JGR Atmospheres

RESEARCH ARTICLE
10.1029/2019JD031380

Mapping and Understanding Patterns of Air Quality Using Satellite Data and Machine Learning

Roland Stirnberg[1,2], **Jan Cermak**[1,2], **Julia Fuchs**[1,2], **and Hendrik Andersen**[1,2]

[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, [2]Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Key Points:
- Drivers of concentrations of particulate matter (PM10) can successfully be quantified in a machine learning model
- Important drivers of PM10 include easterly wind flow, boundary layer height, and temperature
- The relationship between aerosol optical depth (AOD) and PM10 strongly depends on ambient meteorological conditions

Correspondence to:
R. Stirnberg,
roland.stirnberg@kit.edu

Citation:
Stirnberg, R., Cermak, J., Fuchs, J., & Andersen, H. (2020). Mapping and understanding patterns of air quality using satellite data and machine learning. *Journal of Geophysical Research: Atmospheres*, *125*, e2019JD031380. https://doi.org/10.1029/2019JD031380

Received 19 JUL 2019
Accepted 4 FEB 2020
Accepted article online 6 FEB 2020

Author Contributions

Conceptualization: Roland Stirnberg, Jan Cermak
Methodology: Roland Stirnberg, Jan Cermak, Julia Fuchs, Hendrik Andersen
Writing - Original Draft: Roland Stirnberg
Formal Analysis: Roland Stirnberg
Investigation: Roland Stirnberg
Visualization: Roland Stirnberg
Writing - review & editing: Roland Stirnberg, Jan Cermak, Julia Fuchs, Hendrik Andersen

Abstract The quantification of factors leading to harmfully high levels of particulate matter (PM) remains challenging. This study presents a novel approach using a statistical model that is trained to predict hourly concentrations of particles smaller than 10 μm (PM10) by combining satellite-borne aerosol optical depth (AOD) with meteorological and land-use parameters. The model is shown to accurately predict PM10 (overall $R^2 = 0.77$, RMSE = 7.44 $\mu g/m^3$) for measurement sites in Germany. The capability of satellite observations to map and monitor surface air pollution is assessed by investigating the relationship between AOD and PM10 in the same modeling setup. Sensitivity analyses show that important drivers of modeled PM10 include multiday mean wind flow, boundary layer height (BLH), day of year (DOY), and temperature. Different mechanisms associated with elevated PM10 concentrations are identified in winter and summer. In winter, mean predictions of PM10 concentrations >35 $\mu g/m^3$ occur when BLH is below ~500 m. Paired with multiday easterly wind flow, mean model predictions surpass 40 $\mu g/m^3$ of PM10. In summer, PM10 concentrations seemingly are less driven by meteorology, but by emission or chemical particle formation processes, which are not included in the model. The relationship between AOD and predicted PM10 concentrations depends to a large extent on ambient meteorological conditions. Results suggest that AOD can be used to assess air quality at ground level in a machine learning approach linking it with meteorological conditions.

Plain Language Summary In this study, factors leading to severe air pollution are determined using machine learning. In addition, it is tested, to what extent, that the use of satellite data is adequate to derive information on air quality near ground. It is shown that besides human emissions, concentrations of particles in the air are to a large extent driven by meteorological factors such as wind direction, state of the atmospheric boundary layer, and season.

1. Motivation and Research Questions

Extensive research has been conducted in recent years on the adverse health effects of particulate matter (PM) on the human cardiovascular system and the lungs. Cohort studies show that negative effects include emphysema, lung cancer, diabetes, and hypertension (Lelieveld et al., 2019; Lim et al., 2012; Pope et al., 2002; Wichmann et al., 2000), which cause a large number of premature deaths (Lelieveld et al., 2019, 2015). Although these risks are largely known and confirmed, discussions on effective measures to reduce exposure to air pollution are ongoing. Suggested measures range from traffic bans for certain vehicle types (Ellison et al., 2013; Qadir et al., 2013) over a reduced or more efficient use of solid fuel-based residential heating (Chafe et al., 2015) to the expansion of urban vegetation (Bonn et al., 2016). However, the actual effects of those measures are not always evident as not only local emissions but also ambient conditions such as meteorology, vegetation, and other land cover can play a substantial role in determining local PM concentrations. Therefore, understanding and quantifying drivers of PM concentrations is necessary to improve the efficiency of measures toward better air quality. The influence of meteorological conditions on particulate matter concentrations are diverse (see Figure 1) and have been described, for example, by Rost et al. (2009), Reizer and Juda-Rezler (2016), Dupont et al. (2016), Li et al. (2017), and Fuzzi et al. (2015). Boundary layer height (BLH) determines the height up to which particles are distributed in the atmosphere (Gupta & Christopher, 2009a; Schäfer et al., 2012). Precipitation leads to a substantial reduction in PM concentrations by wet scavenging (Li et al., 2015; Rost et al., 2009). Wind regulates particle transport and turbulent mixing away from the surface (Li et al., 2017). Depending on the location of the measuring point, air masses from

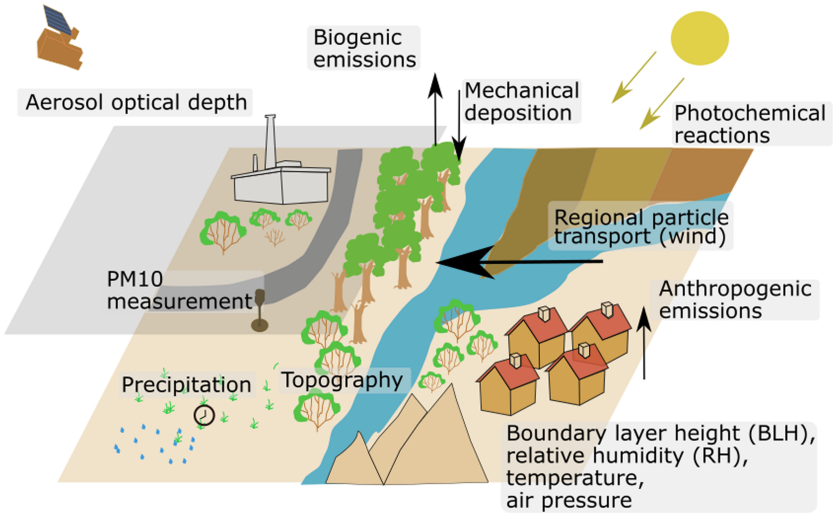

Figure 1. Conceptual design of potential influences on variations in hourly PM10 concentrations. The machine learning model is set up to reproduce these variations.

certain wind directions transport particles and lead to elevated concentrations (Lenschow et al., 2001; Li et al., 2015). Temperature can regulate the particle number in the atmosphere by stimulating photochemical reactions, which transform precursor gases to secondary aerosols (Gupta & Christopher, 2009a; Wang & Martin, 2007) or by causing partitioning of condensed precursor gases (Petit et al., 2014). Various land cover types can act as sinks or sources for particles (Beloconi et al., 2018; Bonn et al., 2016; Churkina et al., 2017). Topography can be of importance for PM concentrations, for example, as particles accumulate in valleys (Emili et al., 2011; Emili et al., 2011).

Satellite AOD provides information on atmospheric particle concentrations and can expand information to areas where station measurements are sparse (Emili et al., 2011), revealing hotspots and spatiotemporal variations of pollution (Cermak & Knutti, 2009; Gupta et al., 2006). However, relying on satellite AOD as a proxy for near-ground air pollution can be misleading, as AOD reflects the extinction of radiation in an atmospheric column, while particulate matter concentrations reflect a highly localized dry mass concentration of particles of a certain size distribution typically measured near ground (Wang & Christopher, 2003). Several studies have trained statistical models on the relationship between AOD and PM, accounting for a range of additional parameters, and mostly with a focus on applications (see review by Rybarczyk, 2018). Methods include linear regression models (Arvani et al., 2016), multiple-additive regression models (Gupta & Christopher, 2009a; Zhang et al., 2018), land-use models (Kloog et al., 2011; Nordio et al., 2013), or a combination of the latter two (Chudnovsky et al., 2014; Hu et al., 2014; Kloog et al., 2012). With increasing data availability and computational power, machine learning methods, for example, artificial neural networks (Gupta & Christopher, 2009b; Di et al., 2016) and random forests (RF) (Brokamp et al., 2017; Chen et al., 2018; Grange et al., 2018) have been applied frequently in recent years. These machine learning models are beneficial as they efficiently reproduce nonlinear relationships and interactions of input features (Brokamp et al., 2017; Elith et al., 2008). In contrast to physical models, machine learning approaches do not require extensive prior process knowledge and thus have the potential to reveal or quantify processes that are as yet undetermined (Knüsel et al., 2019). Multivariate processes can be investigated by isolating certain variables and studying inputs and responses for dominant patterns (Andersen et al., 2017; Cermak & Knutti, 2009; Fuchs et al., 2018).

While numerous air pollution studies apply statistical models mainly to accurately predict PM concentrations (Hu et al., 2017; Kloog et al., 2015; Stafoggia et al., 2017; van Donkelaar et al., 2010; Zhang et al., 2018), recent studies additionally analyzed feature importances and the information content of explanatory variables in the statistical models to infer processes. For example, Park et al. (2019) used a RF model to predict PM10 in South Korea and found large influence of AOD, day of the year (DOY), wind speed, and solar radiation on the modeled PM concentrations. Similarly, Grange et al. (2018) obtained high feature importances for wind speed, back trajectory cluster, DOY, and air temperature, using the RF approach for PM10

Table 1
List of Input Features Stating Data Origin, Data Representation, and Abbreviation

Data set	Variable	Abbreviation
	Input features	
MODIS MAIAC	Aerosol optical depth	AOD
MODIS	Normalized difference Vegetation index	NDVI
NASA Earth at night	Lights at night	Nightlights
EU-DEM, v1.1	Elevation (m),	Elev,
	Topographic position index (m)	TPI
CORINE land cover	Land cover types	1CLC, 2CLC, 3CLC, 4CLC 5CLC, 6CLC
DWD meteorological measurements	Air pressure (sea level) (hPa)	AirPres
	Relative humidity (%)	RH
	Continentality	Conti
RADOLAN	Time since last precip. (hr),	Precip_tsince,
	Magnitude of last precip. (mm/hr),	Precip_magn,
	Cumulative precip. last 24h (mm),	Precip_acc
ERA-Interim	wind vectors u,v (m/s),	u,v
	mean wind vectors of previous hours (m/s)	umean, vmean
	Boundary layer height (m)	BLH
	Temperature (°),	T
	temperature anomalies (K),	Tan
	Convective available potential energy (J/kg)	CAPE
	Surface solar radiation downwards (J/m^2)	SSRD
EEA emission data base	Annual emission of NH_3,	NH3, PM10em,
	PM10, SO_2, NO_x (t/year)	SO2, NOx
Other	Day of year	DOY
	Day of week	Weekday
	Model outcome	
UBA air quality measurements	PM10 concentrations ($\mu g/m^3$)	PM10

measurements in Switzerland. The present study builds upon the approaches applied in these studies but provides a more in-depth analysis of model-inherent relationships. To this end, gradient boosted regression trees (GBRT) are used to understand and quantify the conditions driving air quality, as well as determinants of the relationship between AOD and PM10. GBRT have been successfully applied to study sensitivities of aerosol processes before (cf. Fuchs et al., 2018). Thus, a basis is set for targeted satellite-based analyses of spatial patterns of air quality.

2. Data and Methods

The data basis of this study is comprised of 8 years (2007–2015) including satellite observations from the moderate resolution imaging spectroradiometer (MODIS) and others, model output from the European Centre for Medium-Range Weather Forecasts (ECMWF), and station data from the German Meteorological Service (Deutscher Wetterdienst, DWD) and the German Federal Environmental Agency (Umweltbundesamt, UBA). AOD observations are based on the multi-angle implementation of atmospheric correction (MAIAC) algorithm. A statistical model is set up to predict hourly PM10 concentrations based on a variety of input features, which are summarized in Table 1. Data with high temporal resolution are selected to

reflect the atmospheric situation close to the satellite overpass times. Generally, station measurements were preferred for parameters, which have adequate coverage and which are not expected to vary on small scales (e.g., air pressure).

2.1. Satellite Data

2.1.1. MAIAC AOD

Satellite-borne, high-resolution (1*1 km) MAIAC Collection 6 AOD is used (Lyapustin et al., 2011a; Lyapustin, Martonchik, et al. 2011; Lyapustin et al. 2011b; Lyapustin et al., 2018). The product is based on data from the MODIS sensor aboard the Terra and Aqua satellites. The MAIAC algorithm makes use of look-up tables, explicitly taking into account surface bidirectional reflectance factors (BRF). The calculation of AOD relies on the assumption that surface BRFs remain largely constant over time, considering a time series of 16 consecutive days (Lyapustin et al., 2011a; Lyapustin, Martonchik, et al. 2011). Quality flags are applied for filtering. Pixels were only incorporated when clear conditions are reported, that is, no contamination of the data due to clouds is to be expected (Lyapustin et al., 2018). An additional filter was set up to exclude AOD close to clouds to avoid increased AOD near cloud fringes due to aerosol swelling effects (Schwarz et al., 2017; Várnai et al., 2013). Therefore, the distance to the nearest cloud as classified by the MAIAC algorithm (Lyapustin et al., 2018) was calculated and a threshold of 0.1° was set, which corresponds to ~7 km. This is in the range of what previous studies used as threshold (Emili et al., 2011; Koren et al., 2007). Terra or Aqua satellite overpass times were used as temporal reference for data collocation, that is, other input feature values closest to the overpass times were used. Valid daytime AOD acquisition times as used in this study range from ~9 to ~1 UTC. AOD is an important input to the statistical model as it provides implicit information on the total aerosol loading in the atmosphere, reflecting natural as well as anthropogenic sources.

2.1.2. MODIS NDVI

Sixteen-day NDVI averages obtained from the MODIS MOD13Q1 Version 6 product are used (DOI: 10.5067/MODIS/MOD13Q1.006) to approximate photosynthetically active vegetation (Tucker C. J., 1979). The NDVI was found to influence PM concentrations in several previous works (Beloconi et al., 2018; Chudnovsky et al., 2014; Stafoggia et al., 2017). Vegetation acts as a sink by increasing the aerodynamic roughness and available surface for mechanical deposition (Fuzzi et al., 2015). On the other hand, vegetation can increase background particle concentrations by emitting pollen in spring (Fuzzi et al., 2015) and by enhancing the creation of secondary organic aerosols (SOA, Churkina et al., 2017). Around each PM station coordinate, a window with an edge size of 20 km is established to reflect the local contribution of vegetation to the total particle concentration. An edge size of 20 km is comparable to the mean representativeness of the PM stations (EU, 2008). The mean NDVI of each window was used as a predictor.

2.1.3. NASA Night Lights

The NASA night lights data set is included as a surrogate for population density. Based on the data from the Visible Infrared Imaging Radiometer Suite (VIIRS) onboard the Suomi National Polar-Orbiting Platform (SNPP), the night-time lights product is available at 500 m resolution (Román et al., 2018). Population density is an important factor contributing to PM10 concentrations as it reflects human activity (Beloconi et al., 2018; Park et al., 2019). The mean night light intensity of a 20 km window around each PM station was used as a predictor.

2.1.4. EEA DEM

Topography can have a marked influence on the accumulation of particles (Emili et al., 2011). Here, the v1.1 EU-DEM is used, which is a hybrid product produced by the European Environment Agency (EEA) based on SRTM and ASTER GDEM fused by a weighted averaging approach. It has a spatial resolution of 25*25 m (more information and download at https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1?tab=metadata). Station altitude and the dominant topography around each PM station are incorporated as predictors. A topographic position index (TPI) is computed to provide information on the topography of a pixel relative to its surrounding pixels. It employs the station altitude and subtracts the mean altitude of surrounding pixels in its vicinity. Positive values indicate a summit position, whereas negative values indicate a valley position (Egli et al., 2018). Again, a window size of 20 km was chosen.

2.1.5. EEA Corine Land Cover

Data from the CORINE land cover (CLC) inventory for the years 2006 and 2012 are used (Bossard et al., 2000), accessed via https://land.copernicus.eu/pan-european/corine-land-cover/view. In its finest thematic accuracy, the CLC data set consists of 44 classes. For this study, a more simplistic classification in six land cover types was chosen, which represent the most relevant land cover types influencing air quality (see

Table 2
Original Labels and Aggregation of Corine Land Cover Classes Used in This Study

Code	Label	Original labels included
1	Artificial surfaces	Continuous urban fabric, discontinuous urban fabric, industrial or commercial units, road and rail networks and associated land, port areas, airports, mineral extraction sites, dump sites, construction sites, green urban areas
2	Agricultural areas	Nonirrigated arable land, permanently irrigated land, rice fields, vineyards, fruit trees and berry plantations, olive groves, pastures, annual crops associated with permanent crops, complex cultivation patterns, land principally occupied by agriculture with significant areas of natural vegetation, agro-forestry areas
3	Forest and seminatural areas	Broad-leaved forest, coniferous forest, mixed forest, natural grasslands, moors and heathland, sclerophyllous vegetation, transitional woodland-shrub, beaches & dunes & sands, bare rocks, sparsely vegetated areas, burnt areas, glaciers and perpetual snow
4	Wetlands	Inland marshes, peat bogs, salt marshes, salines, intertidal flats
5	Water bodies	Water courses, water bodies, coastal lagoons, estuaries, sea and ocean
6	No data, unclassified	No data, unclassified land surface, unclassified water bodies, unclassified

Table 2). The data have a spatial resolution of 250 m. Because the CLC data are categorical, a window of 20 km edge size was set up, and the number of occurrences of each type in that window was calculated. The number of occurrences of each feature is used as predictor for the model.

