https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

A hybrid PMF-machine learning approach for resolving urban black carbon sources from multi-wavelength absorption measurements

Xueyan Wang[a,b], Lu Lei[c], Vignesh Prabhu[c], Wei Xu*[a,b], Jian Zhao[d], Chunshui Lin[e], Kirsten N. Fossum[c], Baihua Chen[a], Darius Ceburnis[c], Colin O'Dowd[c], Jurgita Ovadnevaite*[c]

[a]State Key Laboratory of Advanced Environmental Technology, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China
[b]University of Chinese Academy of Sciences, China
[c]School of Natural Sciences, Physics, Ryan Institute's Centre for Climate & Air Pollution Studies, University of Galway, Galway, Ireland
[d]Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland
[e]State Key Laboratory of Loess Science, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, China.

Correspondence: wxu@iue.ac.cn; jurgita.ovadnevaite@universityofgalway.ie

Abstract

Source apportionment of urban black carbon (BC) is challenging because co-emitted combustion sources share overlapping optical and chemical signatures, and conventional Aethalometer approaches require prescribed absorption Ångström exponents (AAEs). Here we combined positive matrix factorization (PMF/ME-2) with a physically constrained machine-learning model to resolve urban BC into policy-relevant emission sectors. At an urban background site in Dublin, Ireland, PMF/ME-2 separated total equivalent BC (eBC) into fossil-fuel (eBC-FF) and solid-fuel (eBC-SF) factors directly from 7-wavelength observations, yielding AAEs of 1.1 and 2.0 without prior assumptions. A neural-network–linear mixed model then apportioned eBC-FF into traffic and oil-heating contributions, reproducing observed eBC-FF (R^2 = 0.79) with a coefficient "k" = 0.730 ± 0.036 that shifted by at most 18% across the predictor sets tested without reordering the three sectors. Traffic, oil heating and solid fuel accounted for 59.2%, 17.2% and 23.5% of total eBC. Collocated ACSM factors for peat and wood burning correlated far more strongly with eBC-SF (r = 0.83, 0.76) than with traffic (0.42, 0.44), independently supporting the solid-fuel assignment of eBC-SF. Solid-fuel BC therefore far exceeded the ~2% of Dublin households reporting solid fuel as their primary heating fuel, showing that energy statistics alone underestimate this source; the oil-heating estimate remains conditional on the assumed proportionality between the two heating fuels.

Key words: black carbon, source apportionment, PMF, machine learning

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

1. Introduction

Black carbon (BC), a typical short-lived climate forcer exhibiting strong light absorption and positive radiative forcing, is known to affect global warming and the melting of snow and ice (IPCC, 2014). It is primarily emitted from the incomplete combustion of biomass and fossil fuels (Saarikoski et al., 2021), with its particle size distribution predominantly exhibiting an accumulation mode, characterized by a small deposition velocity, long residence time in the atmosphere, and a propensity for long-range transport (Change, 2007), BC typically coexists with organic aerosol (OA) as a major component of carbonaceous aerosols (Jimenez et al., 2009; Grigas et al., 2017), which often account for a substantial fraction of fine particulate matter (PM_1 or $PM_{2.5}$, i.e. particulate matter with aerodynamic diameters less than 1 μm and 2.5 μm, respectively) mass (Bond et al., 2013). Owing to its combined impacts on human health and the climate system (Janssen et al., 2012; Chen et al., 2025b), BC has been recognized as a priority air pollutant that requires urgent mitigation, and has attracted increasing attention from both the scientific community and policymakers in recent years (Organization, 2013). Therefore, understanding the sources of BC is critical for cost-effective mitigation. This is especially important in urban environments where multiple combustion sources coexist and their contributions often overlap in both chemical composition and temporal behavior.

In urban environments, BC is commonly associated with combustion sources such as road traffic and residential heating, although the relative importance of these sources varies strongly among cities. Dublin experienced black smoke concentrations exceeding 750 μg/m³ during the 1980s mainly caused by coal combustion. Although the 1990 coal ban brought substantial improvement, extreme wintertime PM_1 episodes exceeding 300 μg/m³ still occur as a result of intense home heating emissions in winter, with black carbon (BC) accounting for a substantial fraction of the particulate mass, alongside other carbonaceous components such as organic aerosol (OA) (Goodman et al., 2009; Lin et al., 2017). These observations indicate that BC remains an important component of urban air pollution in Dublin, particularly during the heating season. In the Dublin urban atmosphere, BC originates primarily from traffic emissions and residential heating, the latter involving both solid fuel and oil combustion.(Lin et al., 2018). This creates a mixed fossil fuel signal — traffic and heating contributions coexist within black carbon from fossil fuel combustion, yet they have similar optical and chemical characteristics, making them difficult to distinguish using conventional source apportionment alone (Lei et al., 2025; Lin et al., 2020). Moreover, energy-use statistics can be misleading: while only a small fraction of Dublin households report solid fuel as their primary heating fuel, due to the disproportionate emission factors of black carbon and unfavorable diffusion conditions at night, even a small fraction of solid fuel usage can dominate ambient BC concentrations (Central Statistics Office, 2021). These complexities underscore the need for a more refined approach capable of resolving the fossil fuel fraction into policy-relevant emission sectors.

A range of methods have been developed to identify BC sources. Radiocarbon analysis can distinguish fossil carbon from contemporary carbon because fossil fuel

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

emissions are depleted in ^{14}C (Szidat et al., 2006; Mouteva et al., 2017), and has been widely applied to quantify fossil fuel and biomass burning contributions to carbonaceous aerosols (Ni et al., 2018; Winiger et al., 2019). Stable carbon isotopes can provide additional information on fuel type and combustion conditions (Ni et al., 2018). Particle-resolved microscopic approaches further identify source-related morphological features (Xu et al., 2020). For example, soot particles formed through high-temperature gas-phase processes commonly appear as chain-like or cluster-like aggregates of small carbonaceous spherules, whereas char-like particles from solid fuel combustion or incomplete combustion often retain more irregular structures (Han et al., 2010). Biomass burning can also produce characteristic carbonaceous particles such as tar balls (Pósfai et al., 2004). Although these approaches provide relatively direct source information, they are often limited by sampling cost, temporal resolution, and representativeness for long-term continuous observations, making them difficult to apply to multi-year high-time-resolution source apportionment (Winiger et al., 2019; Xu et al., 2020).

In contrast, optical and receptor-based methods, such as the Aethalometer and positive matrix factorization (PMF), are more suitable for long-term BC observational datasets. The conventional Aethalometer model separates BC into fossil fuel and biomass combustion components based on wavelength-dependent light absorption. However, this method requires prescribed source-specific absorption Ångström exponents (AAEs), and the apportioned source contributions can be highly sensitive to these assumed values. This is mainly because source optical properties may vary with fuel type, combustion conditions, atmospheric processing, and the mixing of multiple sources, fixed AAE assumptions can introduce substantial uncertainty and may even lead to physically unrealistic source estimates (Sandradewi et al., 2008; Drinovec et al., 2015; Paatero and Tapper, 1994; Hopke, 2016). Positive matrix factorization (PMF), which is commonly used for OA source apportionment (Ulbrich et al., 2009; Paatero and Tapper, 1994), provides a more data-driven receptor-modeling approach by decomposing the observational matrix into factor profiles and temporal contributions based on their covariance structure and uncertainty-weighted residuals. When applied to multi-wavelength absorption coefficients, PMF can resolve source-related absorption profiles directly from the measurements, allowing factor-specific AAEs to be derived as model outputs rather than prescribed inputs. This makes PMF particularly useful for long-term AE33 datasets and provides a stronger physical basis. Although PMF is a powerful receptor model, it has inherent limitations and its ability to separate sources is largely limited when different sources have highly similar factor profiles and/or temporal variations. This is particularly relevant for traffic emissions and oil-heating emissions, both of which contribute to fossil-fuel-related eBC and can share similar optical and chemical characteristics. Therefore, PMF alone cannot further separate eBC-FF into traffic- and heating-specific contributions. However, this distinction is important because traffic and residential oil heating are associated with different activity patterns, emission drivers, and mitigation strategies. Without clear separation, the fossil-fuel fraction remains unresolved at the level of policy-relevant emission sectors, limiting the

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

assessment of targeted traffic and residential-heating controls.

In recent years, machine learning has been increasingly applied in aerosol source apportionment to extract complex relationships and source-related information from multidimensional observational datasets that are difficult to resolve using conventional approaches. A key advantage of such approaches is their ability to fully capture the nonlinear, high-dimensional relationships between atmospheric variables and source contributions, which conventional linear or tracer-based methods cannot readily represent (Wang et al., 2025). In this study, long-term observations from an urban background site in Dublin, together with meteorological, traffic-count, temporal, and pollutant information, were used to further resolve the fossil-fuel black carbon component (eBC-FF) that remains unresolved by traditional source apportionment methods. In Ireland, road traffic has also been identified as an important contributor to urban aerosol pollution, further supporting the need to distinguish traffic-related fossil-fuel BC from heating-related contributions (Lin et al., 2020). To improve physical interpretability, we developed a neural-network-linear mixed model, in which the unknown fossil fuel heating component is represented using the known heating-related product (eBC-SF), while the traffic-related component is estimated from orthogonalized predictors. The objective is to separate (eBC-FF) into traffic- and heating-related contributions, thereby improving the interpretation of black carbon sources at policy-relevant emission-sector levels. By combining PMF-derived source categories with a physically constrained machine-learning framework, this hybrid approach provides more specific scientific evidence for urban BC mitigation.

2. Experiments and methods

2.1 Data sampling

The sampling site was located at two urban background sites in Dublin, Ireland from August 2016 to December 2023 (initially at the Science Center North in University College Dublin (UCD, 53.31 °N, 6.22 °W), and then was relocated to Trinity College Botanical Gardens (TCBG, 53.31 °N, 6.26 °W) from 6th September 2023 onward, approximately 3 km away from the UCD site), which are shown in Fig 1 (Lei et al., 2025). Both sites are urban background locations ~3 km apart, representative of the same urban background air mass, and were combined into a single record. The main instrument was a 7-wavelength Aethalometer (model AE33, Magee Scientific). The AE33 measures light attenuation across seven wavelengths (370, 470, 520, 590, 660, 880, and 950 nm) and employs a patented dual-spot measurement technique in which aerosol is collected on two filter spots at different flow rates, allowing real-time compensation of the aerosol loading effect that can otherwise bias attenuation-based BC measurements. The instrument was operated with a multiple-scattering parameter (C) of 1.39 (Drinovec et al., 2015). Loading-compensated attenuation at 880 nm is converted to equivalent black carbon (eBC) at a 1-minute time resolution, and the multi-wavelength absorption data are

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

also used for source apportionment (Ni et al., 2018).

In addition, a Quadrupole Aerosol Chemical Speciation Monitor (Q-ACSM) was collocated with the AE33, which measures the non-refractory components in PM_1 (Ng et al., 2011). The sampling periods covered August 2016 to March 2018 and 2021 to 2023. Seasons were defined following the meteorological convention as spring (March - May), summer (June - August), autumn (September - November), and winter (December - February). Meteorological parameters (Dublin Airport, https://www.met.ie/climate/available-data/historical-data) and traffic data were obtained from local monitoring platforms in Dublin (https://data.smartdublin.ie/dataset/traffic-signals-and-scats-sites-locations-dcc). The gaseous pollutants data including nitrogen oxides (NO_x) was obtained from a nearby Environment Protection Agency monitoring station (Rathmines; https://eparesearch.epa.ie/safer/index.jsp). All variables were aggregated to a common hourly resolution to match the model time step.

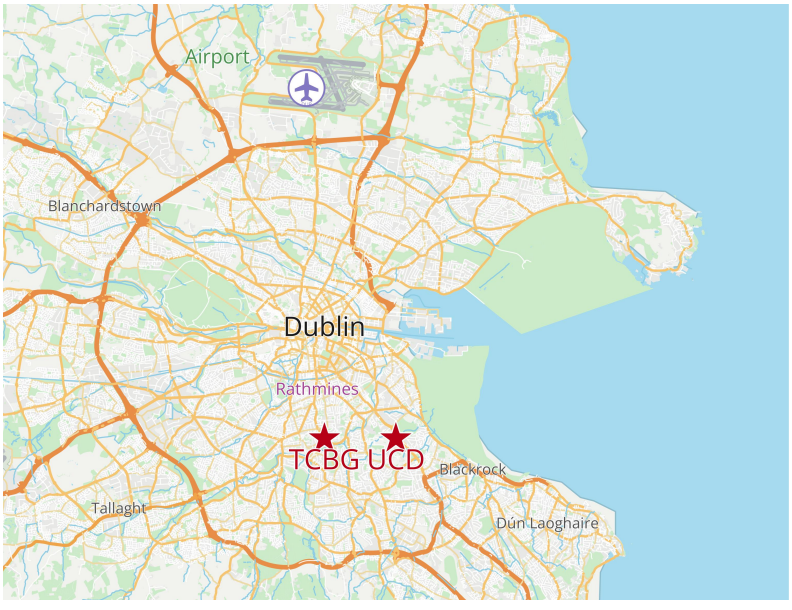


Fig 1. Map of the sampling sites and supporting monitoring stations in Dublin. The AE33 Aethalometer was initially operated at University College Dublin (UCD) and relocated to Trinity College Botanical Gardens (TCBG) in September 2023.

