially when combining satellite-derived data (rich in spatial detail but temporally sparse) with ground station measurements (temporally dense but spatially limited).

To address this, we designed a loss function that explicitly balances the contributions of the different data sources. The aim is to prevent the model from overfitting to satellite data patterns at the expense of learning broader temporal dynamics from the ground stations.

Further details on the loss function and how it handles this trade-off are provided in Section 2.2, as part of our reply to Reviewer #1.

3.9 An intercomparison of spatial distribution of predicted PM2.5 estimates from MERRA-2 with satellite-derived PM2.5 at 100-m from Sentinel observations should be provided to assess the reliability of the proposed model in resolving PM2.5 distributions in Paris.

To compare and highlight the benefit of using satellite data, we trained another model using only ground stations as target data (so removing satellite $PM_{2.5}$ from the training) and keeping the rest of the methodology (e.g. same model architecture we considered for our model).

Figure 7 compares a NOODLESALAD $PM_{2.5}$ map (at single satellite overpass) to our model output and the output obtained removing satellite $PM_{2.5}$ from the training. Further, we considered all NOODLESALAD $PM_{2.5}$ maps contained in the validation set and calculated RMSE values per pixel, in order to estimate how well our model and the model trained without satellite data can reproduce the NOODLESALAD $PM_{2.5}$ spatial patterns (as illustrated in Fig. 8). Averaging the RMSE values per pixel, we obtained 4.57 $\mu g/m^3$ for our model, and 5.69 $\mu g/m^3$ when training without satellite $PM_{2.5}$. Both the model were trained leaving out station 1.

These results suggest that our model is able to capture the spatial information contained in NOODLESALAD $PM_{2.5}$ data.

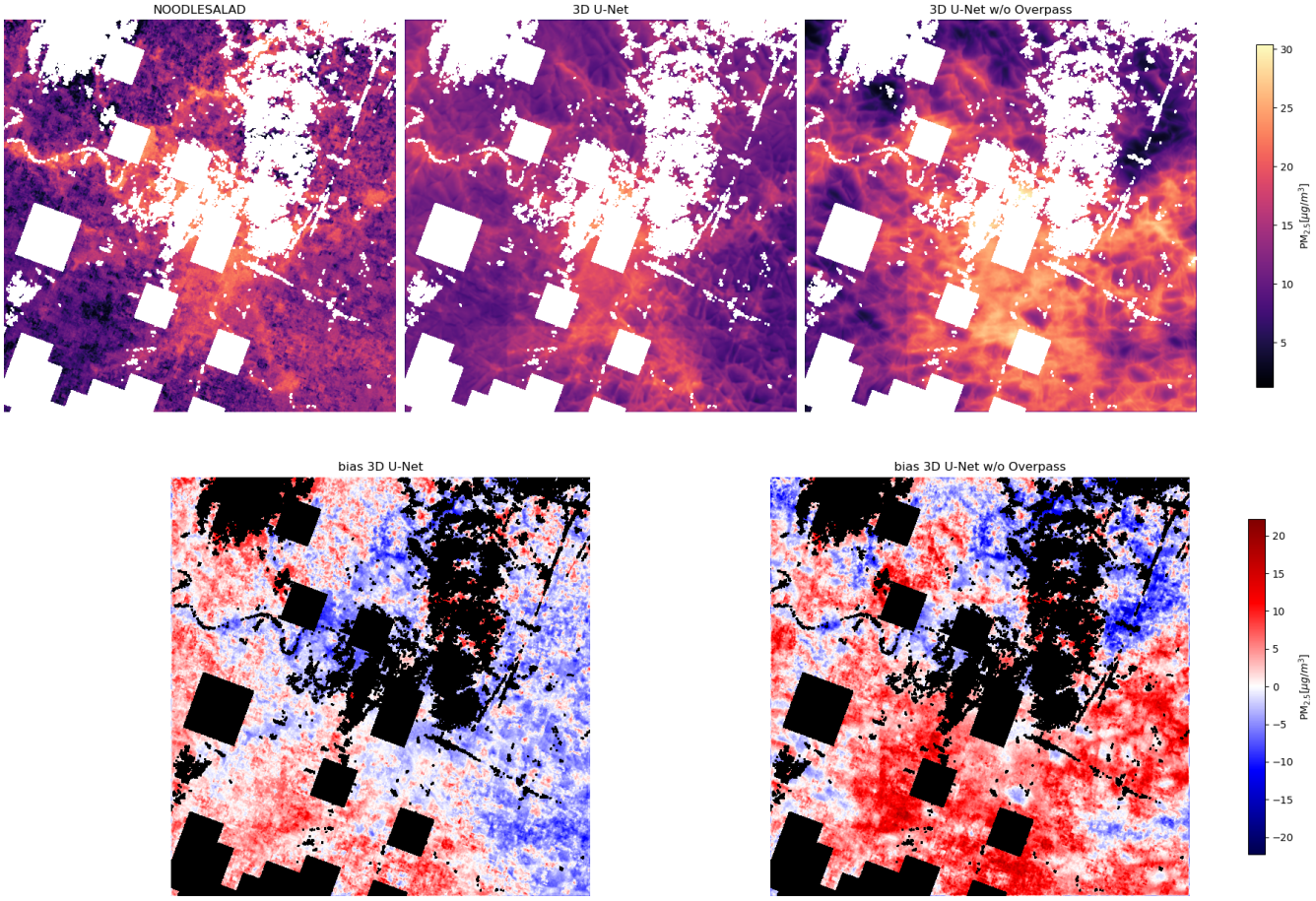

Figure 7. On the top: comparison between NOODLESALAD $PM_{2.5}$ (left), our model (center) and another model trained without satellite $PM_{2.5}$, at one single satellite overpass. On the bottom: bias calculated comparing our model output to the NOODLESALAD $PM_{2.5}$ map (left), bias calculated comparing another model trained without satellite $PM_{2.5}$ to the NOODLESALAD $PM_{2.5}$ map (right). The NOODLESALAD $PM_{2.5}$ map is taken from the validation set.