2.2. Reanalysis Data

The Era-Interim reanalysis data by the ECMWF provides full spatial coverage for the whole study period with an interpolated spatial resolution of 0.125 ° (Dee et al., 2011). The data set has been successfully used in previous air quality studies (cf. Chen et al., 2018; Stafoggia et al., 2017; Zheng et al., 2017). To capture regional transport of particles, ERA-Interim reanalysis wind components (m/s) in east-west and north-south direction (10 m height) are used. Wind direction and speed can influence both particle concentrations (Beloconi et al., 2018; Chudnovsky et al., 2013; Li et al., 2015) and the relationship between AOD and PM10 (Stirnberg et al., 2018; Zheng et al., 2017). Wind direction and speed are included as instantaneous values and as the mean of precedent days (72 hr). BLH data are employed to approximate the vertical distribution of aerosols in the lower troposphere, implying that particles are well mixed within the boundary layer (Ansmann et al., 2000; Gupta & Christopher, 2009a). Dispersion of particles within a high and well-mixed boundary layer leads to reduced particle numbers near ground. If the BLH is low, particles accumulate and increase PM concentrations near ground as they are constrained to a smaller volume (Gupta & Christopher, 2009a; Wagner & Schäfer, 2017). Era-Interim temperature in 2 m height is included as instantaneous values and as temperature anomaly. Temperature anomalies are determined as the deviation of the expected value for each day of the year. Expected values are calculated by averaging daily values over 30 years, then calculating a running mean over 30 days to achieve a smooth sequence of the expected temperature. In addition to temperature, downward surface solar radiation (SSRD) and convective potential energy (CAPE) are included to capture potential secondary aerosol formation based on photochemical transformation processes (Wang & Martin, 2007) and to capture convective mixing in the atmosphere (Chudnovsky et al., 2013).

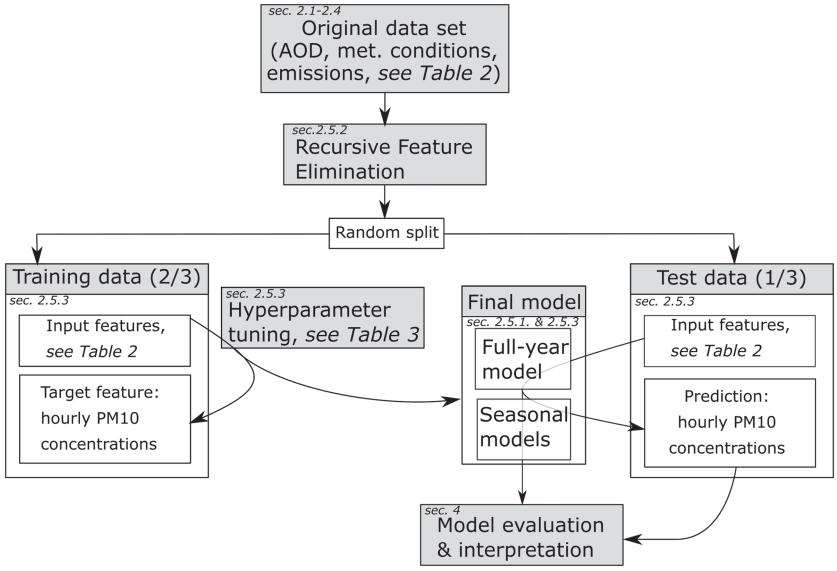

Figure 2. Framework of this study: from input feature selection (i.e., separating important features from noise), hyperparameter tuning, model training, and model evaluation to the interpretation of the results.

2.3. Ground-Based Data

2.3.1. UBA PM10

The focus here is on PM10 rather than PM2.5 because the latter excludes the fraction of larger particles, which are nevertheless accountable for light extinction and thus contribute to AOD measured by the satellite (Emili et al., 2011). In addition, PM10 measurements are more widely available. The PM stations are maintained by the UBA and measure the hourly mean PM10 concentration. PM concentrations are determined by measuring the attenuation of β-radiation by a dust-coated filter (Umweltbundesamt, 2004) or by the principle of oscillating micro weighing (TÜVRheinland, 2012). To avoid condensation, the particle inlet is heated permanently. Thus, measurements are largely uninfluenced by temperature and humidity (Umweltbundesamt, 2004). The uncertainty of the continuous measurements is prescribed to be below 25% (Bundesministerium der Justiz und für Verbraucherschutz, 2010; VDI, 2002). Instructions by the European Union prescribe site locations that avoid microenvironmental effects. Stations are classified as background, industrial, or traffic and are designed to be representative for urban, rural, or suburban areas. Traffic sites are generally situated relatively close to main roads or intersections (EU, 2008). Urban background sites are assumed to capture the contribution of all sources near the site without one particularly dominating source. Suburban background stations need to be placed downwind (referring to the main wind direction) of emission sources. Rural background stations should not be influenced by agglomerations or industrial sites closer than 5 km (EU, 2008). The spatial distribution of station types and their altitudes are shown in Figures A1 and A2. UBA PM station coordinates are used as spatial reference for data collocation, that is, pixels from continuous data grids are collocated with the position of these stations if below a distance threshold of 0.01 ° (~0.7 km). Urban industrial stations are not considered in this study, as they are primarily influenced by local emissions from point sources that cannot be adequately represented with the available data. PM10 concentrations were checked for sudden peaks, which could be due to localized events. These are filtered out by eliminating situations, where the PM10 concentration was more than double the mean of the previous and following hours.

2.3.2. DWD Meteorological Data

Air pressure and relative humidity (RH) data were obtained from the German Meteorological Service (DWD) (DWD Climate Data Center [CDC], 2017). Previous studies found a positive correlation of PM10 concentrations and air pressure. Higher air pressure indicates stable synoptic conditions, which favors the accumulation of particles (Li et al., 2015). RH potentially enhances particle numbers by stimulating the formation of aqueous SOA, forming in cloud or aerosol water (Ervens et al., 2011). In addition, RH can influence the relationship between AOD and PM10, as higher levels of humidity lead to hygroscopic particle growth. Moisture on particles increases their diameter, eventually causing a rise in AOD without affecting PM10 measurements (Schwarz et al., 2017; Titos et al., 2014; Stirnberg et al., 2018).

2.3.3. Continentality Factor

The relationship between AOD and PM10 as well as driving factors of PM10 concentrations may vary in different climatic regions (Di et al., 2016). This is accounted for by the inclusion of a dimensionless continentality factor k, which is calculated following the formula by Conrad (1946)

$$k = (1.7 * A/sin(\phi + 10^{\circ})) - 14, \quad (1)$$

where A corresponds to the difference between the hottest and coldest mean monthly temperature and ϕ to the latitude in decimal degree. Here, k was calculated using the mean temperature of July (1980–2010) and the mean temperature of January (1980–2010). Temperature data are provided by the DWD (DWD Climate Data Center [CDC], 2018).

2.3.4. DWD Radolan Precipitation

The influence of precipitation on PM10 concentrations includes wash-out effects and the limitation of movement of particles after precipitation events (Fuzzi et al., 2015; Li et al., 2015; Rost et al., 2009). In this study, data from the Radar Online Adjustment project (RADOLAN) are used. The data set is produced by the DWD and merges radar measurements with rain gauge data, also including orographic correction (Bartels et al., 2004; Weigl, 2017). The data are available with high spatial coverage and are expected to be better suited for the present analysis than the Era-Interim precipitation product, which has some known biases (de Leeuw et al., 2015). In the statistical model, the time since the last precipitation event (hr) and its magnitude (mm/hr) as well as the accumulated precipitation of the last 24 hr (mm) are included. Hourly means of precipitation around each PM station are averaged within a window of edge size 5 km. The chosen window size is smaller than the previous ones, as precipitation effects typically vary on small scales.

2.4. Other Input Data

2.4.1. EEA Emission Database

Annual emissions of NO_x, PM10, SO_2, and NH_3 (in tonnes, based on the year 2008) are included to approximate background pollution levels. The data are gathered by the European Environment Agency (EEA) and described in Theloke et al. (2009). Diffuse air releases from traffic, agricultural, industrial, and residential sources are covered. Strong emitters that fall under the European Pollutant Release and Transfer Register (E-PRTR) are not included in this data set.

2.4.2. Spatiotemporal Information

Seasonality was shown to be an important factor in previous studies (Grange et al., 2018). Here, DOY is used as seasonal proxy. To mirror the seasonal cycle, DOY was converted to a sine curve with +1 representing summer solstice and −1 representing winter solstice (Park et al., 2019; Stolwijk et al., 1999). To further approximate variability in emission strengths based on human activity, day of the week is included.

2.5. Gradient Boosted Regression Trees

2.5.1. Model Specifications

GBRT as implemented in python's scikit-learn module are used (Hastie et al., 2009; Pedregosa et al., 2012). GBRT merge several statistical approaches found in machine learning applications: decision trees (1) and boosting (2) with gradient descent (3).

1. Decision trees use decision nodes to split the predictor space in subsets, which provide the most homogeneous distribution, that is, the subsets' variance is minimized. For each subset, regression trees fit the mean response of the observations that go into the model (Elith et al., 2008).
2. Similar to the RF method, GBRT consist of an ensemble of decision trees. In GBRT models, however, the construction of the ensemble is different, as decision tree regressors are sequentially added to the ensemble (Elith et al., 2008; Hastie et al., 2009; Rybarczyk, 2018). Each new tree that is added to the ensemble boosts its predecessor with the goal to minimize a loss function, and existing trees are not changed when new trees are added. The trees are fitted on a subset of the complete data set, which induces a random component to the model to reduce overfitting (Elith et al., 2008; Hastie et al., 2009).
3. Each new predictor is fitted to the predecessor's previous residual error using gradient descent (Hastie et al., 2009; Elith et al., 2008).

GBRT capture complex interactions and interactive effects between individual predictors (Brokamp et al., 2017; Elith et al., 2008), which nevertheless need to be considered when interpreting the model outcome. An advantage of tree-based methods such as RF or GBRT is that, compared to deep learning methods,

Table 3
List of Hyperparameters and Parameter Grid That is Applied During the Grid Search

Hyperparameter	Description	Parameter grid
loss	Loss function to be optimized	fixed: least squares regression
learning_rate	Contribution of trees to ensemble	[0.08,0.05,0.01]
n_estimators	Number of boosting iterations	[800, 1,800, 2,500, 4,000]
subsample	Fraction of samples used for individual tree	[0.6, 0.8]
min_samples_split	Minimum number of samples used to split a decision node	[6, 10, 14]
min_samples_leaf	Minimum number of samples required for a leaf node	[14, 18, 20]
max_depth	Maximum depth for individual tree, i.e., maximum number of node layers	[6, 10, 14]
max_features	Fraction of features to be considered when searching for best split	[0.5, 0.8]

Note. Hyperparameters determine the architecture of the GBRT model.

model decisions can be retraced and dependencies of the model outcome to input features can be quantified, allowing for conclusions regarding physical processes. This makes GBRT an interpretable machine learning method. GBRT theoretically produce results more effectively than the RF method, as trees are built systematically and less iterations are required (Elith et al., 2008). GBRT have shown to have good predictive power in previous studies (Elith et al., 2008; Fuchs et al., 2018; Just et al., 2018). The general framework of setting up the model is shown in Figure 2 and includes input feature selection, hyperparameter tuning, model training, and model validation.

2.5.2. Feature Selection: Recursive Feature Elimination

Redundant input features lack information and potentially degrade model performance by inducing misleading information, thus weakening the target orientation of the model (Meyer et al., 2018). Therefore, a feature selection is conducted to eliminate redundant predictors. A base model is initialized with a fixed number of trees (500), a fixed learning rate (0.1), and all other hyperparameters on default settings. One feature is then excluded, and the data set is randomly split into a training data set (2/3) and a test data set (1/3) 50-fold. After 50 repetitions, the decrease in model performance on the test data set due to the exclusion of the feature was determined using R^2 as indicator. In the final model, the magnitude of the last precipitation event, the accumulated rainfall of the last 24 hr, and the annual mean emission of SO_2 are excluded, as their exclusion did not lead to a decrease in model performance.

2.5.3. Hyperparameter Tuning, Model Training, and Model Evaluation

Hyperparameters refer to the model architecture, for example, the number of trees or the number of decision nodes. The determination of adequate model hyperparameters is essential to avoid overfitting of the model but at the same time ensures that the model is able to generalize. A grid search is executed, where several parameter combinations are tested. A list of tested parameters is provided in Table 3.

There are trade-offs between *max_depth*, *n_estimators*, and *learning_rate*. A lower *learning_rate* requires a higher number of trees, as the contribution of each tree is decreased. The contribution of each tree is effectively the step size of the gradient descent. Thus, if *learning_rate* is too large, the model cannot adapt to the training data. The number of decision nodes in a tree (*max_depth*) also affects *n_estimators*: Increasing *max_depth* reduces the number of necessary trees but increases the risk of overfitting.

The model penalizes erroneous predictions to improve accuracy. The calculation of the penalty value is determined by the loss function. In this study, a least squares loss function is chosen. The least squares loss function is sensitive to very high (low) values as it strongly penalizes large deviations between predictions and observations and will adjust the model accordingly. This is desirable, as the model should be able to reproduce high concentrations of PM10. Model performance is validated using two kinds of validation strategies. The integrated scikit-learn split function creates a random training (2/3) and test data set (1/3), ensuring a comparable distribution of both data sets. To test the spatial generalizability of the model,

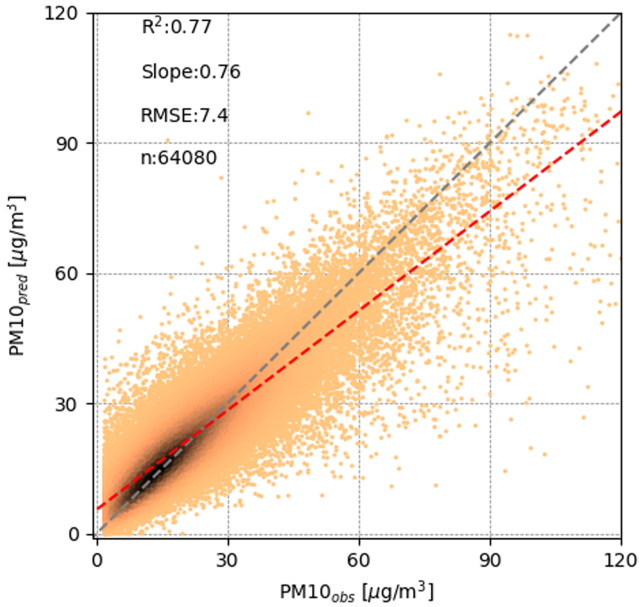

Figure 3. Scatter plot showing the full-year model predictions for hourly PM10 concentrations. Also shown are coefficient of determination (R^2), slope (red dotted line), and root mean square error (RMSE). The color range from black (high) to orange (low) indicates the frequency of occurrence. The relatively high R^2 shows that the model covers the majority of occurring variance. However, an underestimation of higher PM10 concentrations leads to a lower slope.

a leave-location-out (LLO) approach is conducted. Therefore, one third of randomly chosen stations are restrained for validation. If the model performance is lower compared to the random approach, the spatial generalizability of the model is limited (Meyer et al., 2018).

For further analysis, four seasonal models are trained on seasonal subsets of the data (see also Figure 2) using the hyperparameters determined in the full-year grid search. One input feature set representative for all seasonal models and the full-year model is used to ensure their comparability. The only exception is the EEA emission data set, which is only included in the yearly model, since it represents yearly emissions.

2.6. Model Interpretation: Isolation of Feature Contributions

2.6.1. Feature Importance

The relative feature importance reflects the explanatory power a feature provides to the model. Assuming that the model is able to capture physical processes well, the feature importance represents a valuable qualitative measure to determine the relative magnitude of the influence of input features to predicted PM10 concentrations. The feature importance is calculated by repeated permutation of one feature (Strobl et al., 2007).

2.6.2. Partial Dependence

To quantify the influence of input features on the model, the partial dependence (PD) of modeled PM10 concentrations on input features is calculated. PDs express the average effect of one input feature on the modeled PM10 outcome while accounting for average effects of complement input features (Elith et al., 2008; Hastie et al., 2009; Goldstein et al., 2015). The investigated input feature is gridded and the corresponding average PM10 prediction is calculated with respect to complement features, which are varied over their marginalized distributions (e.g., 1st–99th percentile). Thus, PD plots reflect the mean change of average predicted PM10 concentrations based on one input feature. The isolated effects of input features on the model response can be evaluated and put into context regarding their significance to physical and chemical processes determining PM10 concentrations (Grange et al., 2018).