2.2 Positive matrix factorization analysis

The Aethalometer model is commonly used to attribute the total eBC into fossil-fuel combustion and biomass burning based on the wavelength dependence of light absorption (Sandradewi et al., 2008). However, this approach requires priori source-specific AAE values, and the apportioned source contributions can be highly sensitive to these pre-assumed values. Temporal variability in the true source-specific AAEs may therefore introduce considerable uncertainty and may even result in

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

physically unrealistic negative source contributions. In contrast, PMF applied directly to multi-wavelength absorption coefficients can resolve source-related absorption profiles and their temporal contributions from the measurements themselves, without prescribing fixed AAEs. This makes PMF a more flexible and data-driven approach for eBC source apportionment, while the factor-specific AAEs can be derived from the resolved factor profiles and used for source interpretation.

Unlike the Aethalometer model which requires prescribed AAEs, PMF resolves source-related absorption profiles directly from the multi-wavelength measurements. In this study, the 7-wavelength absorption data collected by the Aethalometer at the Dublin urban background sites were prepared for PMF analysis following the procedure proposed in Yttri et al. (2021) (Yttri et al., 2021). In brief, the PMF input matrix was constructed from the AE33 7-wavelength absorption coefficients. The corresponding uncertainty matrix was estimated for each wavelength by considering instrumental uncertainties and the variability of blank attenuation signals derived from regular clean-air tests (Paatero and Tapper, 1994; Yttri et al., 2021). Since the reference eBC-FF profile showed only minor variability (Fig A1), the ME-2 a-value, the ME-2 constraint parameter governing how strongly factor profiles are pulled toward the reference profile, was varied in a relatively narrow range (0-0.2) for all ME-2 analyses. The final solution was selected based on the physical interpretability of the factor profiles, the temporal behavior of the factor time series, the residual structure, and the agreement with independent source indicators. The factor-specific AAE was then calculated from the absorption coefficients at 470 and 950 nm.

In addition to the source apportionment of eBC, PMF analysis was also performed on the organic aerosol (OA) matrix measured by the Q-ACSM. More details of the OA PMF analysis can be found in previous studies (Lin et al., 2021; Lei et al., 2025). Briefly, six OA factors were identified, including four primary OA (POA) factors, namely peat, wood, coal, and hydrocarbon-like OA (HOA), as well as two oxygenated OA (OOA) factors, namely less oxidized OOA (LO-OOA) and more oxidized OOA (MO-OOA). The temporal variations of the OA factors, particularly the POA factors, which are often closely associated with eBC sources, were used as important supporting information to identify local source patterns, including both residential heating and traffic emissions.

2.3 Model development

To further partition eBC-FF into traffic and heating sources, three candidate model formulations were evaluated in this study: a counterfactual formulation, a direct additive formulation, and an orthogonalized additive formulation. These formulations differed in how the traffic-related fossil fuel component was estimated and how the heating-related fossil fuel component was represented. The construction of each formulation is described below, while their diagnostic comparison is presented in the Results and Discussion section.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

2.3.1 Counterfactual model

The first approach was a counterfactual formulation. This approach follows the logic of reference-based source-apportionment methods (Lei et al., 2025; Wu and Yu, 2016; Chen et al., 2025a), in which a period dominated by a specific source is used to define a baseline relationship, which is then extrapolated to infer an unmeasured component. The detailed workflow, screening criteria, and methodological limitations of the counterfactual formulation are provided in Section A1.2 and Fig A2.

2.3.2 Direct additive formulation

The second candidate formulation was a direct additive model trained using the full-year dataset. In this formulation, fossil fuel BC was represented as the sum of a traffic-related component and a heating-related component:

$$eBC\text{-}FF = eBC\text{-}FF\text{-}traffic + eBC\text{-}FF\text{-}heating$$

Since solid fuel BC (eBC-SF) originates entirely from residential heating (Lin et al., 2018; Lin et al., 2019), its temporal variations are expected to co-vary with those of fossil-fuel heating BC (eBC-FF-heating). We therefore approximate the heating-related fossil fuel contribution as proportional to eBC-SF, with the scaling coefficient “k”:

$$eBC\text{-}FF\text{-}heating = k \cdot eBC\text{-}SF$$

As for the traffic emission component, since it has a non-linear relationship with variables such as pollutant concentration, meteorological conditions, and traffic volume, we employ a non-linear neural network model, denoted as f(X), for its estimation. This model utilizes a feature matrix comprising meteorological conditions, temporal variables, traffic volume, and concentrations of related pollutants for prediction. Thus, the black carbon from fossil fuel combustion is represented by a hybrid model:

$$eBC\text{-}FF = f(X) + k \cdot eBC\text{-}SF$$

The design of this model avoids subjective sample selection and simultaneously enhances its physical interpretability. Both sides of the equation represent the total black carbon concentration from fossil fuel combustion, eliminating the need for any screening based on time periods (e.g., heating season) or meteorological conditions.

The nonlinear function f(X) was implemented using a feed-forward neural network and trained jointly with the linear coefficient “k”. The model was optimized in log-transformed space by minimizing the mean squared error between log(1+eBC-FF) and log(1+eBC-FF-pred). To ensure a positive nonlinear contribution, a Softplus transformation, was applied to the neural-network output, constraining the traffic-related component to non-negative concentrations consistent with its physical interpretation:

$$Softplus(x) = ln(1+e^{x})$$

The predictor variables were standardized using the train set statistics, while eBC-SF was kept in its original concentration scale. A stratified random sampling approach was applied, where 80% of the data from each month were used for training and the remaining 20% for testing, ensuring comparable seasonal representation. The detailed neural-network architecture and training hyperparameters are summarized in

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table A1, and the predictor variables are listed in Table A2.

2.3.3 Orthogonalized additive formulation

The same neural-network architecture, loss function, optimization procedure, and training hyperparameters were used in the orthogonalized additive formulation. The only modification was that the original predictor matrix X was replaced by the orthogonalized predictor matrix $X_{\perp}$.

In the orthogonalized additive formulation, the predictor matrix used by the neural network was first decomposed with respect to eBC-SF. For each predictor, the component linearly associated with eBC-SF was removed, and only the residual component orthogonal to eBC-SF was retained. The resulting feature matrix was denoted as $X_{\perp}$. This procedure reduced the linear collinearity between the neural-network predictors and the heating-related linear term k·eBC-SF, thereby limiting competition between the nonlinear traffic-related component and the linear heating-related component. The final orthogonalized additive formulation was therefore expressed as:

$$eBC\text{-}FF = f(X_{\perp}) + k \cdot eBC\text{-}SF$$

To interpret the fitted model, SHAP (SHapley Additive exPlanations) values were calculated for both the neural-network component, $f(X_{\perp})$ (Lundberg and Lee, 2017), and the full additive model, $f(X_{\perp}) + k \cdot eBC\text{-}SF$, to distinguish the drivers of the traffic- and heating-related terms. The mean absolute SHAP value was used to rank the overall importance of each variable.

The overall process of this study is shown in Fig 2:

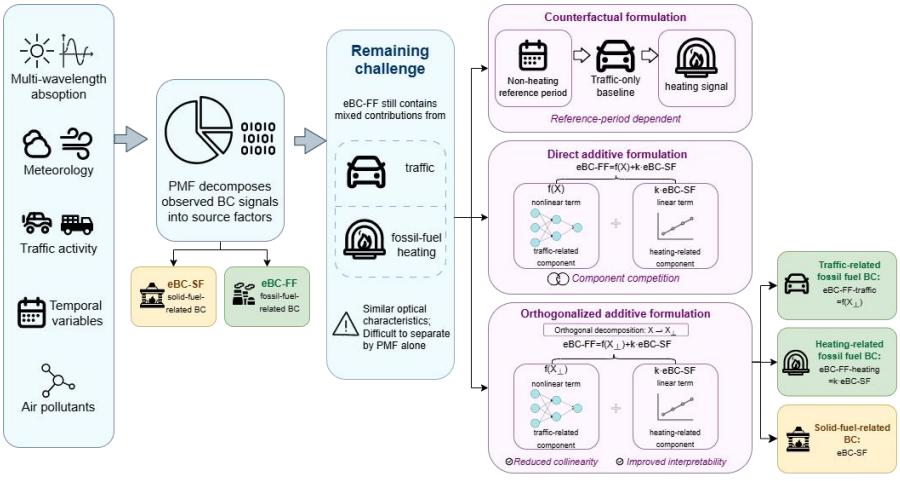


Fig 2. Conceptual framework of this study. Positive matrix factorization (PMF), a source apportionment method based on light absorption properties, applied to multi-wavelength absorption data separates total equivalent black carbon (eBC) into fossil fuel (eBC-FF) and solid fuel (eBC-SF) components. Three machine learning models were applied to further resolve eBC-FF into traffic- and heating-related contributions.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

3. Results and discussion

3.1 PMF-resolved eBC sources and their temporal patterns

The PMF analysis consistently resolved two distinct eBC factors throughout the study period, which were interpreted as fossil-fuel-related eBC (eBC-FF) and solid-fuel-related eBC (eBC-SF). Figure 3(a) shows the average absorption profiles of the two factors derived from all available years. This agrees well with previous studies showing that the spectral dependence of fossil-fuel-derived black carbon is relatively stable, with reported AAE values typically close to unity (Zotter et al., 2017). Nevertheless, all retrieved AAE values fall well within the reported range for biomass-burning emissions, indicating that the PMF-derived factor profiles are physically reasonable.

The temporal behaviour of the two factors further supports their source interpretation. As shown in Fig 3(b), eBC-FF exhibited a characteristic bimodal diurnal cycle, with a smaller morning peak during the traffic rush hour (around 09:00-10:00) and a much more pronounced evening peak. While the evening enhancement is partly associated with increased traffic activity, it is primarily attributed to residential oil heating, one of the dominant domestic heating fuels in Dublin during winter (Lin et al., 2019). This interpretation is further supported by the diurnal patterns in summer, where the morning traffic peak remains evident and comparable, whereas the evening peak is substantially reduced during summer when residential heating emissions are largely absent. In contrast, eBC-SF displayed a pronounced single evening peak that coincided with the residential heating period, consistent with the well-recognized wintertime influence of solid fuel combustion in Dublin (Lin et al., 2018; Lin et al., 2019; Lei et al., 2026).

Overall, the PMF analysis successfully separated the two dominant categories of black carbon sources, namely fossil-fuel combustion and solid-fuel combustion, without relying on predefined spectral assumptions required by the conventional Aethalometer model. This demonstrates the potential of receptor-model-based source apportionment to provide a more observation-driven characterization of black carbon sources. However, the analysis also reveals an inherent limitation of PMF under the present conditions. Because traffic emissions and residential oil heating exhibit highly similar absorption spectra, they cannot be further separated within the PMF framework. Distinguishing these two source sectors is nevertheless important for designing effective air pollution mitigation strategies. Therefore, a machine-learning approach was subsequently developed to further separate the fossil fuel factor into traffic- and oil-heating-related contributions. These PMF-resolved factors provide the basis for the following analysis: eBC-SF was treated as a heating-activity proxy, while eBC-FF was selected as the target for further decomposition into traffic-related and heating-related fossil-fuel BC components.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

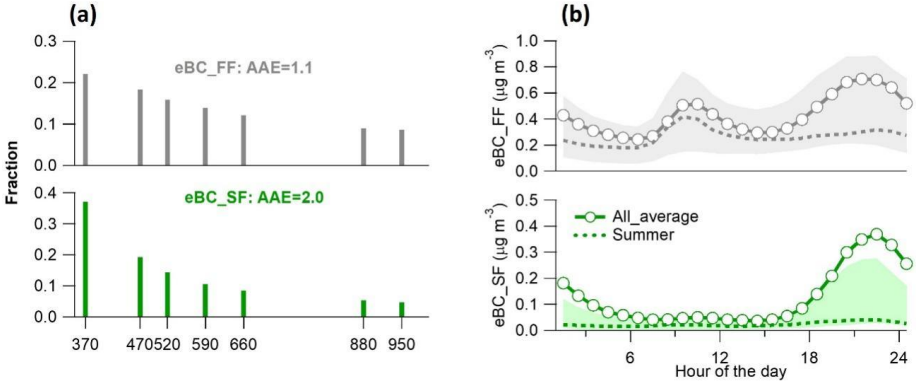


Fig 3. (a) Average absorption profiles of the two equivalent black carbon (eBC) factors identified by the positive matrix factorization analysis, fossil fuel (eBC-FF) and solid fuel (eBC-SF) components, based on the entire dataset (2016-2018 and 2021-2023). The average AAE value for each factor is also shown. (b) Average diurnal variations of the two eBC factors. The shaded areas represent the 25th and 75th percentiles, while the dashed line shows the mean diurnal cycle of eBC-FF during the summer period, when residential heating emissions are minimal.