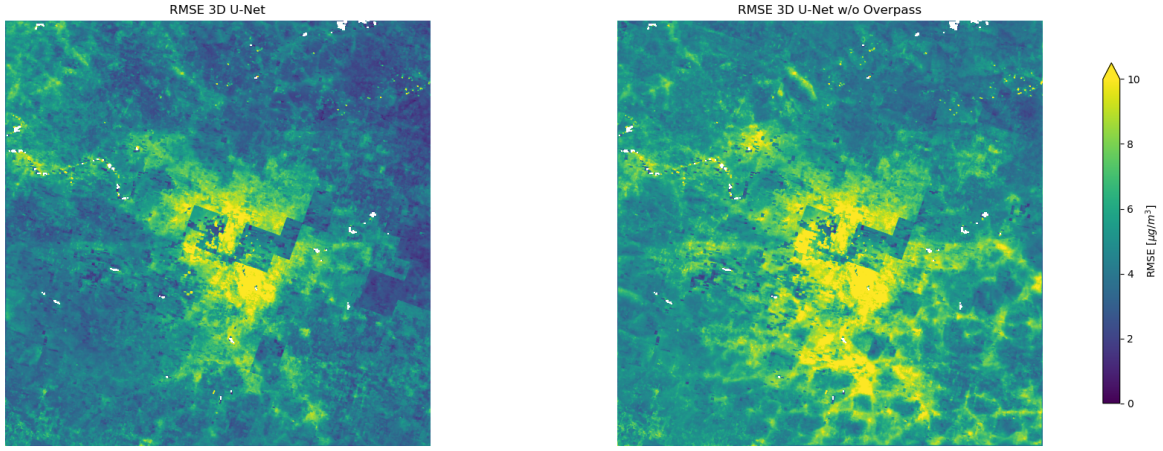

Figure 8. On the left: RMSE per pixel comparing our model to NOODLESALAD $PM_{2.5}$ maps in the validation set. On the right: RMSE per pixel comparing another model trained without satellite $PM_{2.5}$ to NOODLESALAD $PM_{2.5}$ maps in the validation set.

References

Bishop, C. M. and Bishop, H.: Deep Learning: Foundations and Concepts, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-031-45468-4, 2024.

Buchard, V., Randles, C. A., da Silva, A. M., Darmenov, A., Colarco, P. R., Govindaraju, R., Ferrare, R., Hair, J., Beyersdorf, A. J., Ziemba, L. D., and Yu, H.: The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies, Journal of Climate, 30, 6851 – 6872, https://doi.org/10.1175/JCLI-D-16-0613.1, 2017.

Jin, C., Wang, Y., Li, T., and Yuan, Q.: Global validation and hybrid calibration of CAMS and MERRA-2 PM2.5 reanalysis products based on OpenAQ platform, Atmospheric Environment, 274, 118 972, https://doi.org/https://doi.org/10.1016/j.atmosenv.2022.118972, 2022.

Koo, J.-S., Wang, K.-H., Yun, H.-Y., Kwon, H.-Y., and Koo, Y.-S.: Development of PM2.5 Forecast Model Combining ConvLSTM and DNN in Seoul, Atmosphere, 15, 1276, https://doi.org/10.3390/atmos15111276, number: 11 Publisher: Multidisciplinary Digital Publishing Institute, 2024.

Lipponen, A., Reinvall, J., Väisänen, A., Taskinen, H., Lähivaara, T., Sogacheva, L., Kolmonen, P., Lehtinen, K., Arola, A., and Kolehmainen, V.: Deep-learning-based post-process correction of the aerosol parameters in the high-resolution Sentinel-3 Level-2 Synergy product, Atmospheric Measurement Techniques, 15, 895–914, 2022.

Muthukumar, P., Cocom, E., Nagrecha, K., Comer, D., Burga, I., Taub, J., Calvert, C. F., Holm, J., and Pourhomayoun, M.: Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data, Air Quality, Atmosphere, & Health, 15, 1221–1234, https://doi.org/10.1007/s11869-021-01126-3, 2022.

Porcheddu, A., Kolehmainen, V., Lähivaara, T., and Lipponen, A.: Post-process correction improves the accuracy of satellite $PM_{2.5}$ retrievals, Atmospheric Measurement Techniques, 17, 5747–5764, https://doi.org/10.5194/amt-17-5747-2024, 2024.

Thangavel, P., Park, D., and Lee, Y.-C.: Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview, International Journal of Environmental Research and Public Health, 19, 7511, https://doi.org/10.3390/ijerph19127511, 2022.