2.6.3. Individual Conditional Expectation

PD plots reflect the mean model response and neglect model heterogeneity. Model responses to single data instances (i.e., one set of input features related to one PM10 observation) can be unwrapped and bundled in one plot using individual conditional expectation (ICE). ICE plots reflect individual predicted responses as a function of one data instance depending on correspondent feature observations. Model responses are computed by keeping the complement features constant while the investigated feature varies, thus creating new

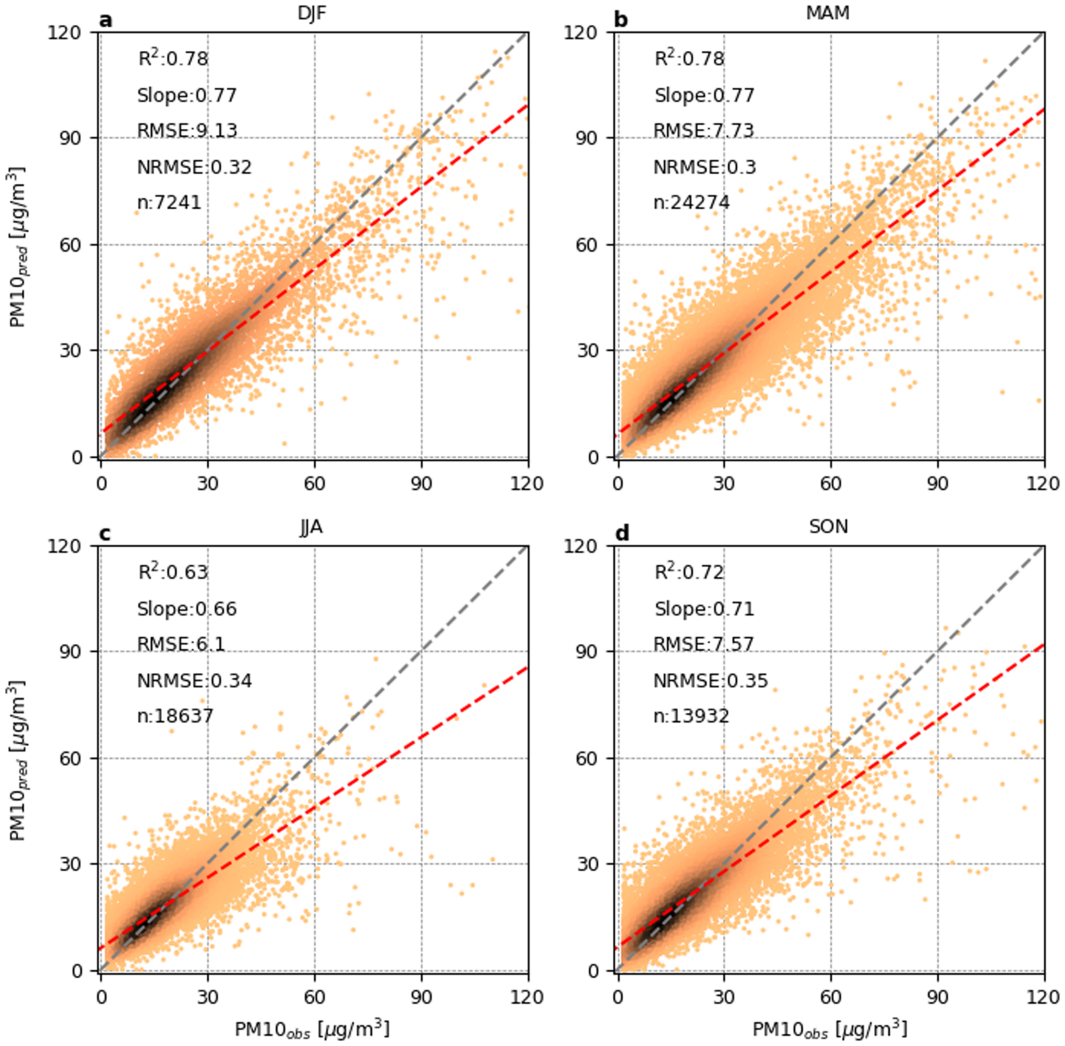

Figure 4. Scatter plot showing seasonal model predictions for hourly PM10 concentrations. Also shown are seasonal R^2, slope (red dotted line), normalized RMSE (NRMSE), and RMSE. The NRMSE was calculated by dividing the RMSE by the mean of the corresponding seasonal subgroup of PM10 observations. Colors as in Figure 3.

data instances and predictions from the model. The average over all ICE lines yields the PD plot, allowing model heterogeneity and mean model response to be depicted simultaneously (Goldstein et al., 2015).

3. Results and Discussion

3.1. Model Performance

The overall model performance is shown in Figure 3, depicting observed PM10 versus predicted PM10 for the validation data. The full-year model explains 77% of the variance. The slope (0.76) shows a slight overestimation of low PM10 concentrations as well as an underestimation of high PM10 concentrations. Presumably, the underestimation is due to processes not captured by the input features, that is, street-scale processes not covered by AOD observations but still influencing PM10 observations, such as increased PM10 emissions due to traffic jams or localized dust resuspension. In addition, the model possibly tends to underestimate higher PM10 observations, because most valid data points are available for medium to low PM10 concentrations. Thus, the model is optimized to best reproduce these observations. This tendency was reduced by the choice of the least squares loss function as describes in chapter 2.5.3 but likely still continues to affect the model accuracy. The model performance is comparable to similar studies, also in its underestimation of PM (Hu et al., 2017; Grange et al., 2018; Stafoggia et al., 2017; Zhang et al., 2018). Tenfold random train/test splits were conducted, resulting in 10 models. Validation of these models revealed very similar performances, which is why only one model is shown and used for subsequent model analysis. Applying the LLO split

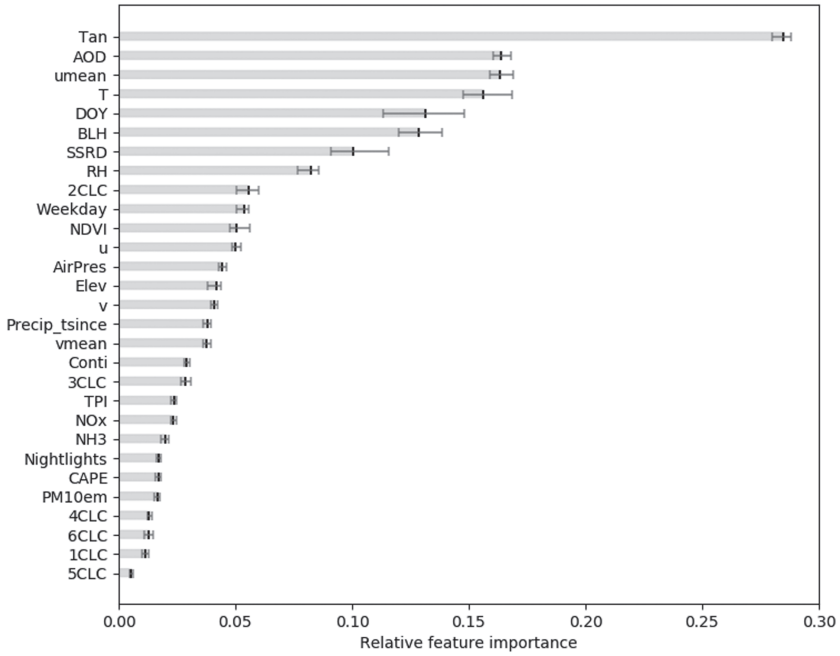

Figure 5. Relative importance of input features based on repeated permutation (see section 2.6). The range represents the standard deviation. Abbreviations correspond to those shown in Table 1.

approach slightly deteriorated model performance. Depending on the restrained stations, R^2 ranged from 0.5 to 0.7. The model is considered as adequate to be used for further investigations. Nevertheless, note that high PM10 values tend to be underestimated by the model (see Figure 3). The spatial distribution of the model skill is shown in Figures B1 and B2.

Model performances vary seasonally (see Figure 4). The model performs best in winter and spring with high R^2 values (0.77) and low NRMSEs (0.32 and 0.3), while R^2 is lowest in summer with an R^2 of 0.63 and a slightly increased NRMSE of 0.34. PM10 concentrations generally show less variance in summer, which reduces the RMSE but possibly provides the model less variance to learn from and deteriorates its skill (i.e., R^2). Obviously, processes governing PM10 concentrations in summer are not as well captured by the model. This will be further addressed in the following chapters.

3.2. Information Content of Input Features

Temperature (anomaly and absolute), AOD, 3-day mean east-west wind component, DOY, and BLH are of high importance to the model (see Figure 5). The importance of the DOY suggests that the model captures the seasonality of PM10 concentrations, which are higher in winter and lower in summer. The relatively high importance of AOD and the good model performance emphasize the suitability of AOD to infer on PM10 concentrations when additional parameters are taken into account. A comparison to similar studies, for example, by Grange et al. (2018) and Park et al. (2019), reveals comparable relative feature importances of AOD, solar radiation (Park et al., 2019), DOY, and BLH (Grange et al., 2018; Park et al., 2019). The high importance of the 3-day mean of the east-west wind component found in this study aligns with the high importance of the back trajectory clusters in the study by Grange et al. (2018). Both parameters reflect regional particle transport. However, there is a discrepancy regarding the importance of wind speed and temperature. Wind speed is considered as instantaneous wind components in this study, which have limited importance. While the importance of absolute air temperature is comparable, the high importance of temperature anomalies found here shows that the approach of splitting temperature information into absolute values and anomalies as pursued in this study provides additional information to the model. Since comparing feature importance values can provide only limited insights into processes behind air pollution patterns, further quantitative analyses are presented in the following chapters using the ICE approach.

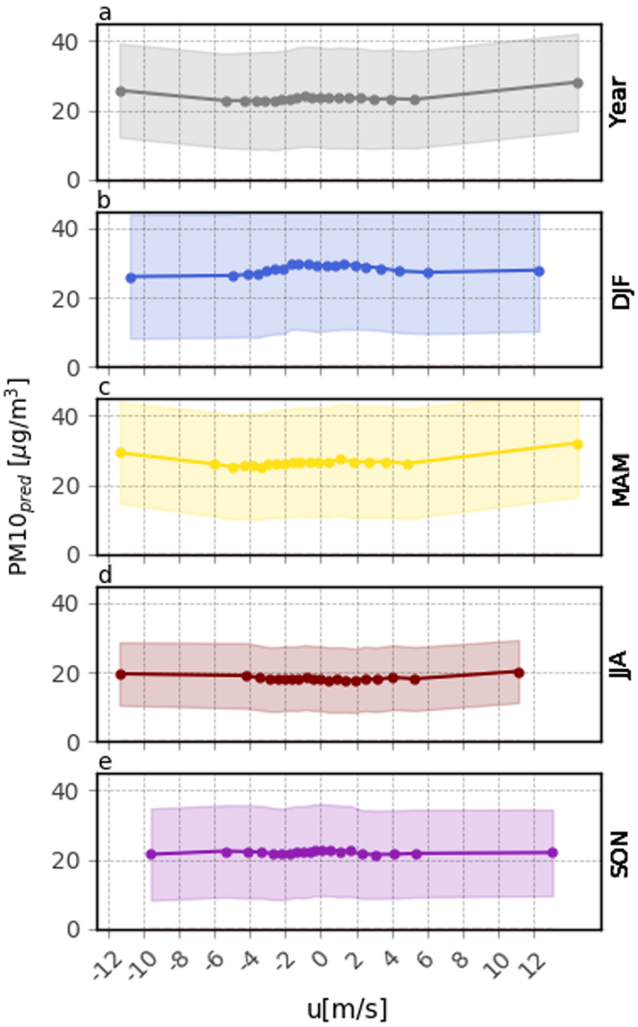

Figure 6. Partial dependence plot showing the mean model response to changes in 3-day mean east-west wind component (m/s) as bold lines for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Shaded areas show the range of the individual conditional expectation (ICE) lines (10th to 90th percentiles). The horizontal distance of dots on the bold line indicates the distribution of valid data points. Negative (positive) values represent dominant inflow of eastern (western) air masses.

3.3. Model Sensitivity

3.3.1. Mesoscale Wind Information

The PD of the 3-day mean east-west wind component shows a consistent pattern throughout all seasons. Positive values (i.e., prevailing western direction of inflow) are associated with reduced concentrations of PM10, whereas a negative east-west wind component (i.e., winds from the east) is associated with increased model PM10 concentrations (Figure 6). For the full-year model the maximum difference in mean PM10 predictions is ~10 $\mu g/m^3$, while in winter, the maximum difference is ~20 $\mu g/m^3$. These numbers agree well with results from van Pinxteren et al. (2019), who quantified the influence of eastern air masses on eastern Germany. They found the contribution of trans-boundary transport from eastern European countries to be 13 $\mu g/m^3$ on average, depending on meteorological conditions. Air masses from continental eastern Europe tend to transport higher amounts of particles, whereas western, more maritime air tends to be cleaner due to precipitation along the trajectories of air masses. Source regions of PM10 include industrial and residential areas in Poland and the Czech Republic with heavy industries or extensive usage of solid fuels for residential heating (Beloconi et al., 2018; Kiesewetter et al., 2015; Reizer & Juda-Rezler, 2016; van Pinxteren et al., 2019). Results by Grange et al. (2018) also show increased values of PM10 for northern and northeastern wind directions, although to a lesser extent. In winter, the effect of particle transportation is strongest, presumably due to increased emissions from domestic heating in eastern Europe (Reizer & Juda-Rezler, 2016;

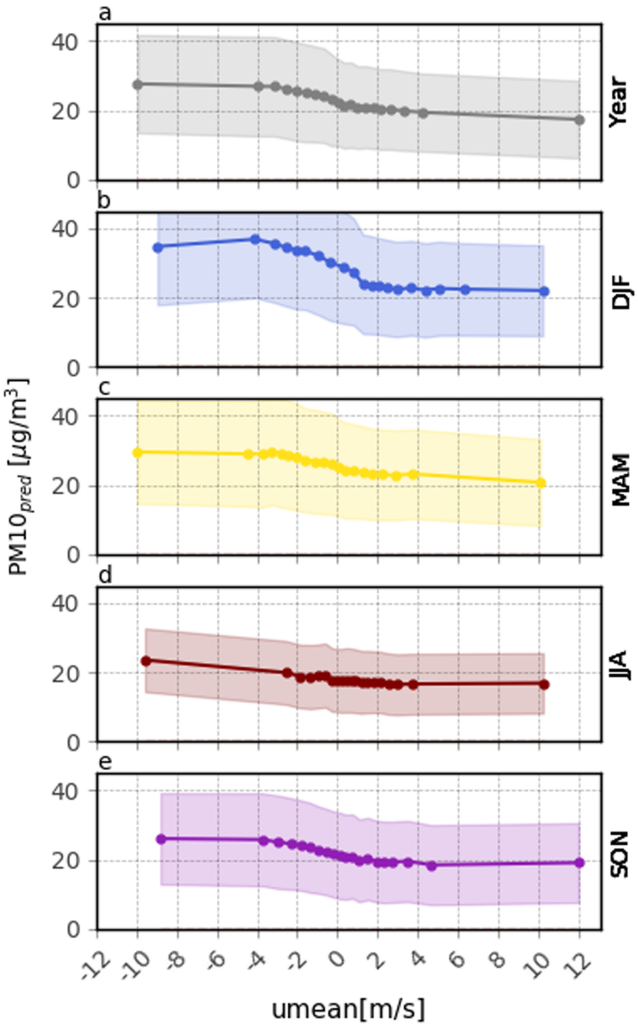

Figure 7. PD plot showing the mean model response to changes in east-west wind component (m/s) as bold lines for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Negative (positive) values represent dominant inflow of eastern (western) air masses. Description as in Figure 6.

van Pinxteren et al., 2019). A slight increase in modeled PM10 at low values of instantaneous east-west wind components is visible (see Figure 7). This could indicate insufficient mixing of the atmosphere, which would lead to an accumulation of particles near ground (Chudnovsky et al., 2013). Overall, instantaneous wind information has little influence on the model, causing the PD to remain relatively constant. The relatively constant PD of instantaneous wind information implies little influence on PM10 predictions. This suggests that wind information needs to be extended to a longer time scale to influence PM predictions.As shown in Grange et al. (2018), wind speed aggregated for a daily period can substantially influence PM10 concentrations, with lower speeds causing higher concentrations. Park et al. (2019) also found the maximum wind speed of previous 3 hr to be of importance for their statistical predictions of PM10. The north-south wind component PDs (instantaneous and 3 days) do not provide clear trends.

3.3.2. BLH and CAPE

The PD of BLH shows that the model is able to reproduce the pattern of decreasing particle concentrations with increasing BLH (Gupta & Christopher, 2009a; Wagner & Schäfer, 2017).The shape of the full-year PD of BLH shown in Figure 8a is similar to that provided by Grange et al. (2018) for observations in Switzerland. They found a reduction of ~8 $\mu g/m^3$ for daily PM10 predictions. A reduction in mean PM10 predictions of ~10 $\mu g/m^3$ for situations with higher BLH can be seen in Figure 8a. This similarity is encouraging and proves the robustness of the modeling approach, since both studies use ERA-Interim BLH in a comparable geographic setting.

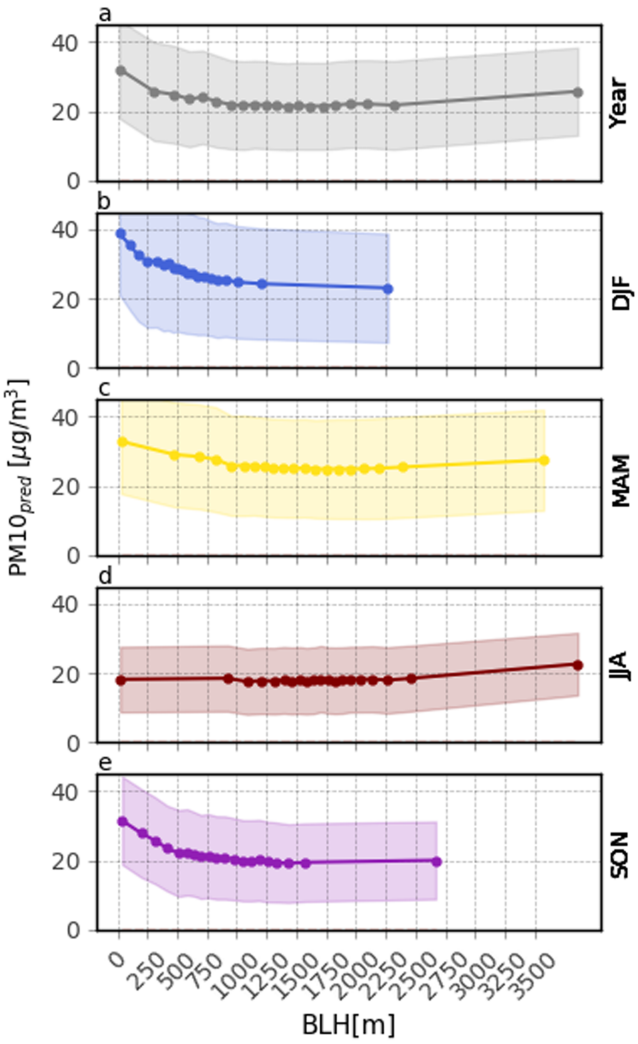

Figure 8. PD plot showing the mean model response to changes in instantaneous BLH (m) for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

For BLH values above ~800 m, the PD remains constant, that is, the influence of the boundary layer on PM concentrations stagnates (Liu et al., 2018). Mean modeled PM10 concentrations increase slightly in conditions with very high BLH (>2,000 m). This pattern could be related to the formation of a deep, convective boundary layer coinciding with high temperatures, enhancing the formation of secondary aerosols (Grange et al., 2018). The abundance of radiation, high temperatures, and precursor gases at excess concentrations would be needed therefore (Fuzzi et al., 2015). Indeed, the pattern is most prominent in summer, when these prerequisites are most likely to be met. The PD of BLH in summer is almost constant, that is, little information is provided to the model. Figure 8d reflects a shift of the frequency of occurrence of data points toward higher BLH: During summer months, medium to high BLHs are more likely to occur due to enhanced convection. As mentioned before, a BLH above 800 m provides only little information to the model, thus reducing its predictive ability. The accumulation effect of a lower BLH is most pronounced in winter with increased mean PM10 predictions of almost 20 $\mu g/m^3$. In addition, PM10 emissions are expected to be higher in wintertime due to combustion of solid fuels for domestic heating. This has been shown for Eastern European countries (Reizer & Juda-Rezler, 2016; van Pinxteren et al., 2019) and likely influences PM10 concentrations in Germany. Thus, with higher locally produced or advected PM10 emissions, the accumulation effect of a low BLH is more distinct in winter than in summer, where reduced emissions are expected and an accumulation of particles is not expected (Wagner & Schäfer, 2017). A dependence of model PM10 on CAPE could not be identified using the PD approach.