3.2 Pollution conditions and temporal patterns

Based on the multi-year PMF analysis, the seasonal variations of eBC composition and sources was explored. Long-term observations at the Dublin urban background site showed clear seasonal variations in eBC components (Fig 4). eBC-SF was mainly elevated during the cold season, with a median concentration of 0.032 μg/m^3 (IQR: 0.006–0.106 μg/m^3), and remained low in summer (0.013 μg/m^3, IQR: 0.003–0.027 μg/m^3). Although eBC-FF was present throughout the year, it also increased during the cold season, with a median concentration of 0.250 μg/m^3 (IQR: 0.130–0.495 μg/m^3), compared with 0.193 μg/m^3 (IQR: 0.114–0.348 μg/m^3) in summer. This suggested that fossil fuel BC did not only reflect year-round traffic emissions but also included heating-related contributions. The monthly mean results further showed an overall inverse relationship between BC concentrations and temperature, indicating the combined influence of low-temperature heating activity and unfavorable dispersion conditions on BC levels.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

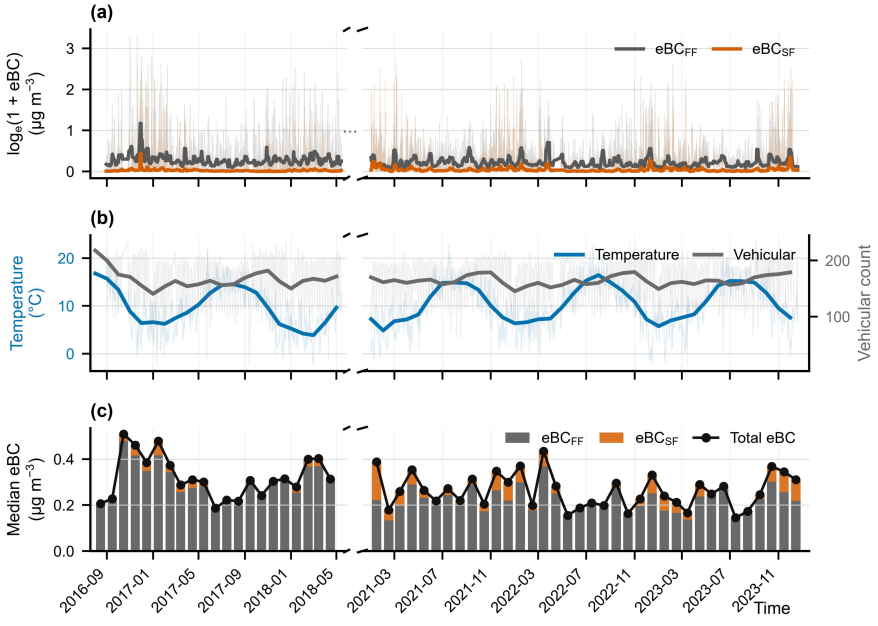


Fig 4. Temporal variations of BC components and related environmental variables. (a) Time series of two components from the positive matrix factorization analysis, fossil fuel (eBC-FF) and solid fuel (eBC-SF); (b) temporal variations of temperature and traffic counts. The solid line represents the hourly median connection, while the light-colored shading indicates the actual values. (c) monthly mean concentrations of BC components.

The seasonal diurnal patterns further reveal the overlap of different source-related signals. Traffic counts showed a clear weekday morning peak at around 09:00, while this peak was much weaker on weekends. eBC-FF mainly followed traffic-related variations in spring and summer, with elevated concentrations during daytime. In autumn and winter, however, eBC-FF also showed a pronounced evening-to-nighttime enhancement, coinciding with the nighttime peaks of eBC-SF. In contrast, eBC-SF remained low in summer but increased substantially during cold-season nights, with the winter nighttime peak being approximately 12.3 times higher than that in summer. These results indicate that BC at the urban background site was influenced by both traffic and heating emissions, making it difficult to further separate traffic- and heating-related fossil-fuel BC contributions from the raw temporal patterns alone.

3.3 Model selection and robustness

Building on the PMF-resolved factors, three candidate machine-learning formulations were diagnostically compared to further resolve eBC-FF into traffic and heating components (Table 1).

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Metrics for the counterfactual model were calculated for the non-heating reference test set, whereas those for the additive models were calculated for the full-year stratified test set.

Model	Data used	Formulation	R^2	RMSE	Validation evidence
Counterfactual	Non-heating reference period	Traffic-only prediction	0.63	0.13	Full-year residual diagnostics
Direct additive	Full-year data	eBC-FF=f(X)+k·eBC-SF	0.75	0.15	LOYO (k): 0.596±0.098 – 1.792±0.070
Orthogonalized additive	Full-year data	eBC-FF=f($X_\perp$)+k·eBC-SF	0.79±0.004	0.14	LOYO (k): 0.685±0.038– 0.956±0.041

The counterfactual formulation used selected non-heating reference samples to train a traffic-only prediction model and then applied this model to the full dataset. Although this approach provided a direct way to construct a traffic-only baseline, its full-period application depended on the definition of reference samples. The resulting residual could therefore be affected not only by heating emissions, but also by seasonal changes in meteorology and model extrapolation, etc (Fig A2 and Section A1.2). This formulation was therefore used as a diagnostic baseline rather than the final model.

The direct additive formulation avoided reference-period screening by using full-year data and explicitly expressing eBC-FF as f(X)+k·eBC-SF. However, without orthogonalization, the nonlinear predictor f(X) and the linear heating term k·eBC-SF could compete in explaining similar variations in eBC-FF. This was reflected by the large variability of the learned coefficient "k" in cross-validation, indicating unstable component allocation between traffic and heating.

Before adopting the nonlinear formulation, we also examined whether a linear traffic term would suffice (Fig A4). Holdout R^2 in log(1+eBC-FF) space rose from 0.672 ± 0.012 for a three-predictor baseline (NO_x, HOA, eBC-SF) to 0.718 ± 0.009 for a full-feature linear additive model with log-scale loss, but remained below the 0.775 ± 0.008 of the additive neural network. A gradient-boosted model predicting total eBC-FF without decomposition reached 0.836 ± 0.005; it is not a competing formulation here, because it returns no sectoral split. The fitted coefficients differed accordingly (0.839, 0.995 and 0.743; Fig A4(b)), showing that the allocation is specification-dependent even as prediction improves with model flexibility.

The orthogonalized additive formulation was therefore adopted as the final model. By using feature components orthogonal to eBC-SF, this formulation reduced the competition between the neural-network term and the heating-related linear term,

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

while producing a more stable coefficient "k". Detailed diagnostic comparisons of prediction accuracy and monthly median residuals among the three candidate formulations are provided in Fig A3, while the LOYO comparison before and after orthogonalization is shown in Fig A5. The performance and robustness of this selected additive formulation will be elaborated upon in the following sections.

Fig 5 assesses the performance of the model from three different perspectives: overall prediction accuracy, stability, and the ability to capture diurnal patterns. First, the scatter plot of observed versus predicted values in Fig 5(a) shows good agreement, with a regression slope close to unity and R^2 of 0.79±0.004, indicating that the model can reasonably reproduce the overall variation in eBC-FF concentrations. Fig 5(b) shows the distribution of the proportional coefficient "k" from 300 repeated trainings under the same model structure and fixed data split. Because the diagnostics below use five seeds, each with its own month-stratified 80/20 split, their fitted k (0.743 ± 0.039) differs slightly from the 300-run mean (0.735) and from the jointly optimized value used for the final apportionment (0.730); these differences are below 2% and are small relative to the specification sensitivity examined later. In these repeated runs, randomness was introduced only through neural-network initialization and mini-batch shuffling during training, with a mean of 0.735. This indicates that the learned coefficient "k" was not strongly affected by random initialization or mini-batch shuffling under the fixed data split. All R^2 values in this section are holdout values in log(1+eBC-FF) space. Table 1, Fig 5(a) and the fitted coefficients above derive from 300 repeated trainings on the full stratified test set, whereas the linear-baseline and ablation comparisons below use five runs evaluated on the early-stopping subset; the latter are therefore about 0.015 lower (0.775 vs 0.79) and should not be compared at the third decimal.

Second, Fig 5(c) presents the seasonal mean diurnal cycles based on hourly median values of observed and predicted eBC-FF. The close agreement between the two curves further demonstrates that the model captures the major temporal features of eBC-FF, including higher concentrations in winter than in summer, low concentrations in the early morning, and a pronounced morning peak at around 09:00. These features are consistent with the understanding that urban fossil-fuel-related black carbon is influenced by both traffic activity and seasonally varying dispersion conditions. Previous studies have also reported that urban BC or eBC-FF commonly exhibits diurnal peaks during traffic rush hours and higher concentrations during cold or winter periods.(Zotter et al., 2017) Therefore, Fig 5 suggests that the model not only provides satisfactory numerical predictive performance, but also preserves temporal patterns consistent with real-world emission processes, providing a basis for the subsequent separation of traffic- and heating-related fossil fuel BC contributions.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

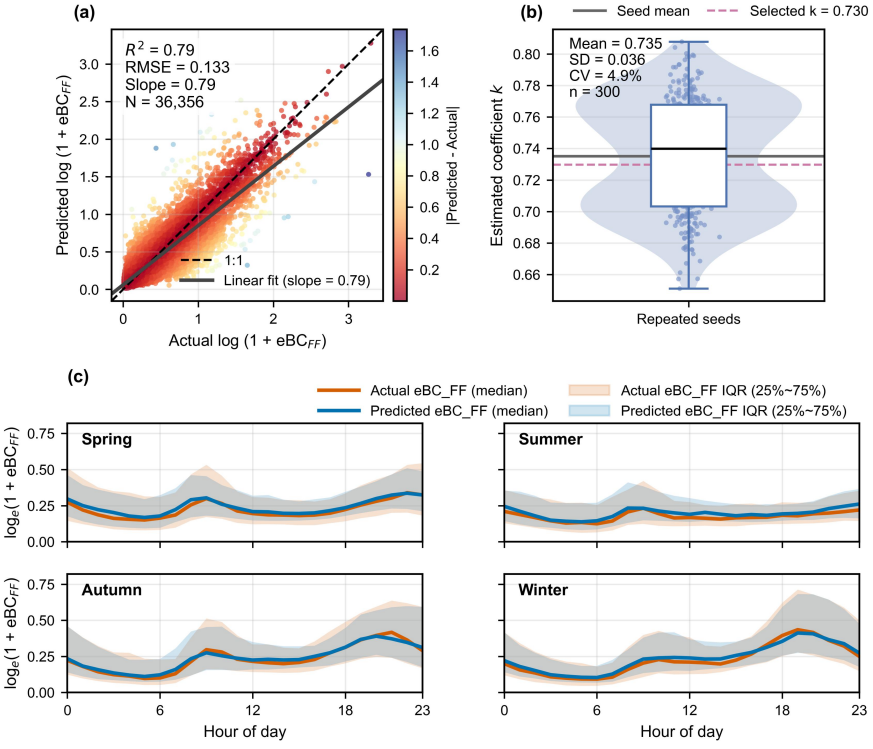


Fig 5. Performance and stability of the selected orthogonalized additive model. (a) Predicted vs. observed log(1 + eBC-FF); point colour denotes the absolute prediction error, |predicted - observed|. (b) Distribution of the fitted coefficient "k" from 300 repeated random seeds; points are individual runs, the violin shows the kernel density, the box indicates the interquartile range and median, whiskers extend to 1.5 × IQR, and horizontal lines indicate the seed mean and selected final "k". (c) Seasonal diurnal cycles of observed and predicted eBC-FF; lines show hourly medians and shaded bands show the 25th-75th percentile range (IQR).

To evaluate the temporal transferability of the selected model, a leave-one-year-out (LOYO) cross-validation was performed. In this approach, each complete year was used once as the test set, while the remaining years were used for model training. As shown in Fig A6, the model maintained reasonable predictive performance across years, with mean train and test R^2 values of 0.778 and 0.696, respectively. To assess the robustness of the learned proportional coefficient, each LOYO fold was further repeated with 10 random seeds. The fold-specific mean coefficient "k" ranged from 0.685 ± 0.038 to 0.956 ± 0.041 for the 2017, 2021, 2022, and 2023 test folds (the four complete calendar years available across the two sampling periods, 2016 - 2018 and 2021 - 2023), respectively. These results indicate that the model had acceptable cross-year predictive capability, while the interannual variability of "k" was larger than the uncertainty caused by random initialization and

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

mini-batch shuffling within each fold. Nevertheless, this cross-year variability of "k" still represented a marked improvement over the non-orthogonalized formulation, whose coefficient stability was substantially poorer (Fig A5).

To test how far the allocation depends on individual predictors, two ablations were run on the same five splits (Fig A7). Removing vehicle counts changed neither holdout R^2 (0.775 ± 0.008 to 0.778 ± 0.006) nor the heating fraction (k: 0.743 to 0.716, heating shift < 1 percentage point). Removing NO_x and HOA together reduced R^2 to 0.648 ± 0.007, raised "k" to 0.863 ± 0.021, and shifted the heating fraction upward by about 3.4 percentage points, with a corresponding decrease in the traffic fraction. Normalized traffic diurnal profiles nevertheless remained similar in shape. The allocation therefore rests mainly on combustion-related predictors rather than on traffic counts, and agreement in diurnal shape alone cannot demonstrate attribution accuracy.

Finally, the estimator was tested against synthetic data in which the heating contribution was known (Fig A8). Using the observed predictors and SF series, eBC-FF was regenerated as $f(X_{\perp})+k_{true}\cdot$eBC-SF multiplied by a lognormal factor, i.e. Gaussian noise added on the log scale, with its standard deviation chosen to give a log-space signal-to-noise ratio comparable to the observations. At $k_{true}=0.73$ the recovered coefficient was 0.732 ± 0.012, but at $k_{true}=0$ a non-zero offset of 0.083 ± 0.002 remained; recovery declined to 1.436 ± 0.033 at $k_{true}=1.50$. Three violations of the assumed structure were then imposed: heating driven by temperature rather than proportional to SF yielded 0.082 ± 0.005; SF leakage into eBC-FF ($\lambda=0.15$) yielded 0.873 ± 0.036 against an expected effective value of 0.88; and traffic–SF coupling ($\beta=0.35$) yielded 0.827 ± 0.004. The decomposition therefore tracks changes in the prescribed coefficient. Under mild leakage, however, the recovered value (0.873) matched the effective coefficient implied by that leakage (0.88) rather than the prescribed heating coefficient, indicating that a leaky channel is not distinguishable from a stronger "k". A small positive offset also remained at zero truth. These tests describe how the estimator behaves under a known template; they do not establish accuracy against the real atmosphere.

Considered together, the predictor-removal configurations moved the fitted coefficient between 0.716 and 0.863 (Fig A7), corresponding to 54–59% for traffic and 17–21% for oil heating. Solid-fuel BC is unaffected in absolute terms; only its normalized share shifts, from 23.5% to about 24.5%, as the other two components change. The ordering of the three sectors and the combined heating share of 41–46% are therefore preserved, although the margin between traffic and heating narrows from about 18 to 9 percentage points. The wider spread of "k" across leave-one-year-out folds (0.685–0.956; Fig A6) presumably reflects year-to-year differences in fuel use, but changes in meteorology, data coverage and model transferability across folds cannot be excluded, so it is not read here as a measure of specification sensitivity. The ± values reported in Section 3.4 describe training randomness under a fixed specification and should not be read as the total uncertainty of the allocation.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

3.4 Source apportionment results

For the selected orthogonalized additive model, the neural-network parameters and the proportional coefficient "k" were jointly optimized during model training by minimizing the difference between observed and predicted eBC-FF. The optimized coefficient was k = 0.730 ± 0.036, and the heating-related fossil fuel BC component was therefore estimated as:

$$eBC\text{-}FF\text{-}heating = (0.730 \pm 0.036) \times eBC\text{-}SF.$$

The traffic-related fossil-fuel BC component was obtained from the neural-network output $f(X_{\perp})$. Based on this calculation, the concentration-based contributions of traffic-related fossil fuel BC, heating-related fossil fuel BC, and solid fuel BC were approximately 59.2% ± 1.4%, 17.2% ± 1.4%, and 23.5%, respectively. The ± values denote the standard deviation across 300 repeated trainings with a fixed data split, i.e. training randomness only; the specification sensitivity is quantified in Section 3.3.

Based on this relationship, the eBC-FF was apportioned into traffic and heating contributions, as shown in Fig 6. Fig 6(a) presents the seasonal diurnal cycles of actual eBC-FF, predicted eBC-FF, traffic-related (eBC-FF-traffic), and heating-related eBC-FF (eBC-FF-heating). Traffic emissions exhibited a clear morning peak and generally remained at a certain concentration level during the daytime, including the period between the morning and evening peaks. This pattern is consistent with the diurnal variation of traffic activity. In contrast, heating-related eBC-FF was mainly observed during nighttime in the heating season, and it corresponded well to the gap between the traffic-related component and the total eBC-FF concentration during these periods.

Fig 6(b) shows the monthly source contributions and mean temperature. The purple bars represent traffic emissions, while the blue and green bars represent fossil-fuel heating and solid fuel, respectively. The total heating contribution (i.e. fossil-fuel heating + solid fuel) was significantly and inversely related to monthly mean temperature (Pearson $r = -0.68$, $p < 0.001$; $n = 58$ months with concurrent observations), rising from ~12% in summer to ~33% in winter — equivalent to an increase of roughly 2.1 percentage points per °C decrease. Overall, the apportioned results indicate that the model can reasonably distinguish eBC-FF into traffic- and heating-related contributions, producing diurnal and seasonal variations that are consistent with expected real-world emission patterns.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

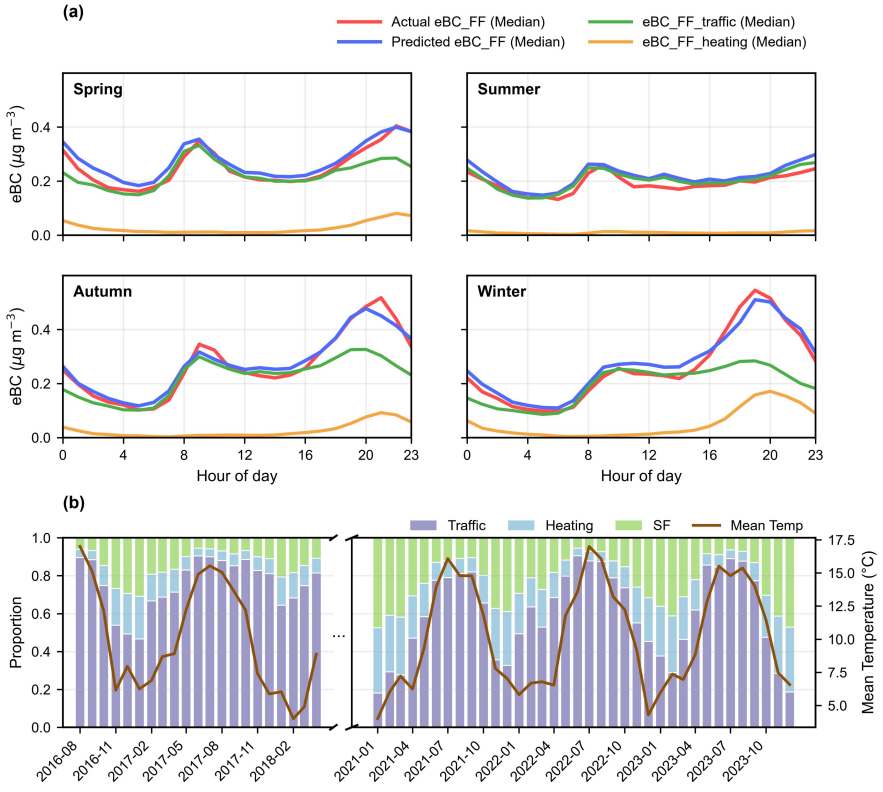


Fig 6. (a) Seasonal diurnal cycles of actual, predicted and apportioned eBC-FF (curves of hourly medians); (b) monthly source contributions of eBC-FF-traffic, eBC-FF-heating and eBC-SF. The curve line shows the pattern of monthly mean temperature.

In this study, SHAP analysis was applied at two levels: the neural-network component $f(X_{\perp})$ and the full mixed model. For $f(X_{\perp})$, which excludes the linear heating term, NO_x, HOA, and PBLH were the most important variables, followed by wind direction, day of year, traffic volume (vehicular) and temperature, etc. This indicates that $f(X_{\perp})$ was mainly controlled by traffic-related combustion markers and atmospheric dispersion conditions, supporting its interpretation as the traffic-related fossil fuel BC component.

When the linear term $k \cdot$eBC-SF was included in the full model, eBC-SF appeared as an additional important variable, reflecting the contribution of the heating-related linear component to the prediction of eBC-FF. Importantly, the SHAP ranking of the remaining orthogonalized predictors was generally consistent with that obtained for $f(X_{\perp})$ alone, and their order does not change with the addition of the linear terms. This consistency indicates that adding the linear heating term did not substantially change the role of the nonlinear predictor $f(X_{\perp})$. Therefore, the two-level SHAP analysis supports the model design in which the heating-related contribution is represented by

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

k·eBC-SF, while the traffic-related variation is mainly captured by $f(X_{\perp})$.

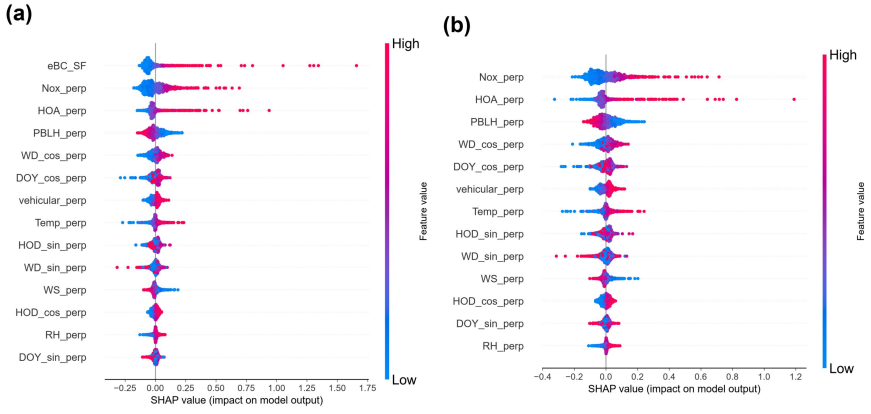


Fig 7. SHAP beeswarm plots for the orthogonalized additive neural network model in log(1 + eBC-FF) space. (a) shows the feature contributions to the full model prediction $f(X_{\perp})$ + k·eBC-SF. (b) isolates the nonlinear neural network component $f(X_{\perp})$, revealing the learned nonlinear dependencies among the 13 perpendicular features alone. In both panels, features are ranked by mean |SHAP| (descending); each point represents one observation, colored by feature value (red = high, blue = low), with horizontal position indicating the SHAP value—positive values shift the prediction above the baseline, negative below it.

Beyond evaluating predictive performance, we examined whether the apportioned components exhibited temporal patterns consistent with their assigned source activities. As shown in Fig A9(a), the seasonally scaled traffic-related eBC-FF component (eBC-FF-traffic) broadly reproduced the characteristic morning peak, evening peak and sustained daytime pattern of vehicle counts across all seasons. The heating-related eBC-FF component (eBC-FF-heating) remained close to zero during summer and showed pronounced evening-to-nighttime enhancements in autumn and winter, consistent with the seasonal timing of residential heating (Fig A9(b)). Weekday–weekend contrasts provide a further activity check (Fig A10). Excluding June–August, the traffic component showed a stronger weekday morning enhancement with a lower weekend daily mean (0.310 vs 0.339 μg/m³, weekends vs weekdays), while the heating component was higher at weekends in both daily mean and daily peak, consistent with occupants being at home. The traffic contrast follows partly from vehicle counts entering the predictor set. The weekend enhancement of the heating component is consistent with occupancy-driven heating activity, but since that component is defined as k·eBC-SF, this comparison provides an activity-consistency check rather than an independent constraint on the oil-heating amount.

Independent chemical support came from the collocated ACSM measurements, whose peat, wood and coal factors were never used as predictors; only the fossil-fuel-related HOA, in its SF-orthogonalized form, enters the predictor matrix (Fig 8). After removing year–month specific hourly means, peat and wood correlated

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

more strongly with eBC-SF (r = 0.830 and 0.756) than with the traffic component (0.418 and 0.440), with paired differences of 0.411 (95% interval 0.241–0.572) and 0.316 (0.118–0.542). Coal showed a smaller but still positive difference (0.097, 0.027–0.168), and the HOA difference was indistinguishable from zero (0.113, −0.017–0.281). September–May diurnal profiles showed evening enhancements shared by eBC-SF, peat and wood, while the traffic component and NO_x retained a morning enhancement. These comparisons support the solid-fuel interpretation on evidence independent of the absorption data. They do not test the oil-heating term, which is proportional to eBC-SF by construction, and residual correlations of about 0.44 between peat/wood and the traffic component indicate co-variation that the present predictor set does not resolve.

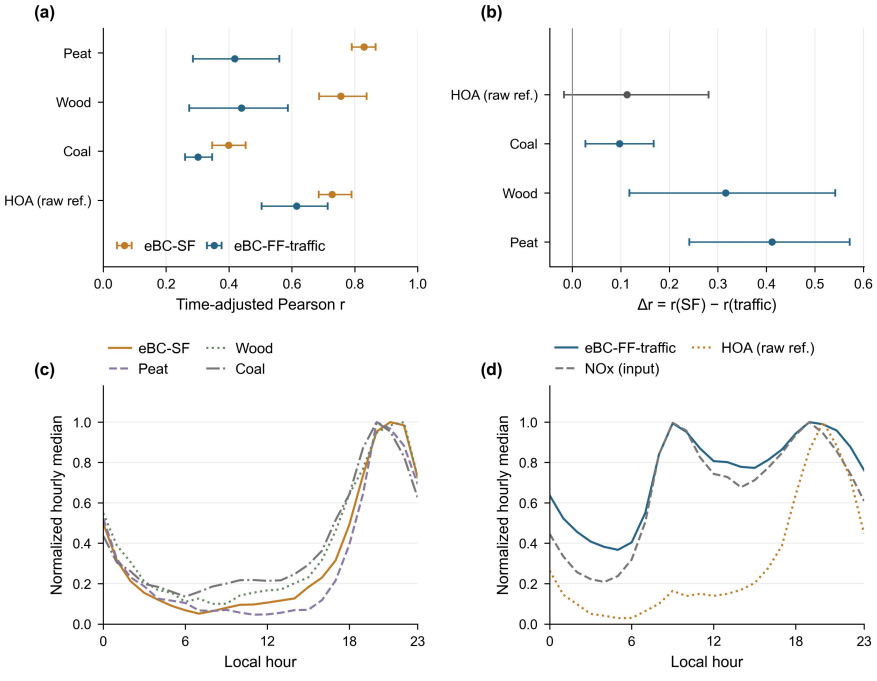


Fig 8. ACSM OA-factor comparisons with BC components. (a) Pearson correlations after removal of year-month-specific hourly means. (b) Paired differences Δr = r(SF) − r(traffic). Intervals are percentile 95% intervals from 2,000 paired calendar-week block resamples. Panels (a,b) use all months. (c,d) September–May hourly medians normalized to each variable's maximum. HOA is shown in its raw form; HOA and NO_x were orthogonalized relative to eBC-SF before NN training. The common complete-case sample contains 36,312 hourly records.

Taken together—holdout performance, repeated-seed and leave-one-year-out stability, model-structure comparison, predictor ablation, synthetic recovery, physical-consistency diagnostics and independent ACSM comparison—the decomposition is reproducible within a bounded range and chemically consistent

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

where an independent measurement exists. The oil-heating term, which no independent measurement constrains, remains conditional on the proportionality assumption.