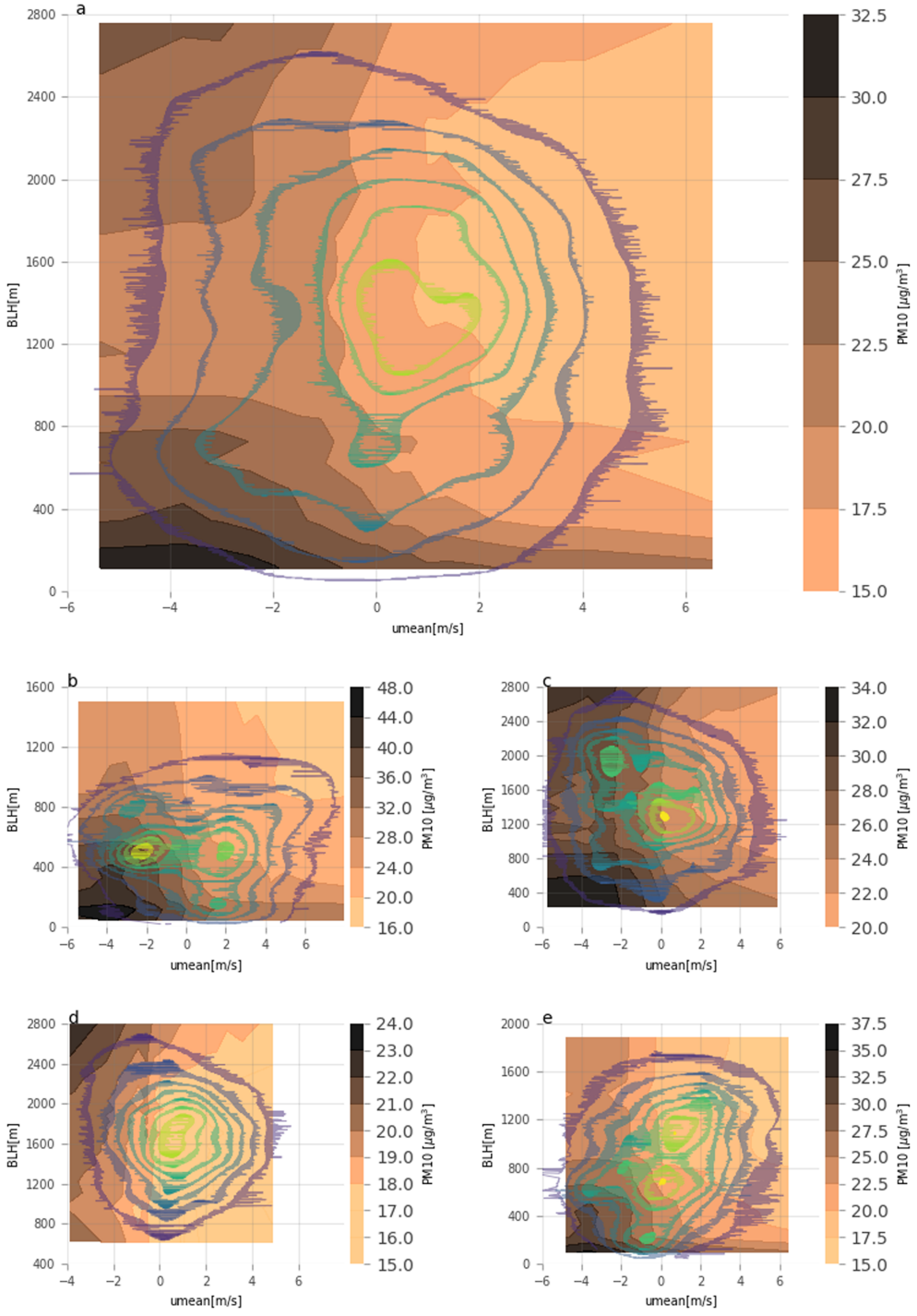

Figure 9. Two-way PD of umean and BLH, full-year model (a) and seasonal models (b–e, DJF, MAM, JJA, and SON). Similar to the previously shown PD, the two-way PD shows the mean modeled PM10 (color coded from light orange to black) to the isolated effects of umean and BLH. On top of the PD isolines, a probability density function (PDF) using Gaussian kernel density estimates is added to indicate the qualitative frequency of occurrence of values from high (yellow) to low (dark blue).

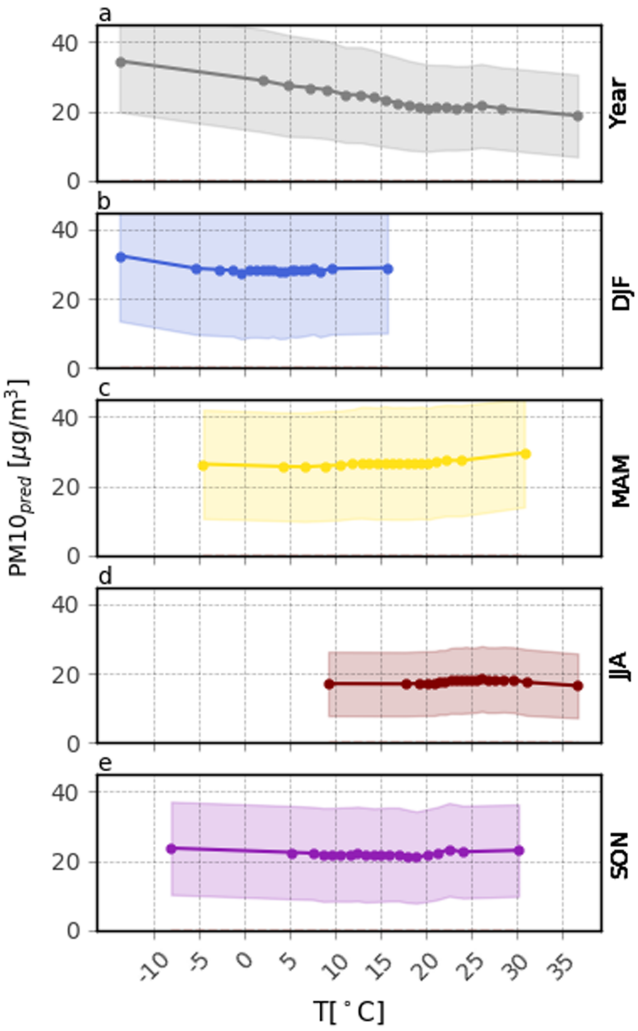

Figure 10. PD plot showing the mean model response to changes in temperature anomalies (K) for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

3.3.3. Two-Way PD: umean and BLH

Depicting the PD of BLH and the 3-day mean east-west wind component simultaneously allows for the quantification of particularly high-polluted situations, considering combined effects of both features. A probability density function (PDF) using Gaussian kernel density estimates is added to provide a qualitative estimate of the frequency of occurrence of PD values. Values outside the PDF estimation as shown in Figure 9 are extrapolated based on the trained model and do not represent observed data.

The two-way PD suggests mean modeled PM10 concentrations double due to changes in BLH and wind flow, referring to the full-year model. Highest mean predictions (~35 $\mu g/m^3$) are modeled when eastern winds coincide with shallow boundary layers, whereas lowest mean predictions occur during medium BLH and western winds (~17 $\mu g/m^3$). Note that due to the tendency of the model to underestimate high PM10 levels, concentrations could be higher in reality. Patterns differ seasonally. In winter, highest mean PM10 predictions surpass 45 $\mu g/m^3$ during shallow BLH conditions and wind flow from the east. In summer, this is not the case. As mentioned in chapter 3.3.2, there is indication of elevated PM10 concentrations during very high BLH (~>2,000 m) conditions, coinciding with eastern wind flow. Highest mean predictions in summer do not surpass ~22 $\mu g/m^3$ (within the limits of the PDF estimation).

3.3.4. Thermal Influence on PM10

In spring, summer, and autumn, positive temperature anomalies cause a marked increase of mean model PM10 predictions (see Figure 10). Presumably, higher temperatures in spring and summer (Figures 10c and 10d) reflect enhanced biogenic activity, as vegetation is generally more active at higher temperatures.

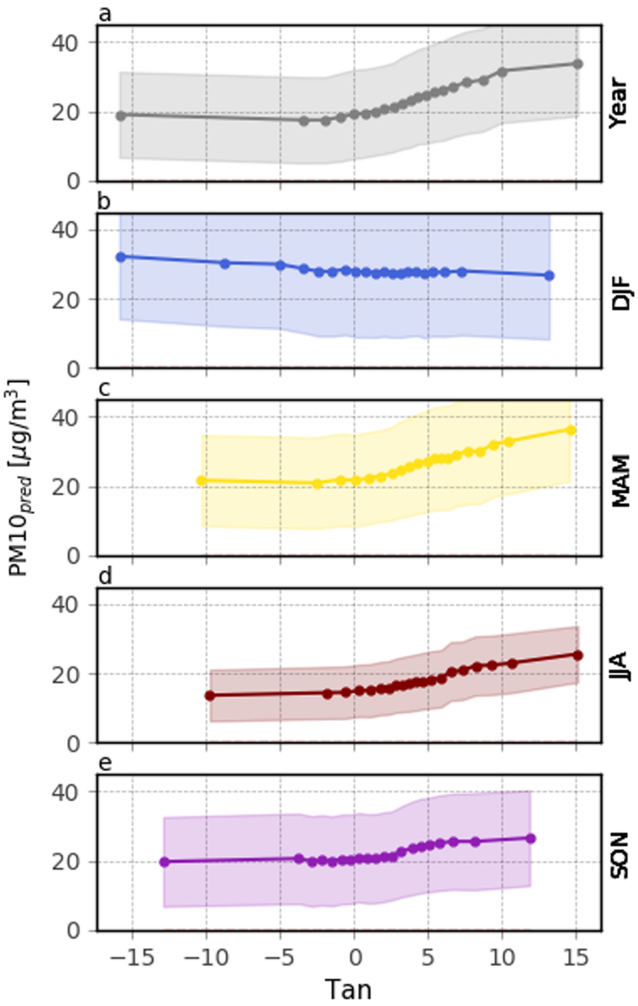

Figure 11. PD plot showing the mean model response to changes in instantaneous temperature (° C) for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

Consequently, emissions of primary particles such as debris or pollen and the emission of biogenic volatile organic compounds (BVOCs) are stimulated (Laothawornkitkul et al., 2009). With higher BVOC emission, an enhancement of secondary SOA formations is leveraged (Churkina et al., 2017; Megaritis et al., 2013). In addition to increased biogenic activities, higher temperatures cause the soil to dry up more quickly, thus increasing dust emissions (Hoffmann & Funk, 2015). In winter, positive anomalies have very little effect on predicted PM10 concentrations (Figure 10b). This supports these hypotheses, since neither increased biogenic activity nor dried-up soils are to be expected in winter. Higher temperatures could reflect increased photochemical oxidation processes, which trigger photochemical reactions leading to new particle formation processes (Birmili & Wiedensohler, 2000; Größ et al., 2018; Wiedensohler, 2000). However, there was no trend in the partial dependence of SSRD, which shows a weak influence of SSRD on PM10 predictions (see Figure C1). Note however that the lacking influence of SSRD could also be due to the fact that this study is confined to cloud-free situations due to the availability of AOD. The variation of SSRD would be higher when including cloudy days, which would likely improve the information content provided to the model by including SSRD. Another possible explanation for increased PM10 concentrations at higher temperatures could be that these situations are associated with stable synoptic conditions (at least in spring, summer, and autumn), causing particles to accumulate in the atmosphere. PD trends for instantaneous temperature (Figure 11) are inverse to those described for temperature anomalies. The full-year model PD shows a decrease in predicted PM10 concentrations for higher temperatures. Presumably, instantaneous temperature reflects the annual cycle of PM10 (similar to DOY), which is why the seasonal PDs show no trends.

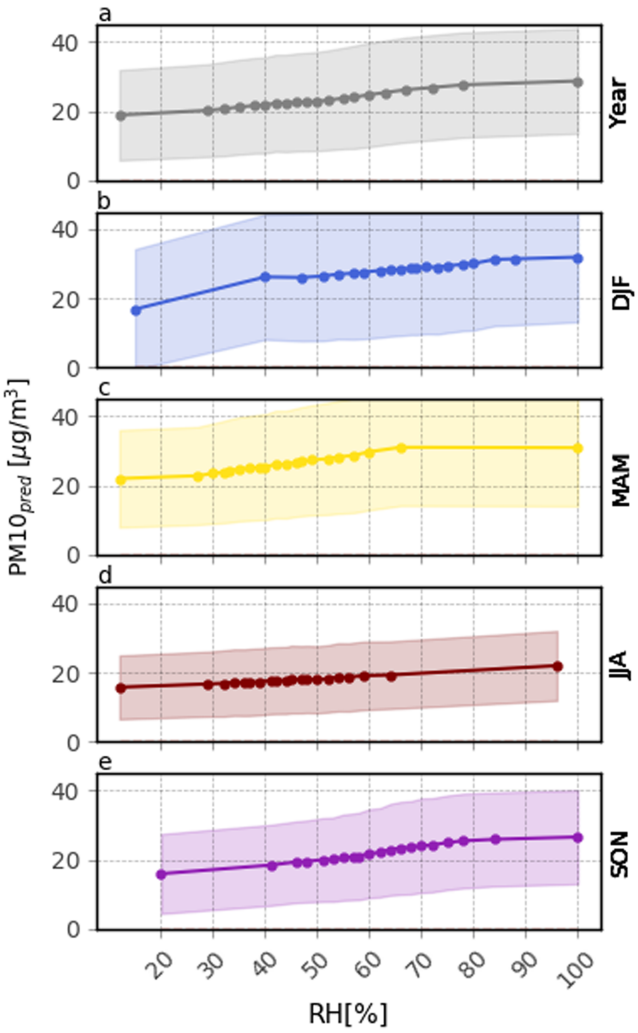

Figure 12. PD plot showing the mean model response to changes in RH (%) for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

PD of air temperature as presented in Grange et al. (2018) reveal a different pattern. In their study, temperatures below freezing are associated with high PM10 concentrations, medium temperature in the range of 0–15 ° C with low PM10 concentrations, and temperature above 15 ° C with high PM10 concentrations. However, when combining the effects of temperature anomalies and temperature presented in Figures 10a and 11a, the emerging pattern would be similar. This suggests that the model presented in this study is able to discriminate between the seasonal component of temperature and the immediate effect of temperature on particle emissions, for example, due to new particle formation (cf. Birmili & Wiedensohler, 2000; Bressi et al., 2013; Größ et al., 2018; Petetin et al., 2014; Wiedensohler, 2000)

3.3.5. RH and Precipitation

Increased RH is associated with higher PM10 predictions (see Figure 12). This is likely related to an increase in AOD due to aerosol swelling in humid conditions (Crumeyrolle et al., 2014; Wang & Christopher, 2003).

Other than increasing AOD, the importance of RH could also be related to a correlation between BLH and RH (see PDF estimate in Figure 13). Higher RH reduces the magnitude of turbulent vertical flux and subsequently reduces the BLH (Adamopoulos et al., 2007; Petäjä et al., 2016), which in turn could increase PM10 predictions. On the other hand, vertical transport of water vapor is impeded by a low BLH, which increases RH and stimulates formation of aqueous secondary aerosols (Liu et al., 2018). To analyze the influences of BLH and RH on predicted PM10, a two-way PD of RH and BLH is calculated (see Figure 13). It shows a

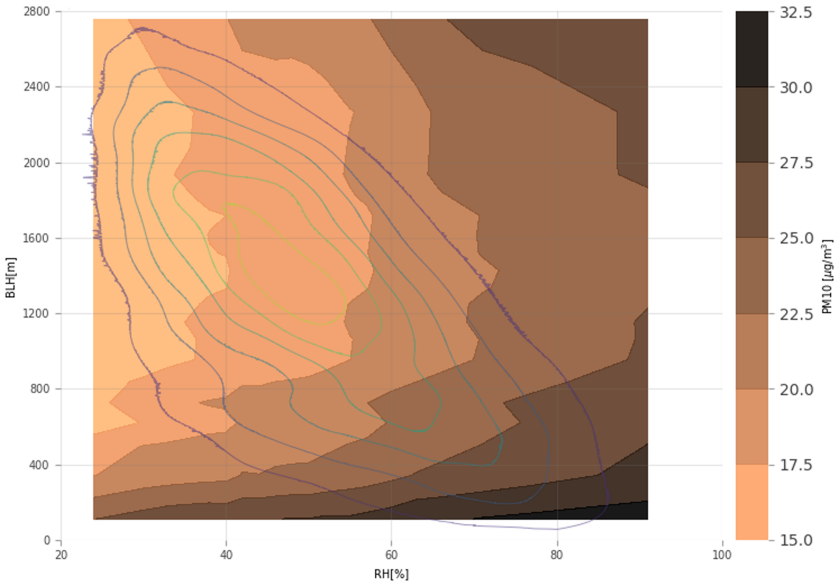

Figure 13. Two-way PD of RH and BLH, full-year model. Description as in Figure 9.

changing pattern with increasing RH, visible in a change of line structure orientation from vertical to more horizontal in the two-way PD plot. While BLH dominates during shallow boundary layer conditions (vertical lines), the influence of RH is more prominent at higher BLH values (horizontal lines). This pattern points to an influence of RH on PM10 predictions, which is decoupled from the BLH, as the influence of BLH above 800 m is marginal. A similar pattern is found for all seasons except for summer. In summer, the model outcome does not show any response to changes in BLH, as there are rarely any BLH values below 800 m (see Figure D1). A study by Belle et al. (2017) conducted in the United States found RH to have positive impact on PM2.5 concentrations during cloud-free conditions due to an increase in sulfate and nitrate masses. However, they found PM2.5 to decrease with increasing RH during cloudy conditions.

The more time passed since the last precipitation event, the higher the PM10 prediction tends to be, reflecting the accumulation of local emissions in the atmosphere. The influence of this effect on the model is not pronounced and stagnates from about 100 hr (see Figure C2). The magnitude of the last precipitation event and accumulated precipitation of last 24 hr were not included in the model due to lacking importance as determined in the feature selection (see chapter 2.5.2). Note however that the low importance of precipitation could be related to the consideration of only cloud-free days in this study. Thus, the immediate effect of rainfall on particles in the atmosphere cannot be investigated. In addition, possible effects of precipitation along the trajectories of advected air masses are not covered.

3.3.6. NDVI, Corine Land Cover, and Spatiotemporal Factors

The NDVI was of minor importance for the prediction of PM10 concentrations. No trends or seasonal differences were found by application of the PD approach. However, with increasing number of pixels in the vicinity of a PM station classified as agricultural areas (2CLC), PM10 concentrations tend to be higher. Likely, this is related to primary emission of dust from arable lands and the application of fertilizers (NH3), which constitute important precursors for secondary particle formation (Hoffmann & Funk, 2015; Wagner et al., 2015). For the other land cover classes, no trends were found using the PD approach. Lower PM10 concentrations are predicted on Saturdays and Sundays, indicating that reduced anthropogenic activity (less traffic, reduced industrial production) has an immediate effect on PM10 concentrations. Increasing altitude slightly reduces the mean PM10 prediction, possibly due to lower population density and more effective pollution dispersion processes (Beloconi et al., 2018; Hu et al., 2014). The PDs of NH3, continentality, PM10 mean annual emissions, and of the TPI do not show a distinct trend. SO_2 was excluded from the model during the feature selection process.

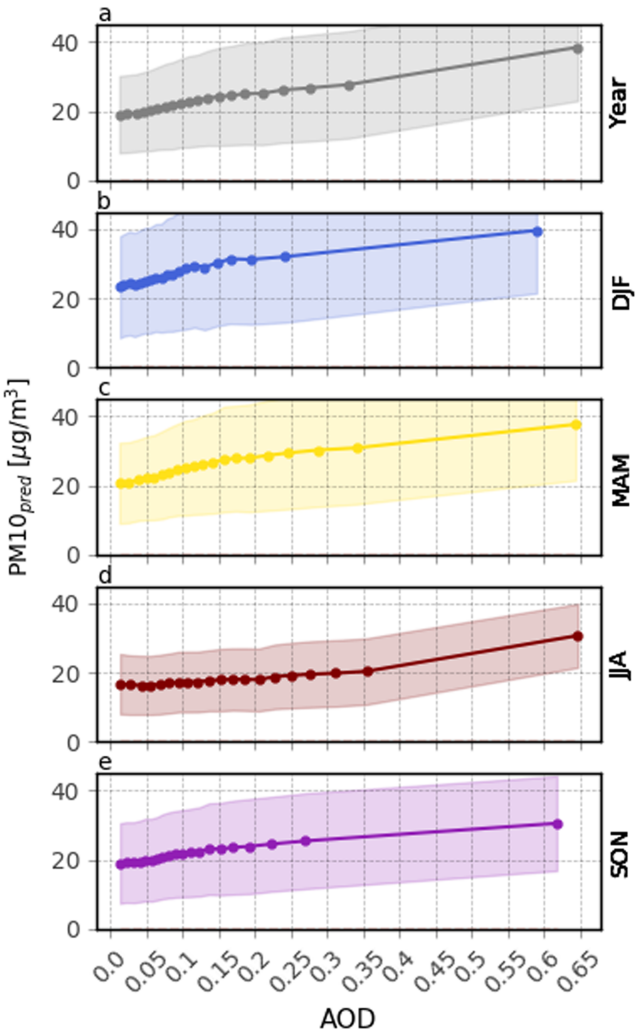

Figure 14. PD plot showing the mean model response to changes in AOD for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

3.4. Determinants of the Relationship Between AOD and PM10

The full-year model and the seasonal models associate increasing AOD with increasing PM10 (see Figure 14). This pattern is less distinct in summer (except for very high AOD) when particles are generally more dispersed within a well-mixed boundary layer, and the AOD is largely determined by particles higher up in the atmosphere, thus weakening the relation between AOD and PM10.

The relationship between AOD and PM10 is not bivariate and can be modified by ambient meteorology (Gupta & Christopher, 2009a; Guo et al., 2009; Sorek-Hamer et al., 2017; Stirnberg et al., 2018). A quantification of this effect is approached here by using the two-way PD method.

The two-way PD of AOD and BLH reveals a dependence of the model on both AOD and BLH (see Figure 15). The importance of interactive effects of these features can be illustrated by the following example: assume an AOD of 0.2 and BLH of 2,000 m versus an AOD of 0.2 and BLH of 200 m. In the latter case, the mean predicted PM10 concentration is ~10 $\mu g/m^3$ higher as the aerosol content determining AOD is closer to the ground and thus more relevant for the PM10 prediction. In other words, a prediction based on AOD (assuming that AOD is largely determined by attenuation in the boundary layer Schäfer et al., 2008) alone would lead to erroneous PM10 predictions, as AOD does not fully capture the particle accumulation effect of a shallow boundary layer (cf. Stirnberg et al., 2018). Similar effects can be observed for the two-way PD of AOD and the 3-day mean east-west wind component (see Figure 15) and for the two-way PD of AOD and

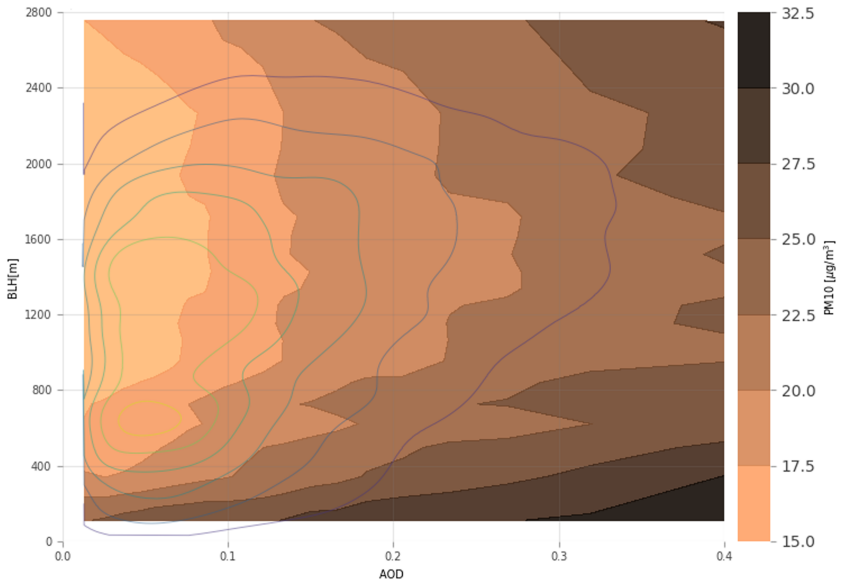

Figure 15. Two-way PD of AOD and BLH, full-year model. Description as in Figure 9.

temperature anomalies (plot not shown). The two-way PD of AOD and BLH shows only a minor seasonal pattern, which is mostly driven by BLH (see Figure D2)

Westerly wind flow (positive umean) leads to substantially lower PM10 predictions when compared to similar AOD values in situations dominated by easterly wind flow (negative umean). The two-way PD suggests this effect to be as large as ~8 $\mu g/m^3$ (see Figure 16). Air masses from the east possibly carry a higher amount of near-ground particles (Beloconi et al., 2018; Bonn et al., 2016; Reizer & Juda-Rezler, 2016), affecting PM10 observations more strongly than AOD. Another reason for the observed effect could be that western air masses carry a relatively large amount of sea salts with a high hygroscopic growth factor. By effectively taking up water, these constituents enhance light scattering, thus increasing AOD without increasing PM10 measurements (Stirnberg et al., 2018; Zieger et al., 2014; Zieger et al., 2013). The seasonality for the two-way PD of AOD and umean is weak (see Figure D3).

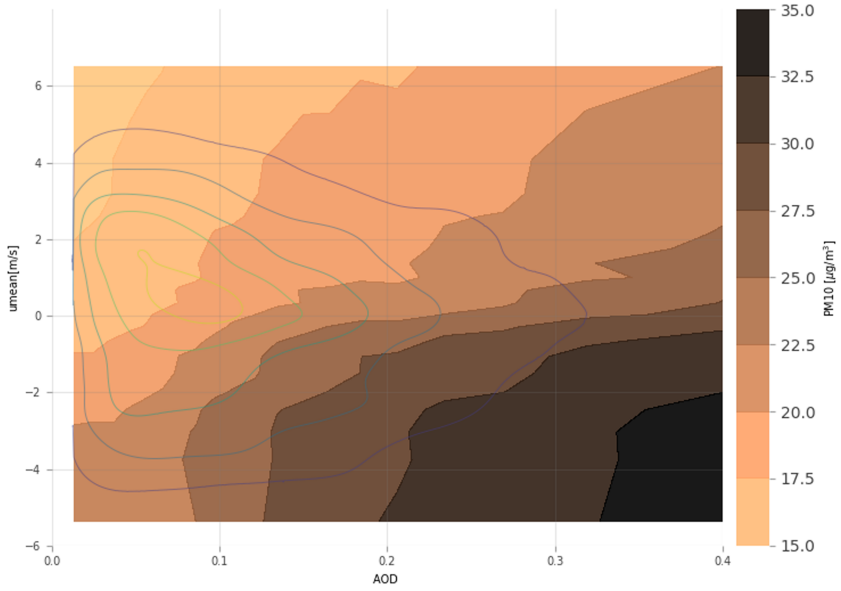

Figure 16. Two-way PD of AOD and umean, full-year model. Description as in Figure 9.

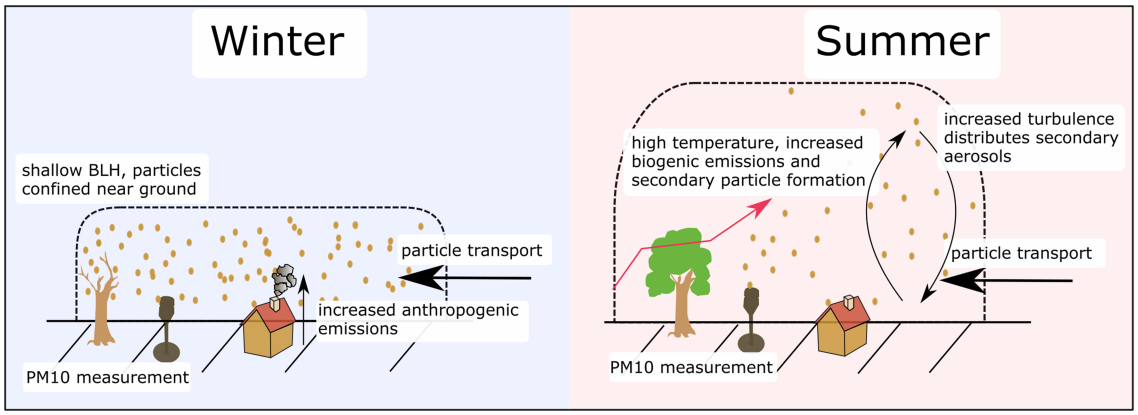

Figure 17. Schematic representation of different processes driving high pollution situations in winter (left) and summer (right).

4. Summary, Conclusions, and Outlook

A machine learning model is used to advance the understanding of drivers of near-ground PM10 and the capability to use satellite AOD to infer on PM10. Parameters pertaining to meteorology, land cover, and satellite-based AOD are considered and related to hourly PM10 concentrations. The model performs well (overall R^2 of 0.77, RMSE = 7.44 $\mu g/m^3$) and provides a basis to assess sensitivities. These allow for the isolation and quantification of effects of ambient conditions on PM10. Overall, the model is more sensitive to meteorological conditions than to land cover parameters. BLH, east-west winds, DOY, temperature, and RH are identified as the important driving factors of PM10 variations. Representing regional particle transport, the 3-day mean of the east-west wind component substantially modifies PM10 concentrations,

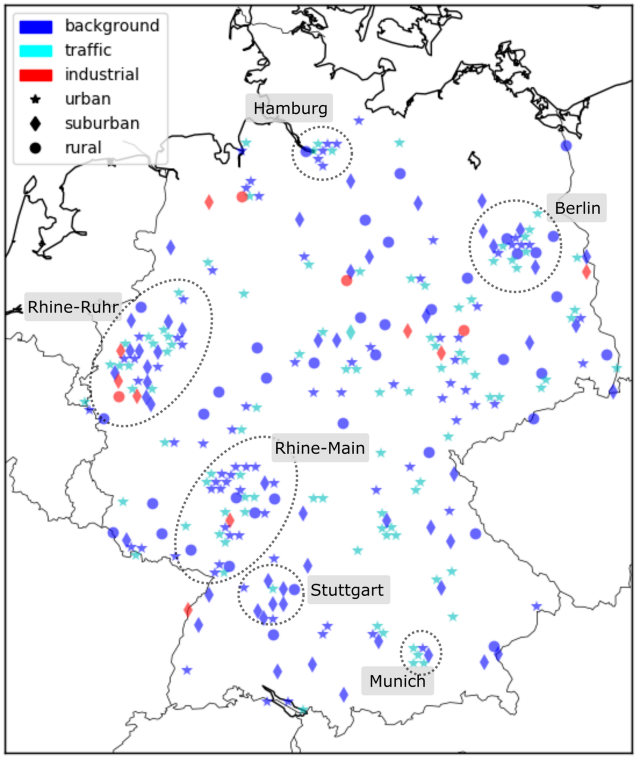

Figure A1. Spatial distribution, type, and representativeness of UBA PM measurement stations in Germany.

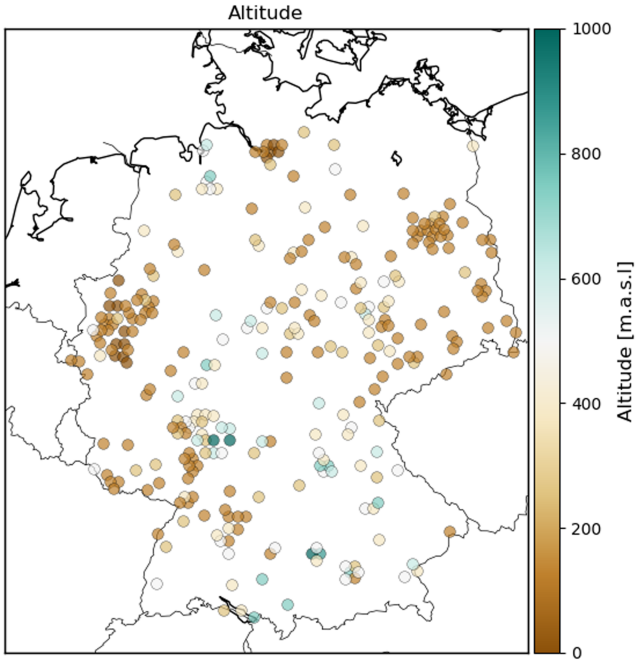

Figure A2. Altitude (m.a.s.l.) of UBA PM measurement stations in Germany.

depending on the direction of inflow. Eastern inflow generally increases PM10 concentrations. Modeled PM10 concentrations were also increased during higher than average temperatures. Possibly, this is due to stimulated vegetation activity, increasing primary particle and precursor gas emissions. The influence of BLH is most prominent at very low (~<500 m) values. However, there is indication that very high BLH values (~<2,500 m) influence PM10 concentrations as well. While the former threshold marks the effects of

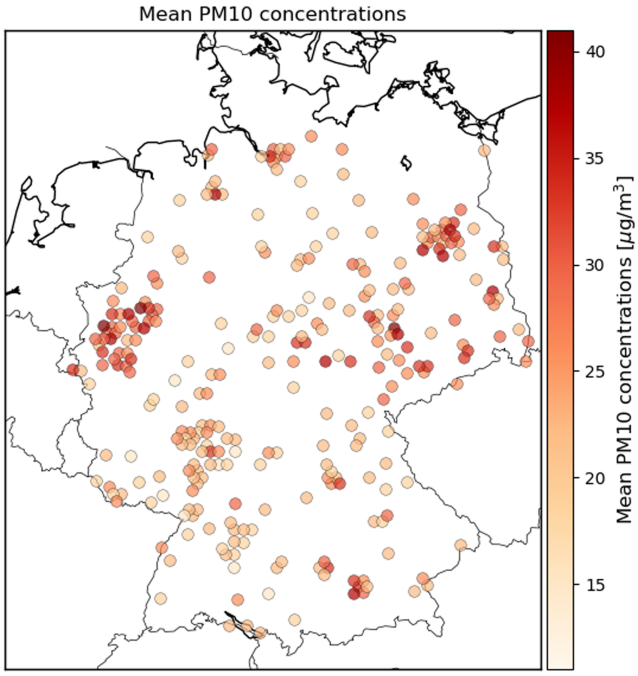

Figure A3. Mean PM10 concentrations for measurement stations used in this study. Time period is the time frame analyzed in this study (2007–2015).