3.5 Discussion

The BC source apportionment model provides an advanced tool for apportioning BC source into policy-informed sectors, traffic and heating, despite their overlapping physicochemical characteristics. Although only about 2% of Dublin households reported solid fuel as their primary heating fuel (Central Statistics Office, 2021), solid-fuel combustion accounted for 23.5% of total eBC — more than the 17.2% from oil heating — and thus for more than half (57.7%) of all heating-related BC. The apportioned BC contributions reveal a disproportionate impact of solid-fuel combustion on urban BC concentrations, originated from the fuel-specific BC emission factors, combustion efficiency, emission timing, and meteorological dispersion, etc. (Denier Van Der Gon et al., 2015; Liu et al., 2023). Solid fuel burning often occurs during cold season evening and nighttime periods, when boundary layers are shallower and dispersion is weaker. As a result, even limited solid fuel use can produce a substantial local BC concentration burden (Herich et al., 2011).

These findings highlight the value of actual source apportionment for identifying air-quality-relevant sources that may be underestimated by energy-use statistics alone. The further separation of fossil-fuel BC into traffic- and heating-related components also provides more sector-specific evidence for urban BC mitigation, particularly for distinguishing traffic emissions from residential heating impacts. Future work should test the transferability of this model at additional sites and incorporate higher-resolution fuel-use, heating-activity, traffic-flow, and dispersion information.

It should be noted that the coefficient “k” in the linear term k·eBC-SF should not be interpreted as a fixed universal emission factor. Rather, it represents an empirical parameter describing the co-variation between fossil fuel heating and solid fuel heating under a given urban residential heating structure. This assumption is reasonable because residential heating activities are generally driven by similar seasonal and diurnal demands, while household fuel-use patterns are not expected to switch abruptly over short timescales. Therefore, eBC-SF can serve as a useful reference for heating activity and for extracting the fossil fuel heating contribution that co-varies with it. Meanwhile, the interannual variation of “k” observed in the LOYO analysis indicates that this relationship may still be affected by changes in fuel-use structure, combustion conditions, and data coverage. The orthogonalized formulation reduced the competition between the nonlinear traffic-related term and the linear heating-related term, making its value more stable and interpretable (“k” varied between 0.716 and 0.863 among the predictor-removal configurations; Section 3.3). Nevertheless, the geographical transferability of this coefficient and model should be further evaluated in other urban environments or with independent heating-activity data.

Direct and independent measurements of traffic-related and heating-related

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

eBC-FF were not available. Moreover, vehicle counts were included among the model predictors, while the heating-related component was defined through the proportional term k·eBC-SF. Therefore, the comparisons in Fig A9 should be interpreted as physical-consistency diagnostics rather than independent source-tracer validation. The resulting components represent a conditional decomposition under the proportionality and orthogonalization assumptions, although their expected temporal behavior and robustness across complementary evaluations support the physical plausibility of the source interpretation. The zero-truth offset of 0.083 found in Section 3.3 would correspond to roughly 2 percentage points of total eBC if it carried over unchanged to the observed coefficient. The synthetic tests were designed to characterise estimator behaviour under a known template, not to bound the real-world deviation, so this figure is an illustration rather than an upper limit.

4. Conclusions and prospects

Building on PMF-resolved source categories, this study achieved a more detailed apportionment of urban black carbon, resolving fossil fuel BC (eBC-FF) into separate traffic and residential oil-heating contributions, a distinction that absorption-based conventional models cannot resolve. The heating component was represented by a linear term k·eBC-SF, and the traffic component by a neural network driven by orthogonalized predictors.

The selected orthogonalized additive model reproduced the observed variability and seasonal diurnal patterns of eBC-FF, with the coefficient “k” learned consistently across training runs (0.730 ± 0.036) and its variation across the predictor sets tested bounded at 18% (Section 3.3). This coefficient yields a heating-related fossil-fuel contribution that was consistently smaller than the solid-fuel contribution. Based on this model, the concentration-based contributions of traffic-related fossil fuel BC, heating-related fossil fuel BC and solid fuel BC were estimated at 59.2% ± 1.4%, 17.2% ± 1.4% and 23.5% of total eBC, respectively. The apportioned components showed physically consistent temporal patterns, with traffic-related eBC-FF persisting across seasons and heating-related eBC-FF mainly appearing during cold season nighttime periods.

Overall, this study resolved urban BC into three policy-relevant sources — road traffic, solid-fuel heating, and oil heating — revealing a much more granular picture of urban combustion emissions than other source apportionment approaches provide. While the model was evaluated at a single urban background site, it provides a basis for future applications at additional sites where long-term BC observations and suitable traffic, meteorological, and pollutant predictors are available. Further work incorporating higher-resolution activity and dispersion information would help improve the transferability and policy relevance of this source apportionment approach. Because road traffic and residential heating are expected to evolve along markedly different trajectories under future decarbonization and electrification scenarios, resolving their separate contributions to urban BC is essential for

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

improving source-specific emission inventories, climate-model inputs, and radiative-forcing estimates.

Appendix A: Supplementary data and methods

A1 Supplementary data and methods

A1.1 Reference eBC absorption profile figure

In addition, a fossil-fuel-dominated reference eBC profile was derived from AE33 measurements at a kerbside site in Dublin, where eBC is predominantly influenced by traffic emissions. The reference eBC-FF absorption profile was determined as the average profile during periods with minimal influence from residential heating, i.e., during the non-heating period. This traffic-dominated reference profile was then used to constrain the fossil-fuel-related factor in the ME-2 analysis. The AE33 was operated using the standard 10-m filter tape (type M8050), with inlet and impactor cleaning every 1 - 3 months following the manufacturer's recommendations (Magee Scientific AE33 User Manual).

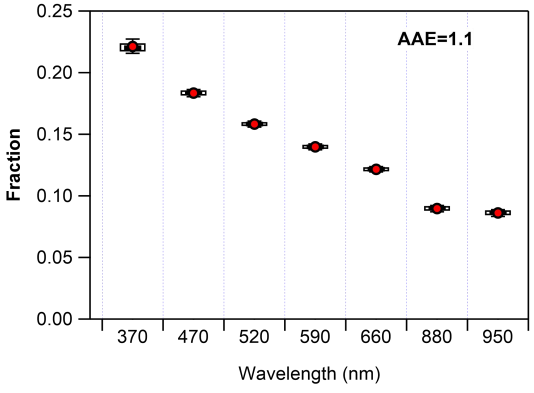


Fig A1. Box plots of the reference eBC absorption profile derived from measurements at a kerbside site in urban Dublin during non-heating periods, defined as summer periods with low POA concentrations (< 0.5 μg/m³). The mean (circle), median (horizontal line), 25th-75th percentiles (box), and 10th-90th percentiles (whiskers) are shown.

A1.2 Overview of the Counterfactual Method

The initial model framework followed a counterfactual method. Samples dominated by traffic emissions were selected from the non-heating season to train an XGBoost model, which was then used to predict eBC-FF under a hypothetical traffic-only scenario. We adopted the screening criteria from previous literature as a

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

reference, but to ensure an adequate sample size, some subjective adjustments were also made: data from April, May, June and September with wind speed below 5 m/s. The difference between actual and predicted eBC-FF was interpreted as the heating-related fossil fuel contribution. However, this approach relies on the assumption that the traffic-emission relationship derived from the non-heating season is transferable to the whole year, and it is sensitive to seasonal meteorological differences and subjective data filtering. The accuracy of the results was also poor in this study, and further optimization did not substantially improve the model performance. Therefore, this counterfactual model was not used as the final method, but is presented in the Supporting Information for methodological comparison.

This counterfactual logic has been widely applied in source apportionment. For example, Wu and Yu used element carbon (EC) as a tracer for co-emitted primary organic carbon (OC), with the residual attributed to secondary formation (Wu and Yu, 2016). More recently, similar ideas have been adopted in ML-based studies, where source-dominated training periods were extrapolated to separate mixed aerosol components, such as local versus transboundary oxygenated organic aerosol (OOA) and primary versus secondary marine OA (Lei et al., 2025; Chen et al., 2025a).

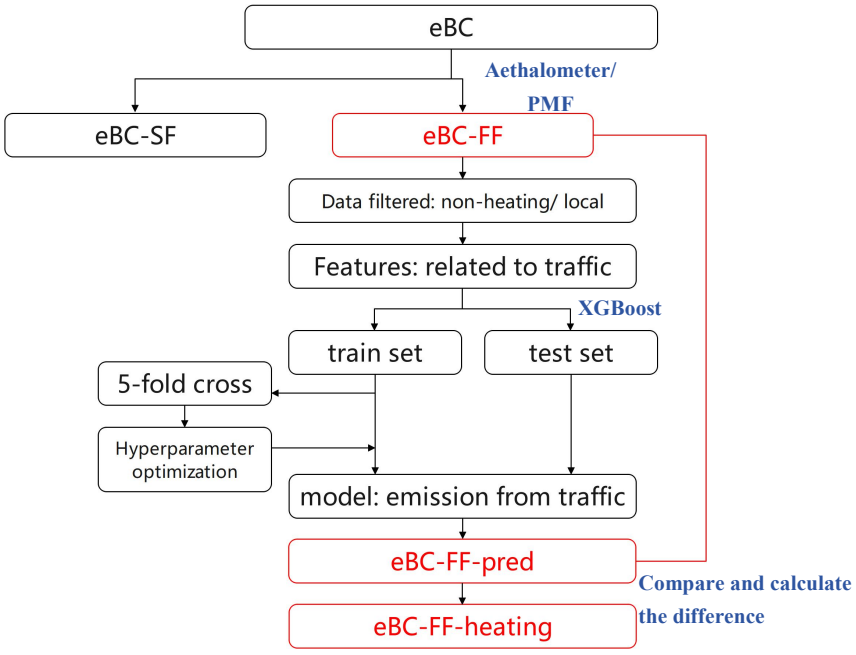


Fig A2. Flow chart of the counterfactual model

A1.3 Hyperparameters and Variable List

The parameters of the neural network-linear mixed model are as shown in the Table A1 below (the parameters remain unchanged before and after orthogonalization):

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table A1. Neural-network architecture and training hyperparameters used in the additive models.

Parameter	Setting
Model structure	Feed-forward neural network + linear additive term
Nonlinear component	f(X) / f($X_{\perp}$)
Linear component	k·eBC-SF
Target	log(1 + eBC-FF)
Loss function	Mean squared error in log space
Input features	13 orthogonalized predictors plus the linear heating term
Data split	Monthly stratified random split
Training/testing ratio	80% / 20%
Feature scaling	Standardization based on training-set mean and standard deviation
Hidden layers	2
Neurons per hidden layer	16
Activation function	ReLU
Output transformation	Softplus
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-5}
Batch size	512
Maximum epochs	200
Early stopping patience	25 epochs
Dropout	0.0
Range of coefficient "k"	0–10

The nonlinear part of the model captures traffic emissions. Therefore, the selected features include pollutants from traffic emissions, traffic flow (vehicular),

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

meteorological conditions, and time-related variables. For variables that are periodic or represented by angles, we include pairs of trigonometric functions as features in the training, which preserves the circular nature of angular variables (e.g., wind direction where 0° ≡ 360°). The final set of features, X, is defined as follows:

Table A2. Features matrix

Variable	Type	Role
NOx	emission-related	traffic tracer
HOA	emission-related	traffic/primary organic aerosol tracer
vehicular	emission-related	traffic activity
PBLH	meteorological	atmospheric dispersion
RH	meteorological	humidity condition
Temp	meteorological	temperature
WS	meteorological	wind speed
WD_sin, WD_cos	meteorological	wind direction
DOY_sin, DOY_cos	temporal	seasonal variation
HOD_sin, HOD_cos	temporal	diurnal variation
eBC-SF	heating-related	linear term for heating-related fossil-fuel component

A1.4 Polynomial functions with different degrees

Models were trained and evaluated in log1p(eBC-FF) space using the same direct additive NN settings and month-stratified 80/20 split. Delta test R^2 is calculated relative to the first-order formulation. The first-order formulation gave the highest test R^2; adding higher-order polynomial terms did not improve predictive performance.

Table A3. Comparison of polynomial formulations for the heating-related fossil-fuel BC term.

Polynomial degree	Heating-term formulation	Train R^2	Test R^2	Delta test R^2 vs linear

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

1	k • eBC-SF	0.757	0.733	0.000
2	$k_1 \cdot eBC\text{-}SF + k_2 \cdot eBC\text{-}SF^2$	0.754	0.727	-0.006
3	$k_1 \cdot eBC\text{-}SF + k_2 \cdot eBC\text{-}SF^2 + k_3 \cdot eBC\text{-}SF^3$	0.738	0.703	-0.030
4	$k_1 \cdot eBC\text{-}SF + k_2 \cdot eBC\text{-}SF^2 + k_3 \cdot eBC\text{-}SF^3 + k_4 \cdot eBC\text{-}SF^4$	0.673	0.617	-0.116

A1.5 Technical Details

During the experiments, we recognized that heating emission, represented by the linear part, is also affected by the same meteorological conditions. When weather conditions change, both parts of the model change simultaneously. This brings difficulties and uncertainty to the predictions of the model. Therefore, to reduce the interaction between the two parts, we adopted a method for removing collinearity, namely orthogonal decomposition.

$$X_{\perp} = X - \frac{X \cdot eBC\text{-}SF}{eBC\text{-}SF \cdot eBC\text{-}SF} eBC\text{-}SF$$

All features were orthogonally decomposed with respect to the linear part, eBC-SF. The parallel components were removed, and only the perpendicular components were retained as the new feature matrix, $X_{\perp}$. This procedure reduced the collinearity between the nonlinear predictor matrix and the linear part, thereby improving the interpretability of the coefficient “k”.

A2 Supplementary results

A2.1 Supplementary of PMF-resolved eBC sources

Because the optimal solution was obtained with a=0 in almost every segmented PMF analysis, the absorption profile of eBC-FF remained essentially identical among years, yielding a constant AAE of 1.1. This agrees well with previous studies showing that the spectral dependence of fossil-fuel-derived black carbon is relatively stable, with reported AAE values typically close to unity (Sandradewi et al., 2008). In contrast, the AAE of eBC-SF exhibited greater interannual variability, ranging from approximately 1.7 to 2.6 (Table A4). Such variability is expected, as numerous laboratory and field studies have demonstrated that the AAE of biomass-burning aerosol is strongly influenced by combustion conditions, fuel composition, and aging processes (Pokhrel et al., 2016).

Table A4. Summary of the AAE values of the two eBC factors identified by PMF over the year.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

Periods	AAE (eBC-FF)	AAE (eBC_BB)
2016-2018	1.1	2.6
2021	1.1	1.7
2022	1.1	2.1
2023	1.1	1.8

A2.2 Effects of models

A2.2.1 Details of comparison of BC source apportionment models

The counterfactual version, trained only on the filtered non-heating season data, also failed to achieve satisfactory prediction performance. In contrast, the neural network-linear mixed model performed much better, increasing the R^2 from 0.63 to 0.75. After applying orthogonal decomposition to the feature matrix, although no significant improvement was observed in the median residual, the R^2 further reached a level of 0.79.

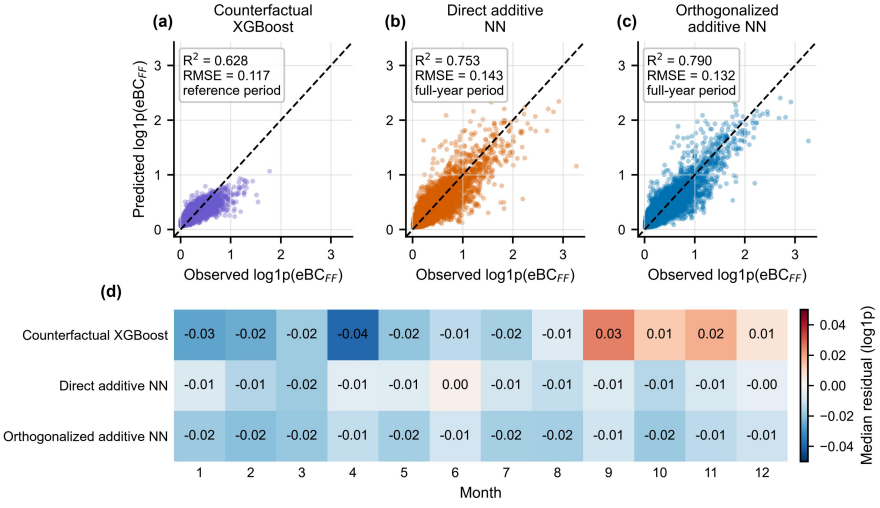


Fig A3. Comparison of the prediction accuracy of the three versions: R^2 of (a) counterfactual model, (b) direct additive NN model, (c) orthogonalized additive NN model; (d) median residual.

A2.2.2 Linear-baseline comparison

To assess whether a simpler linear model could provide predictive performance comparable to the final NN-LM, we evaluated three linear specifications. The first used NO_x and HOA as traffic proxies and eBC-SF as the heating proxy in a multiple

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

linear regression:

$$\mathrm{eBC{-}FF} = b_0 + b_1 \cdot \mathrm{NO_x} + b_2 \cdot \mathrm{HOA} + k \cdot \mathrm{eBC{-}SF}$$

The other two replaced NOx and HOA with the full set of 13 orthogonalized predictors used by NN-LM:

$$\mathrm{eBC{-}FF} = b_0 + \sum_{j=1}^{13} b_j X_{\perp j} + k \cdot \mathrm{eBC{-}SF}$$

The three-variable baseline and the first full-feature model minimized original-scale squared error; the second full-feature model minimized squared error in log(1 + eBC-FF), as did NN-LM. These linear models were compared with the final additive NN-LM and a GBM predicting total eBC-FF without the specified traffic–heating decomposition. The full-feature log-loss linear model matched the NN-LM predictors and loss scale, but did not impose its component non-negativity constraints.

Holdout R^2 increased from 0.672 ± 0.012 for the three-variable baseline to 0.698 ± 0.011 and 0.718 ± 0.009 for full-feature linear models with original-scale and log-scale losses, respectively (Fig A4(a)). The additive NN attained 0.775 ± 0.008, whereas GBM attained 0.836 ± 0.005. Corresponding "k" estimates for the three additive linear specifications were 0.839 ± 0.016, 0.839 ± 0.015 and 0.995 ± 0.009, compared with 0.743 ± 0.039 for the NN (Fig A4(b)).

The tested nonlinear formulation improved prediction, while "k" remained sensitive to model specification. Neither higher total-concentration R^2 nor differences in "k" independently establish attribution accuracy. NN holdout data also served for early stopping, and its across-run variation includes both splitting and training randomness.

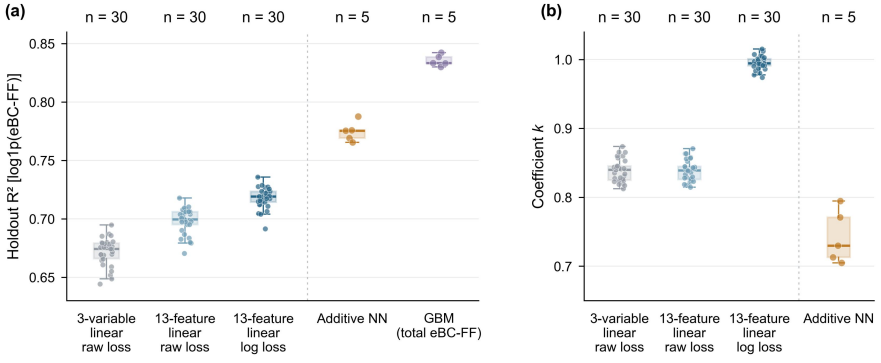


Fig A4. Linear-baseline comparison for eBC-FF. (a) Holdout R^2 in log(1 + eBC-FF) space. (b) Constant coefficient "k" for the additive models. Boxes show the interquartile range and median; whiskers extend to observations within 1.5 interquartile ranges, and points show individual runs. Linear models use 30 runs; additive NN and GBM use five each. The dashed separator distinguishes linear models from reference experiments. Raw loss denotes original-scale squared error; log loss denotes squared error in log(1 + eBC-FF). GBM predicts total eBC-FF without the specified additive decomposition.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

A2.2.3 Leave-one-year-out evaluation

Through leave-one-year-out cross-validation, we found that the initial version of the neural network-linear mixed model performed poorly. When the train and test sets were rotated by year, both the accuracy and the proportional coefficient "k" changed dramatically. After orthogonal decomposition, although some interannual variations due to emission mechanisms and patterns remained, the stability of the model improved significantly. Whether it is test R^2, RMSE or the "k" value, the range of their fluctuations over the years is much smaller than that of the direct additive model.

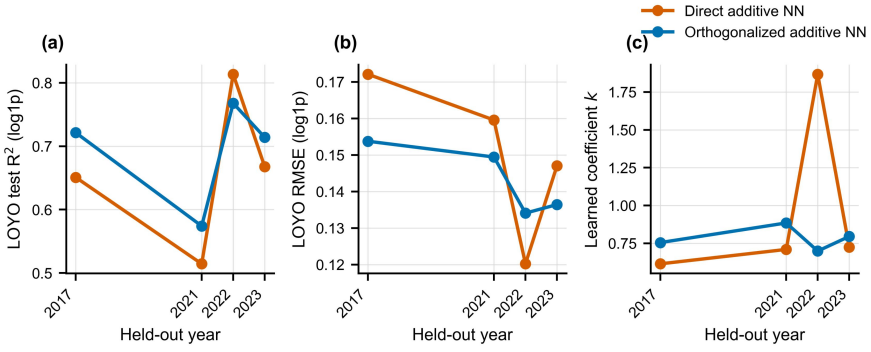
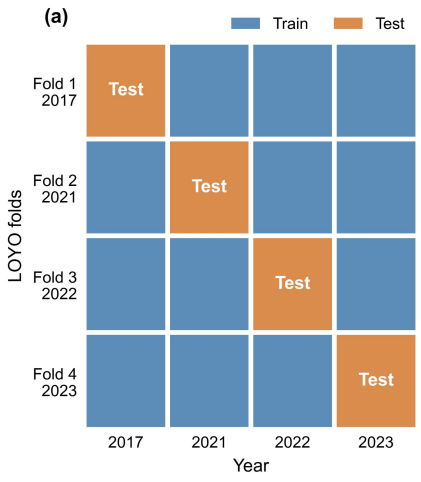


Fig A5. Comparison of the LOYO effect before and after orthogonal decomposition: (a) test R^2; (b) RMSE; (c) "k".

Fig A6 summarizes the leave-one-year-out (LOYO) evaluation of the orthogonalized additive formulation. In each fold, one complete year was held out as the test set and the remaining years were used for training (Fig A6(a)). The corresponding predictive metrics and the learned coefficient "k" are shown in Fig A6(b). To estimate the seed-related uncertainty of "k", each LOYO fold was repeated with 10 random seeds; the reported "k" values are expressed as mean ±SD across these repeated trainings.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.



(b)

Train R^2	Test R^2	RMSE	k
0.778 ±0.006	0.721 ±0.004	0.154 ±0.001	0.753 ±0.010
0.793 ±0.013	0.577 ±0.004	0.149 ±0.001	0.956 ±0.041
0.764 ±0.015	0.764 ±0.003	0.135 ±0.001	0.685 ±0.038
0.775 ±0.010	0.721 ±0.006	0.135 ±0.001	0.784 ±0.010

Mean across folds: Train R^2 = 0.778; Test R^2 = 0.696
'k' mean range = 0.685–0.956

Fig A6. Leave-one-year-out (LOYO) evaluation of the orthogonalized additive formulation. (a) Schematic of the LOYO cross-validation design. (b) Predictive performance and learned coefficient "k" for each test year. the "k" values are reported as mean ± SD from 10 repeated trainings with different random seeds.

A2.2.4 Sensitivity to predictor removal

To assess sensitivity to predictor choice, we refitted eBC-FF = $f(X_{\perp}) + k \cdot$eBC-SF after removing vehicle counts or jointly removing NO_x and HOA from the predictor set (13 orthogonalized predictors plus the linear heating term). Model settings and the 36,356 hourly observations were unchanged. Five seeds were used per configuration, with identical month-stratified 80/20 splits across configurations for each seed. Splits and training randomness varied between seeds. Holdout R^2 was evaluated in log(1 + eBC-FF) space on the subset also used for early stopping; uncertainties are across-run SDs. Contributions were calculated from full-record concentration sums, normalized by the three-component sum.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

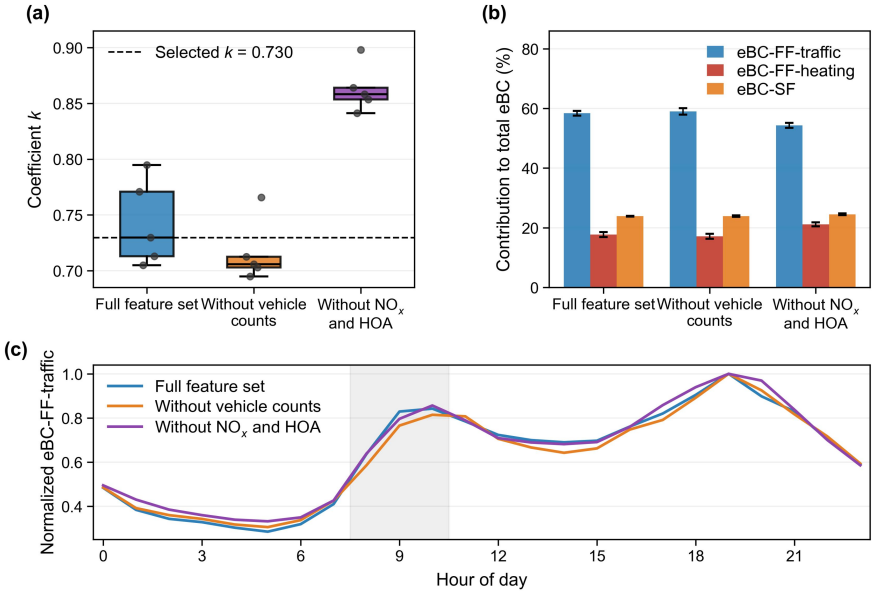


Fig A7. Feature-ablation analysis of the orthogonalized additive model. (a) Coefficient "k" across five runs; boxes show the interquartile range and median, whiskers extend to observations within 1.5 interquartile ranges, and points show individual runs. The dashed line marks k = 0.730. (b) Source contributions (mean ± SD), normalized by the summed three components. (c) Heating-season (October–March) traffic profiles: predictions averaged across runs, summarized by hourly medians and normalized to each profile's maximum. Shading marks 08:00–10:00.

The results show that removing vehicle counts had little effect on either predictive performance or component allocation (Fig A7). Holdout R^2 changed from 0.775 ± 0.008 to 0.778 ± 0.006, while "k" changed from 0.743 ± 0.039 to 0.716 ± 0.028. The estimated heating-related contribution changed from 17.73 ± 0.86% to 17.12 ± 0.86%. In contrast, jointly removing NO_x and HOA reduced holdout R^2 to 0.648 ± 0.007 and increased "k" to 0.863 ± 0.021. The corresponding heating-related contribution increased to 21.16 ± 0.64%, whereas the traffic-related contribution decreased from 58.39 ± 0.78% to 54.33 ± 0.86%. These changes represent shifts of +3.43 and −4.06 percentage points, respectively. Despite these changes in allocation, normalized heating-season traffic profiles retained broadly similar diurnal shapes, including the morning enhancement and evening maximum.