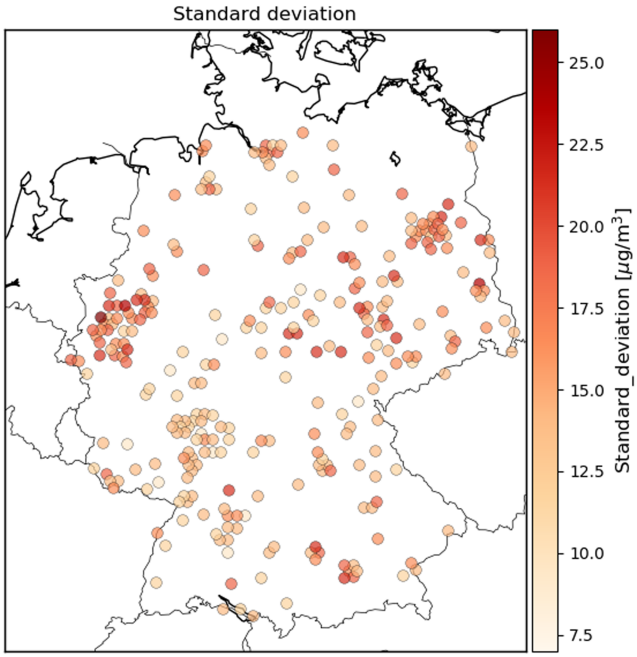

Figure A4. Standard deviation for PM10 concentrations for measurement stations used in this study. Time period is the time frame analyzed in this study (2007–2015).

particle accumulation within a shallow boundary layer, the latter threshold could indicate the formation of a deep boundary layer with stimulated formation of secondary aerosols as suggested by Grange et al. (2018). If BLH is between these thresholds, its explanatory power is limited. In these situations, other processes determine PM10 concentrations. Overall, the model outcome suggests that there are different meteorological boundary conditions that potentially cause elevated PM10 concentrations in winter and summer (Figure 17).

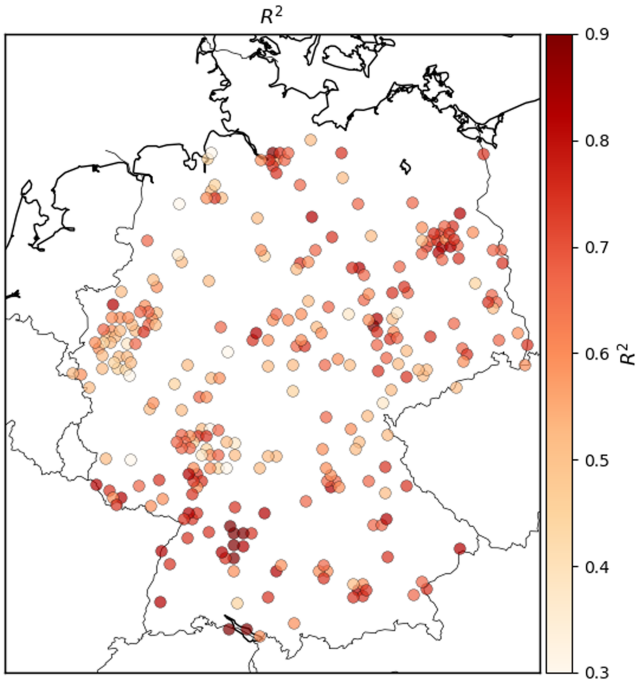

Figure B1. Spatial distribution of R^2.

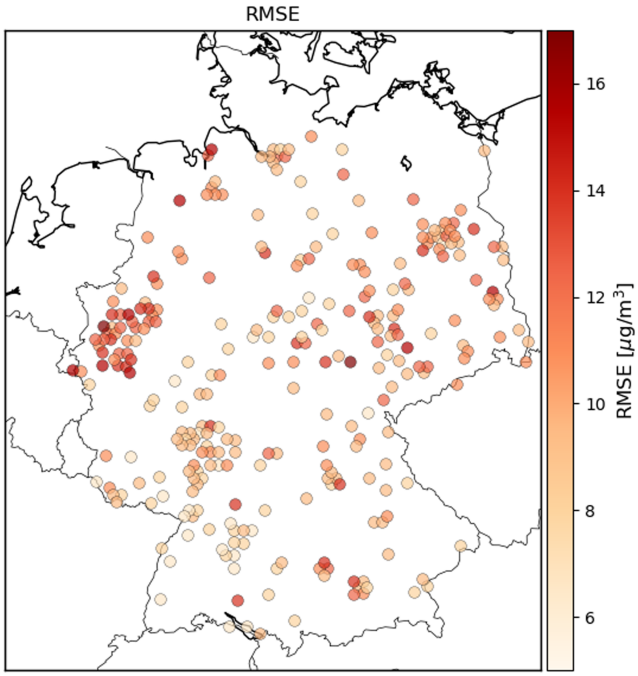

Figure B2. Spatial distribution of the RMSE.

In winter, very shallow boundary layers, coinciding with multiday easterly wind flow cause highest mean PM10 predictions (>30 $\mu g/m^3$). Mean PM10 predictions for these conditions are as high as 40 $\mu g/m^3$ (see Figure 9). This is probably related to higher anthropogenic emissions in winter and frequently low BLH.

In summer, higher temperatures associated with increased formation of secondary aerosols and coinciding with multiday easterly wind flow lead to mean predicted PM10 concentrations >27 $\mu g/m^3$ (figure not shown). High PM10 concentrations in summer appear to be largely uncoupled from changes in BLH (see Figure 8d). The model R^2 decreases in summer, suggesting that the statistical model does not as well resolve these processes. In addition, the relationship between AOD and PM10 is weaker in summer.

Results presented in this study suggest that meteorology plays a substantial role in the development of high pollution situations. This has potential implications for plans toward better air quality in high-polluted areas, as meteorological conditions need to be taken into account, for example, for temporary traffic bans. In addition, there is a need to introduce measures to reduce air pollution on a regional scale. Measures limited to city scales can only decrease pollution levels associated with local emission sources, which can be superimposed by transported particles.

The importance of AOD for the statistical model highlights the suitability of AOD for air quality studies. However, potential implications and limitations for the use of satellite AOD for air quality studies are described. This study has shown that satellite-derived AOD can be used to infer street-level PM10 concentrations, if ambient meteorological conditions are taken into account explicitly. In particular, temperature anomalies, the east-west regional wind component, and BLH modify the relationship between PM10 and AOD. A drawback of including AOD is the restriction to cloud-free situations, which potentially introduces a bias due to nonrandom data gaps (Belle et al., 2017). Depending on the situation and location, both an overestimation or underestimation of PM10 could be the consequence (Belle et al., 2017). In addition, the influence of certain meteorological variables could be underestimated due to important processes under cloudy conditions, which are not covered (Belle et al., 2017; Brokamp et al., 2018). The use of GBRT proved fruitful to understand interconnected processes and the approach presented here can be potentially expanded to other research questions focusing on the understanding multivariate processes. Future efforts will further address the determination of mechanisms leading to high pollution events using machine learning not only for total PM10 concentrations but for individual aerosol species.

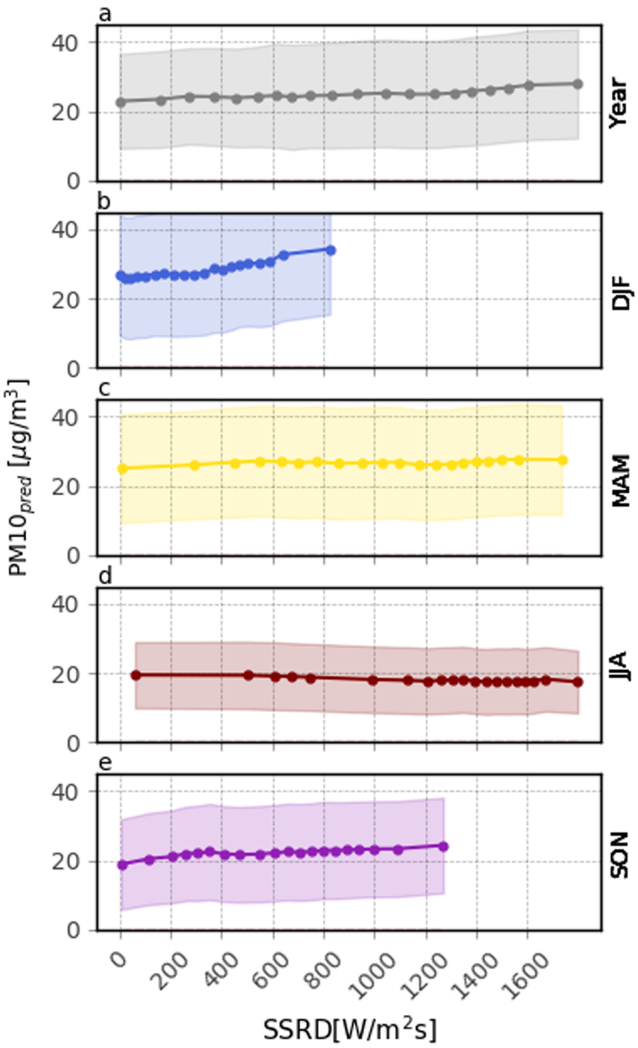

Figure C1. PD plot showing the mean model response to changes in SSRD (W/m^2s) for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

Appendix A: Distribution of UBA PM10 Station Types, Altitude, Mean Concentrations, and Standard Deviations of PM10 Concentrations

The spatial distribution of PM10 measurement stations is shown in Figure A1. Overall, stations are distributed relatively homogeneously over the area of Germany. The number of stations is higher in urban agglomerations. The majority of stations are classified as "urban" or "suburban." "Rural" stations are relatively rare. Most stations are labelled as representative for background conditions by the data provider. Station altitudes (m.a.s.l.) are shown in Figure A2. Altitudes range from 0 to 970 m.a.s.l.

Furthermore, mean PM10 concentrations and standard deviations of all stations for the study period 2007–2015 are shown in Figures A3 and A4, respectively. Mean concentrations are highest in the Rhine-Ruhr area (north-west) and other urban areas such as Munich to the south, Berlin to the northeast, and Hamburg to the north.

Appendix B: Spatial Distribution of Model Skill

Figures B1 and B2 show the spatial distribution of the coefficient of determination and the RMSE, respectively. In the southwest and the northeast, R^2 tends to be higher, and RMSE tends to be lower. In the

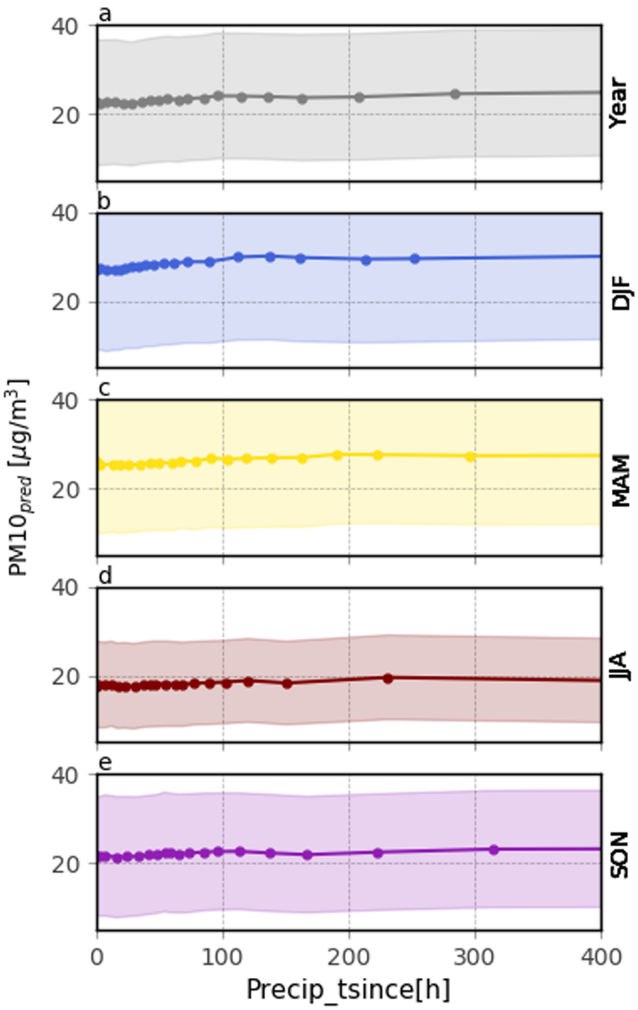

Figure C2. PD plot showing the mean model response to changes in time since last precipitation (hr) for the full-year model (a) and each season separately (b–e, DJF, MAM, JJA, and SON). Description as in Figure 6.

Rhine-Ruhr region (northwest), performance seems to be generally worse. These stations generally have high mean PM10 concentrations (see Figure A3). However, it appears that other urban areas, which also have high mean PM10 concentrations can be modeled quite well (e.g., Berlin in the northeast or Hamburg in the north).

Appendix C: Further Individual Conditional Expectation Plots

Solar radiation (Figure C1) and time since the last precipitation (Figure C2) were analyzed using the ICE method as described in chapter 2.6.3. Both input features show only minor influence on mean PM10 predictions. The ICE functions for time since last precipitation show particularly large variability of model responses of the individual data instances (shown by the shaded areas beneath the bold lines), indicating strong interactions with other features.

Appendix D: Seasonal Two-Way Partial Dependence Plots

In addition to Figure 13, which shows the full-year model two-way partial dependence of RH and BLH, Figure D1 additionally depicts the two-way partial dependence of the seasonal models. Similarly, Figures D2 and D3 show the seasonal two-way partial dependence of AOD and BLH and AOD and umean, respectively.

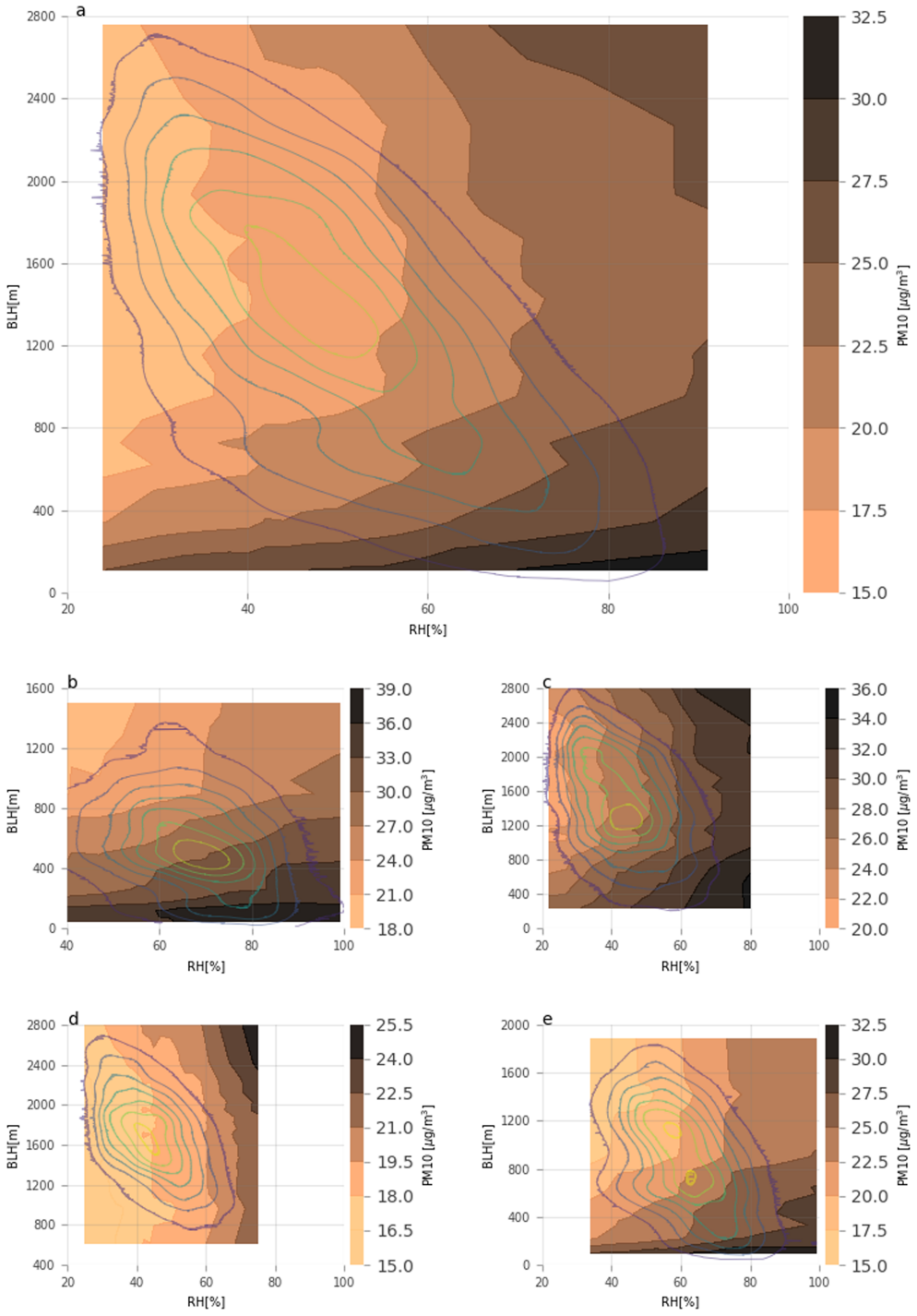

Figure D1. Two-way PD of RH and BLH, full-year model (a) and seasonal models (b–e, DJF, MAM, JJA, and SON). Description as in Figure 9.