The decomposition was insensitive to removing vehicle counts but depended more strongly on the combined NO_x and HOA information. Similar normalized diurnal shapes despite shifted contributions show that temporal agreement alone does not establish attribution accuracy. These shifts indicate predictor sensitivity, not measured attribution bias. Joint removal cannot separate the individual roles of NO_x and HOA.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

A2.2.5 Synthetic recovery and assumption sensitivity

To test whether the model could recover a known heating contribution, and how its estimates would change when the underlying assumptions were violated, we generated synthetic eBC-FF as:

$$[T + k_{true} \cdot \text{eBC-SF}]e^{\varepsilon - \frac{\sigma^2}{2}}$$

Here, (T) is the traffic component predicted by the original model, and (k_{ture}) is a prescribed heating coefficient. We retained 36,356 observed hourly predictor and SF records and multiplied the synthetic eBC-FF by a lognormal factor exp(ε), equivalent to adding Gaussian noise in log space, with its standard deviation chosen to give a log-space signal-to-noise ratio of approximately 0.79. Six coefficients (0–1.50, including zero heating) were tested. We also examined three departures from the assumed decomposition: heating driven by temperature rather than proportional to SF, additional SF leakage into eBC-FF, and traffic varying with SF. Each setting was fitted three times.

At k_{true} = 0.73, the recovered coefficient was 0.732 ± 0.012; at zero truth it was 0.083 ± 0.002, revealing a non-zero recovery offset (Fig A8(a); ± denotes across-run SD). At k_{true} = 1.50, recovery was 1.436 ± 0.033. Temperature-driven heating yielded 0.082 ± 0.005, while SF leakage yielded 0.873 ± 0.036, close to the expected effective coefficient 0.88. Traffic–SF coupling yielded 0.827 ± 0.004, 13.3% above the prescribed heating coefficient 0.73 (Fig A8(b)).

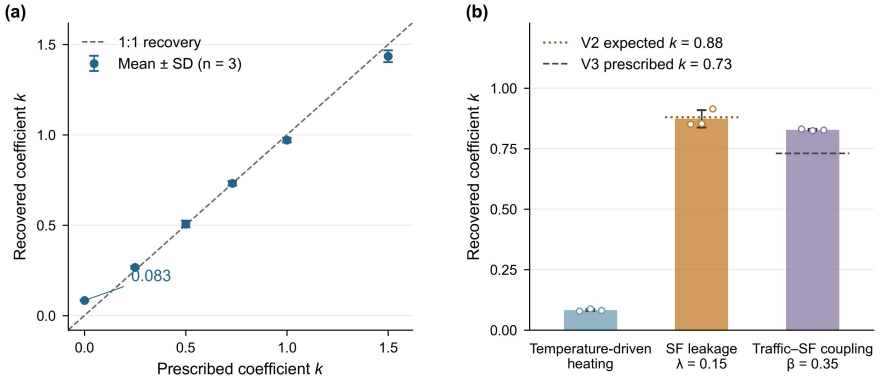


Fig A8. Synthetic coefficient recovery and sensitivity to prescribed assumption violations. (a) Recovered versus prescribed "k"; points and error bars show mean ± SD across three runs, and the dashed line denotes 1:1 recovery. (b) Temperature-driven heating (V1), SF leakage into eBC-FF (V2, λ = 0.15), and traffic–SF coupling (V3, β = 0.35). Bars and error bars show mean ± SD; open points show individual runs. Local reference lines mark the V2 expected effective coefficient (0.88) and V3 prescribed heating coefficient (0.73). V1 has no single prescribed proportional coefficient.

The estimator responded to prescribed coefficient changes, but retained a zero-truth offset and was sensitive to the imposed confounding. A small "k" under temperature-driven heating does not imply negligible actual heating. These

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

model-template simulations assess conditional recovery; they neither establish observational attribution accuracy nor bound real-world bias.

A2.3 Physical-consistency diagnostics of the apportioned components

A2.3.1 Comparison with the patterns of the emission activities

To assess whether the apportioned components exhibited temporal behavior consistent with their assigned source activities, we compared their seasonal diurnal patterns with source-related indicators (Fig A9). For Fig A9(a), traffic-related eBC-FF (eBC-FF-traffic) and vehicle counts were each divided by their season-specific mean before the hourly statistics were calculated, allowing their diurnal shapes to be compared despite their different units and absolute scales. The traffic-related component broadly reproduced the morning peak, evening peak and sustained daytime activity observed in vehicle counts across all seasons. Because vehicle counts were included among the model predictors, this comparison evaluates the physical consistency of the apportioned temporal pattern rather than providing fully independent validation.

Fig A9(b) compares heating-related eBC-FF (eBC-FF-heating) with eBC-SF in concentration units. Heating-related eBC-FF remained close to zero during summer and showed clear evening-to-nighttime enhancements in autumn and winter, consistent with the expected seasonal timing of residential heating. Because eBC-FF-heating was calculated using the optimized proportional relationship eBC-FF-heating = k·eBC-SF, similarity between the two temporal patterns is expected by construction. This panel therefore demonstrates the seasonal plausibility and implemented behavior of the heating-related term rather than serving as an independent tracer comparison.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

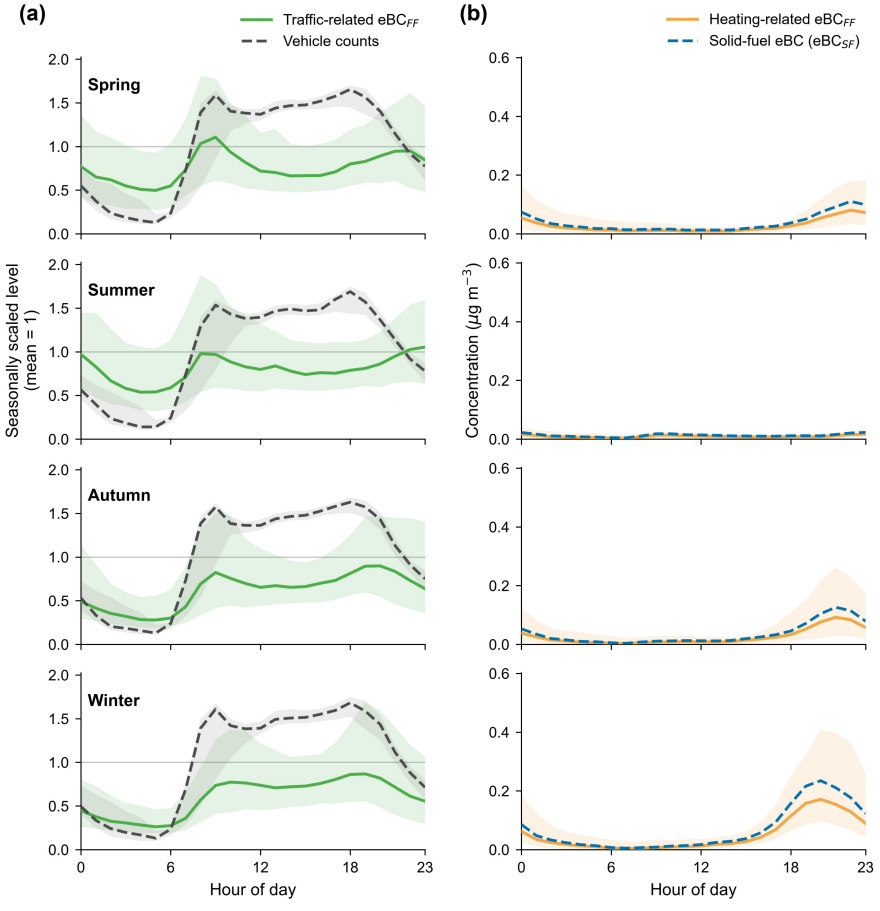


Fig A9. Seasonal diurnal patterns of the apportioned BC components and source-related indicators. (a) eBC-FF-traffic and vehicle counts, with each series divided by its season-specific mean before calculating the hourly statistics. (b) eBC-FF-heating and eBC-SF in concentration units; eBC-FF-heating was calculated as 0.730 × eBC-SF. Lines represent hourly medians, and shaded bands indicate interquartile ranges. These comparisons provide physical-consistency diagnostics of the apportioned temporal patterns rather than independent source-tracer validation.

A2.3.2 Comparison with ACSM OA factors

We paired ACSM OA factors with BC components using 36,312 complete hourly records. Peat, wood and coal factors were absent from the predictor matrix. Raw HOA and NO_x were retained as contextual references, because their SF-orthogonalized components entered the NN. Correlations were calculated after subtracting year-month-specific hourly means. Paired calendar-week block resampling (2,000 replicates) provided intervals for both correlations and their differences; raw, meteorology-adjusted and alternative block-length results were also examined.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

Peat and wood correlated more strongly with eBC-SF (r = 0.830 and 0.756) than with eBC-FF-traffic (0.418 and 0.440; Fig 8(a)). Corresponding Δr values were 0.411 (95% interval 0.241–0.572) and 0.316 (0.118–0.542). Coal showed a smaller difference, 0.097 (0.027–0.168), while the HOA interval included zero, 0.113 (−0.017–0.281). September–May diurnal profiles showed evening enhancements in SF, peat and wood, while the traffic component and NOx also displayed a morning enhancement (Fig 8(c, d)).

These comparisons support the solid-fuel interpretation of eBC-SF. However, peat and wood retained partial correlations of approximately 0.44 with the traffic component after controlling time, selected meteorology and SF, indicating unresolved co-variation. The results do not establish absence of source mixing or validate k: heating is proportional to SF, so its correlations provide no separate test of oil-heating attribution.

A2.3.3 Weekdays vs. Weekends

Weekday–weekend contrasts were used to test whether the apportioned components follow the activity patterns of their assigned sources (Fig A10). June–August were excluded because residential heating is minimal in summer, so a weekend effect on the heating term would not be resolvable; days with fewer than 18 valid hourly records were discarded. Over the remaining months, the traffic component showed a higher daily mean on weekdays than at weekends (0.339 vs 0.310 μg/m³) and a sharper morning enhancement between 08:00 and 10:00, whereas the timing and amplitude of its evening maximum were nearly unchanged. The heating component behaved in the opposite sense: both its daily mean and its daily maximum hourly concentration were higher at weekends, consistent with more time spent at home. Because vehicle counts were included among the neural-network predictors, the weekday enhancement of the traffic component is partly reproduced by construction. The weekend enhancement of the heating component is not constructed in the same way, since it follows the observed weekend behaviour of eBC-SF rather than an imposed term; the heating component is nevertheless simply k·eBC-SF, so this contrast carries no information on the ratio between oil and solid-fuel heating. This comparison is accordingly an activity-consistency check on the heating proxy rather than an independent constraint on the oil-heating amount.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

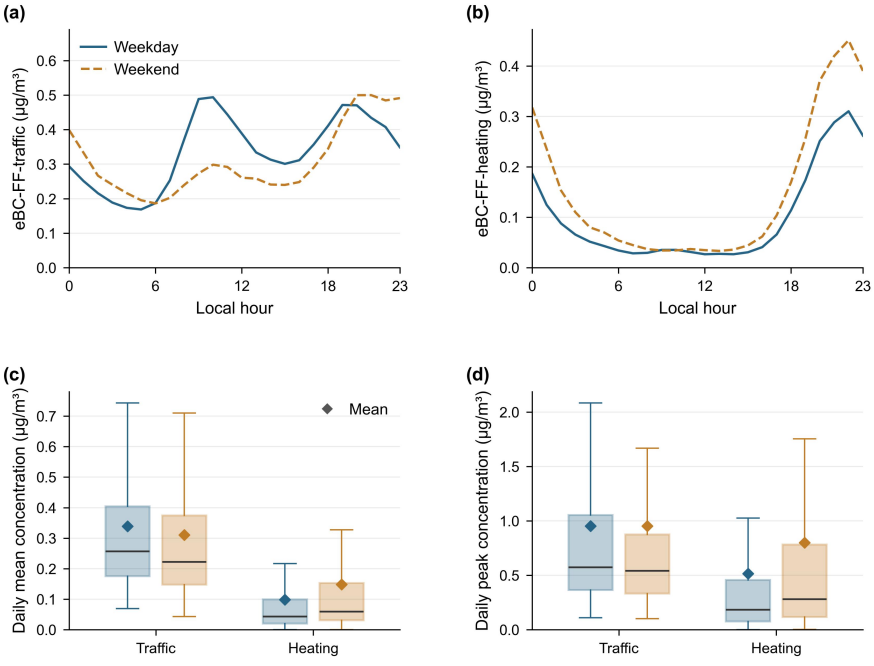


Fig A10. Weekday–weekend comparison excluding June–August. Mean diurnal profiles of (a) eBC-FF-traffic and (b) eBC-FF-heating; distributions of (c) daily mean and (d) daily maximum hourly concentrations. Blue and orange denote weekdays and weekends. Boxes show interquartile ranges and medians; whiskers extend to observations within 1.5 interquartile ranges, and diamonds indicate means. Outliers are omitted from display but included in means. Days require at least 18 valid hourly records.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

Data and materials availability:

The data, analysis code, and model files supporting this study, including those for the counterfactual, direct additive, and orthogonal additive models, have been prepared for deposition in Zenodo under the reserved DOI: https://doi.org/10.5281/zenodo.22700012. The repository record is currently unpublished, and the materials are available to the editor and referees through a private review-access link during peer review.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (No. 42577115); Natural Science Foundation of Fujian Province of China (No. 2025J01261); State Environmental Protection Key Laboratory of Formation and Prevention of Urban Air Pollution Complex (No. 2025080172); International Partnership Program of the Chinese Academy of Sciences (Grant No. 322GJHZ2025063FN); State Key Laboratory of Advanced Environmental Technology (Grant No. SKLAET2025-LH03); the Guiding Project of Seizing the Commanding Heights of City Self-Purifying Technology (No. IUE-CERAE-202403); Chinese Academy of Sciences (No. ZDBS-LY-DQC001); European Union's Horizon Europe Research and Innovation programme under HORIZON-CL5-2022-D1-02 (grant no. 101081430 – PARIS), Taighde Éireann – Research Ireland under Grant number [22/FFP-A/10611], the EPA-Ireland and the Department of Climate, Energy and the Environment.

Author Contributions:

Conceptualization: WX, JO, LL, XYW, JZ
Data curation: LL, JO, VP, JZ, CHL, KF, DC, CO
Methodology: XYW, WX, LL, JO, JZ, BHC
Investigation: LL, JO, XYW, WX, VP
Visualization: XYW, LL
Supervision: WX, JO
Writing—original draft: XYW, LL
Writing—review & editing: WX, LL, VP, JO, JZ, CHL, KF, BHC, DC, CO

Competing interest:

The authors declare no competing interests.

Reference

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., and Koch, D.: Bounding the role of black

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

carbon in the climate system: A scientific assessment, Journal of geophysical research: Atmospheres, 118, 5380-5552, 2013.
Household Environmental Behaviours – Energy Use Quarter 3 2021: https://www.cso.ie/en/releasesandpublications/er/hebeu/householdenvironmentalbehaviours-energyusequarter32021/, last access: 20 May.
Change, O. C.: Intergovernmental panel on climate change, World Meteorological Organization, 52, 1, 2007.
Chen, B., Lei, L., Chevassus, E., Xu, W., Zhen, L., Zhong, H., Wang, L., Lin, C., Huang, R.-J., and Ceburnis, D.: Differentiation of primary and secondary marine organic aerosol with machine learning, Atmospheric Chemistry and Physics, 25, 14205-14219, 2025a.
Chen, J., Yu, W., Cui, X., Zhang, Q., Li, T., Wang, X., Ren, Y., Wang, H., and Wang, Q.: Black carbon in urban Jinan: Variations, health risks, and driving factors analyzed with machine learning, Journal of Environmental Sciences, 2025b.
Denier Van Der Gon, H., Bergström, R., Fountoukis, C., Johansson, C., Pandis, S., Simpson, D., and Visschedijk, A. J.: Particulate emissions from residential wood combustion in Europe–revised estimates and an evaluation, Atmospheric Chemistry and Physics, 15, 6503-6519, 2015.
Drinovec, L., Močnik, G., Zotter, P., Prévôt, A. S. H., Ruckstuhl, C., Coz, E., Rupakheti, M., Sciare, J., Müller, T., Wiedensohler, A., and Hansen, A. D. A.: The "dual-spot" Aethalometer: an improved measurement of aerosol black carbon with real-time loading compensation, Atmospheric Measurement Techniques, 8, 1965-1979, 10.5194/amt-8-1965-2015, 2015.
Goodman, P. G., Rich, D. Q., Zeka, A., Clancy, L., and Dockery, D. W.: Effect of air pollution controls on black smoke and sulfur dioxide concentrations across Ireland, Journal of the Air & Waste Management Association, 59, 207-213, 2009.
Grigas, T., Ovadnevaite, J., Ceburnis, D., Moran, E., McGovern, F., Jennings, S. G., and O'Dowd, C.: Sophisticated clean air strategies required to mitigate against particulate organic pollution, Scientific Reports, 7, 44737, 2017.
Han, Y., Cao, J., Lee, S., Ho, K., and An, Z.: Different characteristics of char and soot in the atmosphere and their ratio as an indicator for source identification in Xi'an, China, Atmospheric Chemistry and Physics, 10, 595-607, 2010.
Herich, H., Hueglin, C., and Buchmann, B.: A 2.5 year's source apportionment study of black carbon from wood burning and fossil fuel combustion at urban and rural sites in Switzerland, Atmospheric Measurement Techniques, 4, 1409-1420, 2011.
Hopke, P. K.: Review of receptor modeling methods for source apportionment, Journal of the Air & Waste Management Association, 66, 237-259, 10.1080/10962247.2016.1140693, 2016.
IPCC: Climate Change 2014: Synthesis Report, Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2014.
Janssen, N. A., Gerlofs-Nijland, M. E., Lanki, T., Salonen, R. O., Cassee, F., Hoek, G., Fischer, P., Brunekreef, B., and Krzyzanowski, M.: Health effects of black carbon, World Health Organization. Regional Office for Europe2012.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.



Jimenez, J. L., Canagaratna, M., Donahue, N., Prevot, A., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., and Ng, N.: Evolution of organic aerosols in the atmosphere, science, 326, 1525-1529, 2009.

Lei, L., Xu, W., Lin, C., Chen, B., Fossum, K. N., Ceburnis, D., O'Dowd, C., and Ovadnevaite, J.: Enhancing Differentiation of Oxygenated Organic Aerosol: A Machine Learning Approach to Distinguish Local and Transboundary Pollution, ACS Es&t Air, 2, 891-902, 2025.

Lei, L., Xu, W., Lin, C., Fossum, K. N., Ceburnis, D., Gallagher, J., D'Dowd, C., and Ovadnevaite, J.: Measurement Report: Evolving Sources and Composition of Urban Submicron Aerosols in Dublin: Impacts of Emission Reductions and Transboundary Transport, EGUsphere, 2026, 1-25, 2026.

Lin, C., Ceburnis, D., Trubetskaya, A., Xu, W., Smith, W., Hellebust, S., Wenger, J., O'Dowd, C., and Ovadnevaite, J.: On the use of reference mass spectra for reducing uncertainty in source apportionment of solid-fuel burning in ambient organic aerosol, Atmospheric Measurement Techniques, 14, 6905-6916, 2021.

Lin, C., Ceburnis, D., Xu, W., Heffernan, E., Hellebust, S., Gallagher, J., Huang, R.-J., O'Dowd, C., and Ovadnevaite, J.: The impact of traffic on air quality in Ireland: insights from the simultaneous kerbside and suburban monitoring of submicron aerosols, Atmospheric Chemistry and Physics, 20, 10513-10529, 2020.

Lin, C., Ceburnis, D., Hellebust, S., Buckley, P., Wenger, J., Canonaco, F., Prévôt, A. S. H., Huang, R.-J., O'Dowd, C., and Ovadnevaite, J.: Characterization of primary organic aerosol from domestic wood, peat, and coal burning in Ireland, Environmental science & technology, 51, 10624-10632, 2017.

Lin, C., Ceburnis, D., Huang, R.-J., Xu, W., Spohn, T., Martin, D., Buckley, P., Wenger, J., Hellebust, S., and Rinaldi, M.: Wintertime aerosol dominated by solid-fuel-burning emissions across Ireland: insight into the spatial and chemical variation in submicron aerosol, Atmospheric Chemistry and Physics, 19, 14091-14106, 2019.

Lin, C., Huang, R.-J., Ceburnis, D., Buckley, P., Preissler, J., Wenger, J., Rinaldi, M., Facchini, M. C., O'Dowd, C., and Ovadnevaite, J.: Extreme air pollution from residential solid fuel burning, Nature Sustainability, 1, 512-517, 2018.

Liu, H., Pan, X., Lei, S., Zhang, Y., Du, A., Yao, W., Tang, G., Wang, T., Xin, J., and Li, J.: Vertical distribution of black carbon and its mixing state in the urban boundary layer in summer, Atmospheric Chemistry and Physics, 23, 7225-7239, 2023.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, 30, 2017.

Mouteva, G. O., Randerson, J. T., Fahrni, S. M., Bush, S. E., Ehleringer, J. R., Xu, X., Santos, G. M., Kuprov, R., Schichtel, B. A., and Czimczik, C. I.: Using radiocarbon to constrain black and organic carbon aerosol sources in Salt Lake City, Journal of Geophysical Research: Atmospheres, 122, 9843-9857, 2017.

Ng, N. L., Herndon, S. C., Trimborn, A., Canagaratna, M. R., Croteau, P., Onasch, T. B., Sueper, D., Worsnop, D. R., Zhang, Q., and Sun, Y.: An Aerosol Chemical Speciation Monitor (ACSM) for routine monitoring of the composition and mass

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.

concentrations of ambient aerosol, Aerosol Science and Technology, 45, 780-794, 2011.
Ni, H., Huang, R.-J., Cao, J., Liu, W., Zhang, T., Wang, M., Meijer, H. A., and Dusek, U.: Source apportionment of carbonaceous aerosols in Xi'an, China: insights from a full year of measurements of radiocarbon and the stable isotope 13 C, Atmospheric Chemistry and Physics, 18, 16363-16383, 2018.
Organization, W. H.: Health Effects of Black Carbon, World Health Organization Regional Office for Europe, Copenhagen, 2013.
Paatero, P. and Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5, 111-126, 1994.
Pokhrel, R. P., Wagner, N. L., Langridge, J. M., Lack, D. A., Jayarathne, T., Stone, E. A., Stockwell, C. E., Yokelson, R. J., and Murphy, S. M.: Parameterization of single-scattering albedo (SSA) and absorption Ångström exponent (AAE) with EC/OC for aerosol emissions from biomass burning, Atmospheric Chemistry and Physics, 16, 9549-9561, 2016.
Pósfai, M., Gelencsér, A., Simonics, R., Arató, K., Li, J., Hobbs, P. V., and Buseck, P. R.: Atmospheric tar balls: Particles from biomass and biofuel burning, Journal of Geophysical Research: Atmospheres, 109, 2004.
Saarikoski, S., Niemi, J. V., Aurela, M., Pirjola, L., Kousa, A., Rönkkö, T., and Timonen, H.: Sources of black carbon at residential and traffic environments obtained by two source apportionment methods, Atmospheric Chemistry and Physics, 21, 14851-14869, 2021.
Sandradewi, J., Prévôt, A. S. H., Szidat, S., Perron, N., Alfarra, M. R., Lanz, V. A., Weingartner, E., and Baltensperger, U.: Using Aerosol Light Absorption Measurements for the Quantitative Determination of Wood Burning and Traffic Emission Contributions to Particulate Matter, Environmental Science & Technology, 42, 3316-3323, 10.1021/es702253m, 2008.
Szidat, S., Jenk, T. M., Synal, H. A., Kalberer, M., Wacker, L., Hajdas, I., Kasper-Giebl, A., and Baltensperger, U.: Contributions of fossil fuel, biomass-burning, and biogenic emissions to carbonaceous aerosols in Zurich as traced by 14C, Journal of Geophysical Research: Atmospheres, 111, 2006.
Ulbrich, I., Canagaratna, M., Zhang, Q., Worsnop, D., and Jimenez, J.: Interpretation of organic components from Positive Matrix Factorization of aerosol mass spectrometric data, Atmospheric Chemistry and Physics, 9, 2891-2918, 2009.
Wang, Y., Xu, W., Zhong, H., Chen, B., Yang, L., Wang, L., Ren, J., Duan, J., Lin, C., and Huang, R.-J.: Advancing Aerosol Chemistry with Machine Learning: A Short Review, ACS ES&T Air, 2, 2323-2341, 2025.
Winiger, P., Barrett, T. E., Sheesley, R. J., Huang, L., Sharma, S., Barrie, L. A., Yttri, K. E., Evangeliou, N., Eckhardt, S., Stohl, A., Klimont, Z., Heyes, C., Semiletov, I. P., Dudarev, O. V., Charkin, A., Shakhova, N., Holmstrand, H., Andersson, A., and Gustafsson, Ö.: Source apportionment of circum-Arctic atmospheric black carbon from isotopes and modeling, Science Advances, 5, eaau8052, doi:10.1126/sciadv.aau8052, 2019.

https://doi.org/10.5194/egusphere-2026-7038
Preprint. Discussion started: 20 November 2026
© Author(s) 2026. CC BY 4.0 License.



Wu, C. and Yu, J. Z.: Determination of primary combustion source organic carbon-to-elemental carbon (OC/EC) ratio using ambient OC and EC measurements: secondary OC-EC correlation minimization method, Atmospheric Chemistry and Physics, 16, 5453-5465, 2016.

Xu, L., Fukushima, S., Sobanska, S., Murata, K., Naganuma, A., Liu, L., Wang, Y., Niu, H., Shi, Z., and Kojima, T.: Tracing the evolution of morphology and mixing state of soot particles along with the movement of an Asian dust storm, Atmospheric Chemistry and Physics Discussions, 2020, 1-26, 2020.

Yttri, K. E., Canonaco, F., Eckhardt, S., Evangeliou, N., Fiebig, M., Gundersen, H., Hjellbrekke, A.-G., Lund Myhre, C., Platt, S. M., and Prévôt, A. S.: Trends, composition, and sources of carbonaceous aerosol at the Birkenes Observatory, northern Europe, 2001–2018, Atmospheric Chemistry and Physics, 21, 7149-7170, 2021.

Zotter, P., Herich, H., Gysel, M., El-Haddad, I., Zhang, Y., Močnik, G., Hüglin, C., Baltensperger, U., Szidat, S., and Prévôt, A. S.: Evaluation of the absorption Ångström exponents for traffic and wood burning in the Aethalometer-based source apportionment using radiocarbon measurements of ambient aerosol, Atmospheric Chemistry and Physics, 17, 4229-4249, 2017.