Acronyms

AOD aerosol optical depth
BLH boundary layer height
BRF bidirectional reflectance factor
CAPE convective available potential energy
CLC Corine land cover
DEM digital elevation model

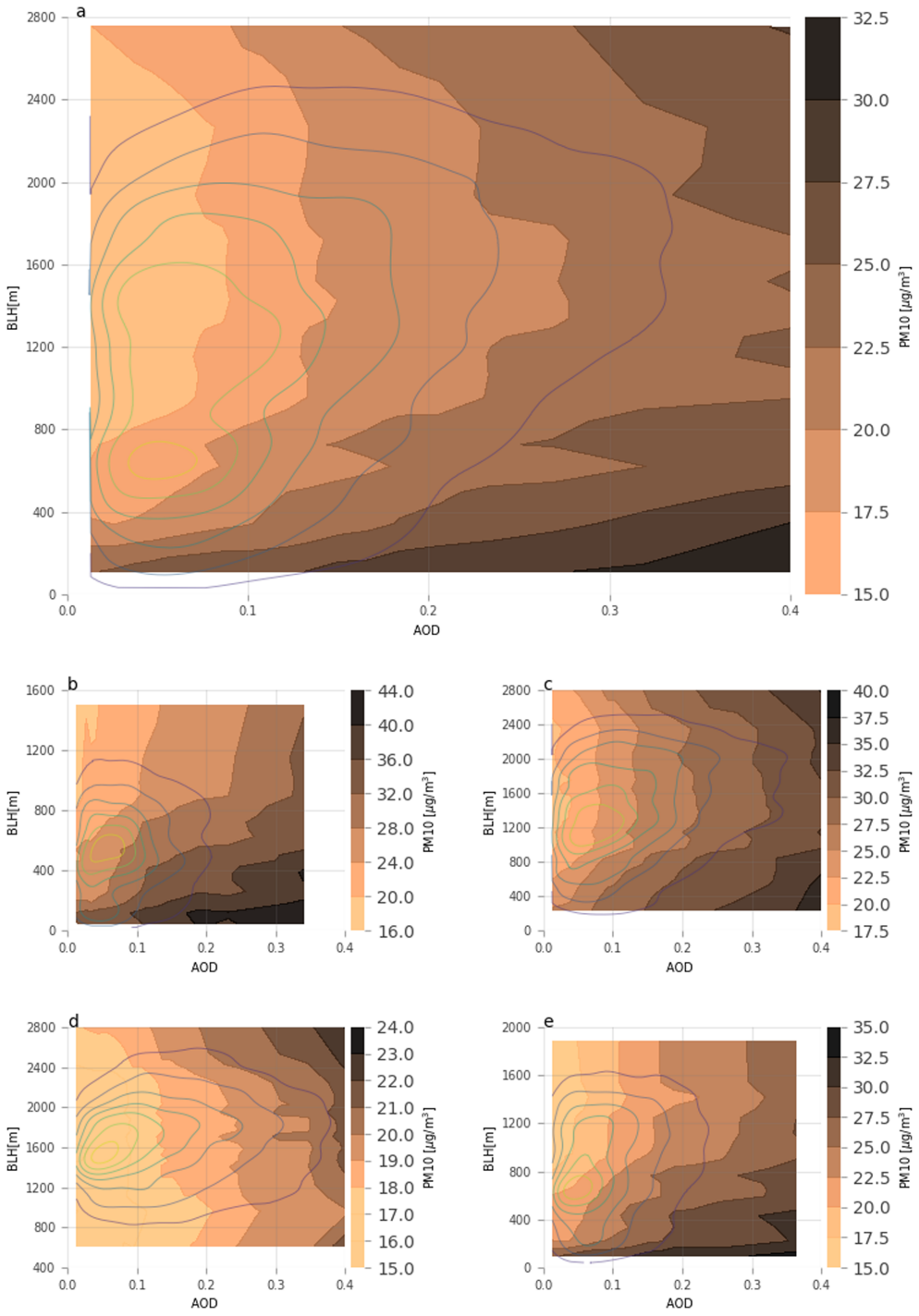

Figure D2. Two-way PD of AOD and BLH, full-year model (a) and seasonal models (b–e, DJF, MAM, JJA, and SON). Description as in Figure 9.

DOY day of year
DWD German Meteorological Service
EEA European Environmental Agency
ECMWF Centre for Medium-Range Weather Forecasts
GBRT gradient boosted regression trees
LLO leave location out
MAIAC multi-angle implementation of atmospheric correction

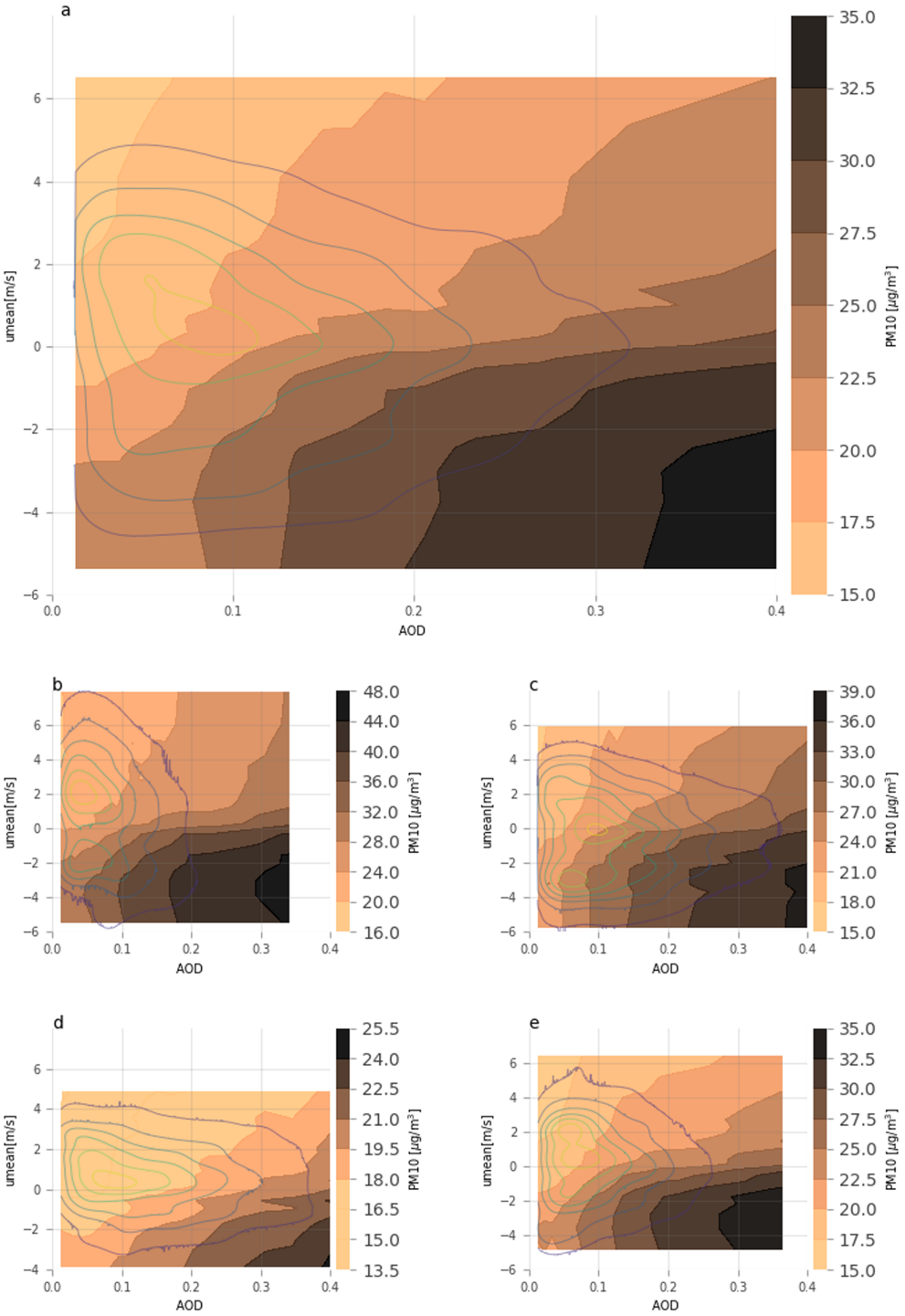

Figure D3. Two-way PD of AOD and umean, full-year model (a) and seasonal models (b–e, DJF, MAM, JJA, and SON). Description as in Figure 9.

MODIS moderate resolution imaging spectroradiometer
NDVI normalized difference vegetation index
PDF probability density function
PM particulate matter
RADOLAN Radar-Online-Aneichung
RF random forest

RH relative humidity
RMSE root mean square error
SOA secondary organic aerosols
SSRD surface solar radiation downwards
TPI topographic position index
UBA German Environmental Agency
VIIRS Visible Infrared Imaging Radiometer Suite
SNPP Suomi National Polar-Orbiting Platform,

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments
RS was supported by the KIT Graduate School for Climate and Environment (GRACE). The authors would like to thank Miae Kim and Frank-Michael Göttsche for helpful discussions. The authors gratefully acknowledge the UBA for providing hourly PM10 measurements and the DWD for the provision of the RADOLAN data set (https://www.dwd.de/DE/leistungen/radolan/radolan.html) and hourly meteorological measurements (https://opendata.dwd.de/climate_environment/CDC/). Furthermore, the EEA is acknowledged for the CORINE land cover (https://land.copernicus.eu/pan-european/corine-land-cover), the EU-DEM (https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-eu-dem), and the EEA emission data base (https://www.eea.europa.eu/data-and-maps/data/european-pollutant-release-and-transfer-register-e-prtr-regulation-art-8-diffuse-air-data#tab-gis-data). ERA-Interim were obtained from the ECMWF website (https://apps.ecmwf.int/datasets/). MAIAC AOD can be accessed via Earth Explorer (https://earthexplorer.usgs.gov/). Nasa Earth at Night data was accessed via the NASA Earth Observatory website (https://earthobservatory.nasa.gov/features/NightLights), and the MODIS NDVI was accessed via the Land Processes Distributed Active Archive Center (LPDAAC, https://lpdaac.usgs.gov/products/mod13q1v006/).

References

Adamopoulos, A. D., Kambezidis, H. D., Kaskaoutis, D. G., & Giavis, G. (2007). A study of aerosol particle sizes in the atmosphere of Athens, Greece, retrieved from solar spectral measurements. *Atmospheric Research*, *86*(3-4), 194–206. https://doi.org/10.1016/j.atmosres.2007.04.003

Andersen, H., Cermak, J., Fuchs, J., Knutti, R., & Lohmann, U. (2017). Understanding the drivers of marine liquid-water cloud occurrence and properties with global observations using neural networks. *Atmospheric Chemistry and Physics*, *17*(15), 9535–9546. https://doi.org/10.5194/acp-17-9535-2017

Ansmann, A., Althausen, D., Wandinger, U., Franke, K., Müller, D., Wagner, F., & Heintzenberg, J. (2000). Vertical profiling of the Indian aerosol plume with six-wavelength lidar during INDOEX: A first case study. *Geophysical Research Letters*, *27*(7), 963–966. https://doi.org/10.1029/1999GL010902

Arvani, B., Pierce, R. B., Lyapustin, A. I., Wang, Y., Ghermandi, G., & Teggi, S. (2016). Seasonal monitoring and estimation of regional aerosol distribution over Po valley, northern Italy, using a high-resolution MAIAC product. *Atmospheric Environment*, *141*(June), 106–121. https://doi.org/10.1016/j.atmosenv.2016.06.037

Bartels, H., Weigl, E., Reich, T., Lang, P., Wagner, A., Kohler, O., & Gerlach, N. (2004). Projekt RADOLAN Routineverfahren zur Online-Aneichung der Radarniederschlagsdaten mit Hilfe von automatischen Bodenniederschlagsstationen (Ombrometer). Deutscher Wetterdienst. Retrieved from https://www.dwd.de/DE/leistungen/radolan/radolan_info/

Belle, J. H., Chang, H. H., Wang, Y., Hu, X., Lyapustin, A., & Liu, Y. (2017). The potential impact of satellite-retrieved cloud parameters on ground-level PM2.5mass and composition. *International Journal of Environmental Research and Public Health*, *14*(10). https://doi.org/10.3390/ijerph14101244

Beloconi, A., Chrysoulakis, N., Lyapustin, A., Utzinger, J., & Vounatsou, P. (2018). Bayesian geostatistical modelling of PM10 and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products. *Environment International*, *121*(August), 57–70. https://doi.org/10.1016/j.envint.2018.08.041

Birmili, W., & Wiedensohler, A. (2000). New particle formation in the continental boundary layer: Meteorological and gas phase parameter influence. *Geophysical Research Letters*, *27*(20), 3325–3328. https://doi.org/10.1029/1999GL011221

Bonn, B., Von Schneidemesser, E., Andrich, D., Quedenau, J., Gerwig, H., Lüdecke, A., et al. (2016). BAERLIN2014—The influence of land surface types on and the horizontal heterogeneity of air pollutant levels in Berlin. *Atmospheric Chemistry and Physics*, *16*(12), 7785–7811. https://doi.org/10.5194/acp-16-7785-2016

Bossard, M., Feranec, J., & Otahel, J. (2000). CORINE land cover technical guide: Addendum 2000 (Tech. Rep. 40). European Environment Agency.

Bressi, M., Sciare, J., Ghersi, V., Bonnaire, N., Nicolas, J. B., Petit, J. E., et al. (2013). A one-year comprehensive chemical characterisation of fine aerosol (PM2.5) at urban, suburban and rural background sites in the region of Paris (France). *Atmospheric Chemistry and Physics*, *13*(15), 7825–7844. https://doi.org/10.5194/acp-13-7825-2013

Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental Science & Technology*, *52*, 4173–4179. https://doi.org/10.1021/acs.est.7b05381

Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., & Ryan, P. (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment*, *151*, 1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066

Bundesministerium der Justiz und für Verbraucherschutz (2010). Neununddreißigste Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Verordnung über Luftqualitätsstandards und Emissionshöchstmengen - 39. BImSchV). Anlage 1. Bundesministerium der Justiz und für Verbraucherschutz. Retrieved from https://www.gesetze-im-internet.de/bimschv_39/anlage_1.html

Cermak, J., & Knutti, R. (2009). Beijing Olympics as an aerosol field experiment. *Geophysical Research Letters*, *36*, L10806. https://doi.org/10.1029/2009GL038572

Chafe, Z., Brauer, M., Héroux, M.-E., Klimont, Z., Lanki, T., Salonen, R. O., & Smith, K. R. (2015). Residential heating with wood and coal: Health impacts and policy options in Europe and North America. Copenhagen: World Health Organization.

Chen, G. L., Guang, J., Li, Y., Che, Y. H., & Gong, S. Q. (2018). Retrieval of atmospheric particulate matter using satellite data over central and eastern China. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-3*, 147–153. https://doi.org/10.5194/isprs-archives-XLII-3-147-2018

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., et al. (2018). A machine learning method to estimate PM2.5concentrations across China with remote sensing, meteorological and land use information. *Science of the Total Environment*, *636*, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251

Chudnovsky, A., Koutrakis, P., Kloog, I., Melly, S., Nordio, F., Lyapustin, A., et al. (2014). Fine particulate matter predictions using high resolution aerosol optical depth (AOD) retrievals. *Atmospheric Environment*, *89*, 189–198. https://doi.org/10.1016/j.atmosenv.2014.02.019

Chudnovsky, A., Tang, C., Lyapustin, A., Wang, Y., Schwartz, J., & Koutrakis, P. (2013). A critical assessment of high-resolution aerosol optical depth retrievals for fine particulate matter predictions. *Atmospheric Chemistry and Physics*, *13*(21), 10,907–10,917. https://doi.org/10.5194/acp-13-10907-2013

Churkina, G., Kuik, F., Bonn, B., Lauer, A., Grote, R., Tomiak, K., & Butler, T. M. (2017). Effect of VOC emissions from vegetation on air quality in Berlin during a heatwave. *Environmental Science & Technology*, *51*, 6120–6130. https://doi.org/10.1021/acs.est.6b06514

Conrad, V. (1946). *Methods in Climatology* (pp. 296–300). Cambridge, Massachusetts: Harvard University Press.

Crumeyrolle, S., Chen, G., Ziemba, L., Beyersdorf, A., Thornhill, L., Winstead, E., et al. (2014). Factors that influence surface PM2.5values inferred from satellite observations: Perspective gained for the US Baltimore-Washington metropolitan area during DISCOVER-AQ. *Atmospheric Chemistry and Physics*, *14*(4), 2139–2153.

DWD Climate Data Center (CDC) (2017). Historical hourly station observations of 2m air temperature and humidity, version v004, 2016. (Tech. Rep.). Offenbach. Retrieved from ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/hourly/air_temperature/historical/.

DWD Climate Data Center (CDC) (2018). Grids of monthly averaged daily air temperature (2m) over Germany Version v1.0. Offenbach. Retrieved from ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/hourly/air_temperature/historical/.

de Leeuw, J., Methven, J., & Blackburn, M. (2015). Evaluation of ERA-Interim reanalysis precipitation products using England and Wales observations. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 798–806. https://doi.org/10.1002/qj.2395

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental Science & Technology*, *50*(9), 4712–4721. https://doi.org/10.1021/acs.est.5b06121

Dupont, J.-C., Haeffelin, M., Badosa, J., Elias, T., Favez, O., Petit, J. E., et al. (2016). Role of the boundary layer dynamics effects on an extreme air pollution event in Paris. *Atmospheric Environment*, *141*, 571–579. https://doi.org/10.1016/j.atmosenv.2016.06.061

EU (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Vol. 152). European Parliament, European Council. doi: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF

Egli, S., Thies, B., & Bendix, J. (2018). A hybrid approach for fog retrieval based on a combination of satellite and ground truth data. *Remote Sensing*, *10*(4), 628. https://doi.org/10.3390/rs10040628

Elith, J., Leathwick, JR, & Hastie, T. (2008). A working guide to boosted regression trees, *77*, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Ellison, R. B., Greaves, S. P., & Hensher, D. A. (2013). Five years of London's low emission zone: Effects on vehicle fleet composition and air quality. *Transportation Research Part D: Transport and Environment*, *23*, 25–33. https://doi.org/10.1016/j.trd.2013.03.010

Emili, E., Lyapustin, A., Wang, Y., Popp, C., Korkin, S., Zebisch, M., et al. (2011). High spatial resolution aerosol retrieval with MAIAC: Application to mountain regions. *Journal of Geophysical Research*, *116*, D23211. https://doi.org/10.1029/2011JD016297

Emili, E., Popp, C., Wunderle, S., Zebisch, M., & Petitta, M. (2011). Mapping particulate matter in alpine regions with satellite and ground-based measurements: An exploratory study for data assimilation. *Atmospheric Environment*, *45*(26), 4344–4353. https://doi.org/10.1016/j.atmosenv.2011.05.051

Ervens, B., Turpin, B. J., Weber, R. J., Brunswick, N., & Sciences, A. (2011). Physics secondary organic aerosol formation in cloud droplets and aqueous particles (aqSOA): A review of laboratory, field and model studies, *11*, 11,069–11,102. https://doi.org/10.5194/acp-11-11069-2011

Fuchs, J., Cermak, J., & Andersen, H. (2018). Building a cloud in the Southeast Atlantic: Understanding low-cloud controls based on satellite observations with machine learning. *Atmospheric Chemistry and Physics*, *18*, 16,537–16,552. https://doi.org/10.5194/acp-18-16537-2018

Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., et al. (2015). Particulate matter, air quality and climate: lessons learned and future needs. *Atmospheric Chemistry and Physics*, *15*(14), 8217–8299. https://doi.org/10.5194/acp-15-8217-2015

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*, 44–65.

Größ, J., Hamed, A., Sonntag, A., Spindler, G., & Manninen, H. E. (2018). Atmospheric new particle formation at the research station Melpitz, Germany : Connection with gaseous precursors and meteorological parameters. *Atmospheric Chemistry and Physics*, *18*, 1835–1861.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM 10 trend analysis. *Atmospheric Chemistry and Physics*, *18*(9), 6223–6239. https://doi.org/10.5194/acp-18-6223-2018

Guo, J. P., Zhang, X. Y., Che, H. Z., Gong, S. L., An, X., Cao, C. X., et al. (2009). Correlation between PM concentrations and aerosol optical depth in eastern China. *Atmospheric Environment*, *43*(37), 5876–5886. https://doi.org/10.1016/j.atmosenv.2009.08.026

Gupta, P., & Christopher, S. A. (2009a). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *Journal of Geophysical Research*, *114*, D14205. https://doi.org/10.1029/2008JD011496

Gupta, P., & Christopher, S. A. (2009b). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2. A neural network approach. *Journal of Geophysical Research*, *114*, D20205. https://doi.org/10.1029/2008JD011497

Gupta, P., Christopher, S., Wang, J., Gehrig, R., Lee, Y., & Kumar, N. (2006). Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmospheric Environment*, *40*(30), 5880–5892. https://doi.org/10.1016/j.atmosenv.2006.03.016

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*, Springer Series in Statistics, vol. 26. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hoffmann, C., & Funk, R. (2015). Diurnal changes of PM10-emission from arable soils in NE-Germany. *Aeolian Research*, *17*, 117–127. https://doi.org/10.1016/j.aeolia.2015.03.002

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM2.5 concentrations in the conterminous United States using the random forest approach, *51*, 6936–6944. https://doi.org/10.1021/acs.est.7b01210

Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., et al. (2014). Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sensing of Environment*, *140*, 220–232. https://doi.org/10.1016/j.rse.2013.08.032

Just, A. C., De Carli, M. M., Shtein, A., Dorman, M., Lyapustin, A., & Kloog, I. (2018). Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM2.5 in the Northeastern USA. *Remote Sensing*, *10*(5), 803. https://doi.org/10.3390/rs10050803

Kiesewetter, G., Borken-Kleefeld, J., Schöpp, W., Heyes, C., Thunis, P., Bessagnet, B., et al. (2015). Modelling street level PM10 concentrations across Europe: Source apportionment and possible futures. *Atmospheric Chemistry and Physics*, *15*(3), 1539–1553. https://doi.org/10.5194/acp-15-1539-2015

Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J., & Schwartz, J. (2011). Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric Environment*, *45*(35), 6267–6275. https://doi.org/10.1016/j.atmosenv.2011.08.066

Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2012). Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the mid-atlantic states. *Environmental Science & Technology*, *46*(21), 11,913–11,921. https://doi.org/10.1016/j.atmosenv.2011.08.066

Kloog, I., Sorek-Hamer, M., Lyapustin, A., Coull, B., Wang, Y., Just, A. C., et al. (2015). Estimating daily PM 2.5 and PM 10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. *Atmospheric Environment*, *122*(March 2016), 409–416. https://doi.org/10.1016/j.atmosenv.2015.10.004

Knüsel, B., Zumwald, M., Baumberger, C., Hirsch Hadorn, G., Fischer, E. M., Bresch, D. N., & Knutti, R. (2019). Applying big data beyond small problems in climate research. *Nature Climate Change*, *9*(3), 196–202. https://doi.org/10.1038/s41558-019-0404-1

Koren, I., Remer, L. A., Kaufman, Y. J., Rudich, Y., & Martins, J. V. (2007). On the twilight zone between clouds and aerosols. *Geophysical Research Letters*, *34*, L08805. https://doi.org/10.1029/2007GL029253

Laothawornkitkul, J., Taylor, J. E., Paul, N. D., & Hewitt, C. N. (2009). Biogenic volatile organic compounds in the Earth system (vol 183, pg 27, 2009). *New Phytologist*, *184*(1), 276.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, *525*(7569), 367–371. https://doi.org/10.1038/nature15371

Lelieveld, J., Klingmüller, K., Pozzer, A., Pöschl, U., Fnais, M., Daiber, A., & Münzel, T. (2019). Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions. *European Heart Journal*, *40*, 1590–1596. https://doi.org/10.1093/eurheartj/ehz135

Lenschow, P., Abraham, H. J., Kutzner, K., Lutz, M., Preu, J. D., & Reichenbcher, W. (2001). Some ideas about the sources of PM10. *Atmospheric Environment*, *35*, S23–S33.

Li, Y., Chen, Q., Zhao, H., Wang, L., & Tao, R. (2015). Variations in PM10, PM2.5 and PM1.0 in an urban area of the Sichuan Basin and their relation to meteorological factors. *Atmosphere (Basel)*, *6*(1), 150–163. https://doi.org/10.3390/atmos6010150

Li, Z., Guo, J., Ding, A., Liao, H., Liu, J., Sun, Y., et al. (2017). Aerosol and boundary-layer interactions and impact on air quality. *National Science Review*, *4*(6), 810–833. https://doi.org/10.1093/nsr/nwx117

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, *380*(9859), 2224–2260. https://doi.org/10.1016/S0140-6736(12)61766-8

Liu, Q., Jia, X., Quan, J., Li, J., Li, X., Wu, Y., et al. (2018). New positive feedback mechanism between boundary layer meteorology and secondary aerosol formation during severe haze events. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-24366-3

Lyapustin, A., Martonchik, J., Wang, Y., Laszlo, I., & Korkin, S. (2011). Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. *Journal of Geophysical Research*, *116*, D03210. https://doi.org/10.1029/2010JD014985

Lyapustin, A., Wang, Y., Korkin, S., & Huang, D. (2018). MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques Discussions*, *11*, 5741–5765. https://doi.org/10.5194/amt-2018-141

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., et al. (2011a). Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research*, *116*, D03211. https://doi.org/10.1029/2010JD014986

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., et al. (2011b). Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research*, *116*, D03211. https://doi.org/10.1029/2010JD014986

Megaritis, A. G., Fountoukis, C., Charalampidis, P. E., Pilinis, C., & Pandis, S. N. (2013). Response of fine particulate matter concentrations to changes of emissions and temperature in Europe. *Atmospheric Chemistry and Physics*, *13*(6), 3423–3443. https://doi.org/10.5194/acp-13-3423-2013

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Nordio, F., Kloog, I., Coull, B. A., Chudnovsky, A., Grillo, P., Bertazzi, P. A., et al. (2013). Estimating spatio-temporal resolved PM10 aerosol mass concentrations using MODIS satellite data and land use regression over Lombardy, Italy. *Atmospheric Environment*, *74*, 227–236. https://doi.org/10.1016/j.atmosenv.2013.03.043

Park, S., Shin, M., Im, J., Song, C.-k., Choi, M., Kim, J., et al. (2019). Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea. *Atmospheric Chemistry and Physics*, *19*(2), 1097–1113. https://doi.org/10.5194/acp-19-1097-2019

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: Machine Learning in Python. Retrieved from http://arxiv.org/abs/1201.0490

Petäjä, T., Järvi, L., Kerminen, V.-M., Ding, A. J., Sun, J. N., Nie, W., et al. (2016). Enhanced air pollution via aerosol-boundary layer feedback in China. *Scientific Reports*, *6*(1), 18998. https://doi.org/10.1038/srep18998

Petetin, H., Beekmann, M., Sciare, J., Bressi, M., Rosso, A., Sanchez, O., & Ghersi, V. (2014). A novel model evaluation approach focusing on local and advected contributions to urban PM2.5 levels—Application to Paris, France. *Geoscientific Model Development*, *7*(4), 1483–1505. https://doi.org/10.5194/gmd-7-1483-2014

Petit, J. E., Favez, O., Sciare, J., Canonaco, F., Croteau, P., Močnik, G., et al. (2014). Submicron aerosol source apportionment of wintertime pollution in Paris, France by double positive matrix factorization (PMF2) using an aerosol chemical speciation monitor (ACSM) and a multi-wavelength aethalometer. *Atmospheric Chemistry and Physics*, *14*(24), 13,773–13,787. https://doi.org/10.5194/acp-14-13773-2014

Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama*, *287*, 1132–1141. https://doi.org/10.1001/jama.287.9.1132

Qadir, R. M., Abbaszade, G., Schnelle-Kreis, J., Chow, J. C., & Zimmermann, R. (2013). Concentrations and source contributions of particulate organic matter before and after implementation of a low emission zone in Munich, Germany. *Environmental Pollution*, *175*(2), 158–167. https://doi.org/10.1016/j.envpol.2013.01.002

Reizer, M., & Juda-Rezler, K. (2016). Explaining the high PM10 concentrations observed in Polish urban areas. *Air Quality, Atmosphere and Health*, *9*(5), 517–531. https://doi.org/10.1007/s11869-015-0358-z

Román, M. O., Wang, Z., Sun, Q., Kalb, V., Miller, S. D., Molthan, A., et al. (2018). NASA's Black Marble nighttime lights product suite. *Remote Sensing of Environment*, *210*(November 2017), 113–143. https://doi.org/10.1016/j.rse.2018.03.017

Rost, J., Holst, T., Sahn, E., Klingner, M., Anke, K., Ahrens, D., & Mayer, H. (2009). Variability of PM10 concentrations dependent on meteorological conditions. *International Journal of Environment and Pollution*, *36*(March 2014), 3–18. https://doi.org/10.1504/IJEP.2009.021813

Rybarczyk, Y. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, *8*(12), 2570. https://doi.org/10.3390/app8122570

Schäfer, K., Harbusch, A., Emeis, S., Koepke, P., & Wiegner, M. (2008). Correlation of aerosol mass near the ground with aerosol optical depth during two seasons in Munich. *Atmospheric Environment*, *42*(18), 4036–4046. https://doi.org/10.1016/j.atmosenv.2008.01.060

Schäfer, K., Wagner, P., Emeis, S., Jahn, C., Münkel, C., Suppan, P., et al. (2012). Mixing layer height and air pollution levels in urban area. *Proceedings of SPIE*, *8534*, 1–10. https://doi.org/10.1117/12.974328

Schwarz, K., Cermak, J., Fuchs, J., & Andersen, H. (2017). Mapping the twilight zone—What we are missing between clouds and aerosols. *Remote Sensing*, *9*(6), 1–10. https://doi.org/10.3390/rs9060577

Sorek-Hamer, M., Broday, D. M., Chatfield, R., Esswein, R., Stafoggia, M., Lepeule, J., et al. (2017). Monthly analysis of PM ratio characteristics and its relation to AOD. *Journal of the Air & Waste Management Association*, *67*(1), 27–38. https://doi.org/10.1080/10962247.2016.1208121

Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., et al. (2017). Estimation of daily PM10 concentrations in Italy (2006-2012) using finely resolved satellite data, land use variables and meteorology. *Environment International*, *99*, 234–244. https://doi.org/10.1016/j.envint.2016.11.024

Stirnberg, R., Cermak, J., & Andersen, H. (2018). An analysis of factors influencing the relationship between satellite-derived AOD and ground-level PM10. *Remote Sensing*, *10*, 1353. https://doi.org/10.3390/rs10091353

Stolwijk, A. M., Straatman, H., & Zielhuis, G. A. (1999). Studying seasonality by using sine and cosine functions in regression analysis. *Journal of Epidemiology and Community Health*, *53*(4), 235–238. https://doi.org/10.1136/jech.53.4.235

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25. https://doi.org/10.1186/1471-2105-8-25

TÜVRheinland (2012). Report on the suitability test of the ambient air quality measuring system TEOM 1405-DF ambient particulate monitor with PM10 pre-separator and virtual impactor of the company Thermo Fisher Scientific for the components PM10 and PM2.5: TÜV Rheinland Energie und Umwelt GmbH.

Theloke, J., Thiruchittampalam, B., Orlikova, S., Uzbasich, M., & Gauger, T. (2009). Methodology development for the spatial distribution of the diffuse emissions in Europe.

Titos, G., Jefferson, A., Sheridan, P. J., Andrews, E., Lyamani, H., Alados-Arboledas, L., & Ogren, J. A. (2014). Aerosol light-scattering enhancement due to water uptake during the TCAP campaign. *Atmospheric Chemistry and Physics*, *14*(13), 7031–7043. https://doi.org/10.5194/acp-14-7031-2014

Tucker C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150.

Umweltbundesamt (2004). Qualitätssicherungshandbuch des UBA-Messnetzes. Berlin: Federal Environment Agency. Retrieved from http://www.umweltbundesamt.de/uba-info-daten/daten/mbm

Várnai, T., Marshak, A., & Yang, W. (2013). Multi-satellite aerosol observations in the vicinity of clouds. *Atmospheric Chemistry and Physics*, *13*(8), 3899–3908. https://doi.org/10.5194/acp-13-3899-2013

VDI (2002). VDI-RICHTLINIEN: Minimum requirements for suitability tests of automated ambient air quality measuring systems Point-related measurement methods of gaseous and particulate pollutants. Retrieved from https://www.umweltbundesamt.de/sites/default/files/medien/1/dokumente/vdi_4202_1_de.pdf

van Donkelaar, A., Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives*, *118*(6), 847–855. https://doi.org/10.1289/ehp.0901623

van Pinxteren, D., Mothes, F., Spindler, G., Fomba, K. W., & Herrmann, H. (2019). Trans-boundary PM10: Quantifying impact and sources during winter 2016/17 in eastern Germany. *Atmospheric Environment*, *200*(March 2019), 119–130. https://doi.org/10.1016/j.atmosenv.2018.11.061

Wagner, S., Angenendt, E., Beletskaya, O., & Zeddies, J. (2015). Costs and benefits of ammonia and particulate matter abatement in German agriculture including interactions with greenhouse gas emissions. *Agricultural Systems*, *141*, 58–68. https://doi.org/10.1016/j.agsy.2015.09.003

Wagner, P., & Schäfer, K. (2017). Influence of mixing layer height on air pollutant concentrations in an urban street canyon. *Urban Climate*, *22*(May 2018), 64–79. https://doi.org/10.1016/j.uclim.2015.11.001

Wang, J., & Christopher, S. (2003). Intercomparison between satellite-derived aerosol optical thickness and PM 2.5 mass: Implications for air quality studies. *Geophysical Research Letters*, *30*(21), 2095. https://doi.org/10.1029/2003GL018174

Wang, J., & Martin, S. T. (2007). Satellite characterization of urban aerosols: Importance of including hygroscopicity and mixing state in the retrieval algorithms. *Journal of Geophysical Research*, *112*, D17203. https://doi.org/10.1029/2006JD008078

Weigl, E. (2017). RADOLAN: Radar online adjustment. Radar based quantitative precipitation estimation products. 63067 Retrieved from https://www.dwd.de/DE/leistungen/radolan/radolan_info/radolan_poster_201711_en_pdf.pdf?__blob=publicationFile&v=2

Wichmann, H., Spix, C., Tuch, T., Wölke, G, Peters, A., Heinrich, J., et al. (2000). Daily mortality and fine and ultrafine particles in Erfurt, Germany part I: Role of particle number and particle mass. *Research report (Health Effects Institute)*, *98*(December 2000), 5–86.

Wiedensohler, A. (2000). New particle formation in the continental boundary layer ' meteorological and gas phase parameter. *Geophysical Research Letters*, *27*(20), 3325–3328.

Zhang, X., Chu, Y., Wang, Y., & Zhang, K. (2018). Predicting daily PM2.5concentrations in Texas using high-resolution satellite aerosol optical depth. *Science of the Total Environment*, *631-632*, 904–911. https://doi.org/10.1016/j.scitotenv.2018.02.255

Zheng, C., Zhao, C., Zhu, Y., Wang, Y., Shi, X., Wu, X., et al. (2017). Analysis of influential factors for the relationship between PM2.5 and AOD in Beijing. *Atmospheric chemistry and physics, Discussions*, *17*, 13,473–13,489. https://doi.org/10.5194/acp-2016-1170

Zieger, P., Fierz-Schmidhauser, R., Poulain, L., Müller, T., Birmili, W., Spindler, G., et al. (2014). Influence of water uptake on the aerosol particle light scattering coefficients of the Central European aerosol. *Tellus B: Chemical and Physical Meteorology*, *66*(1), 105462. https://doi.org/10.3402/tellusb.v66.22716

Zieger, P., Fierz-Schmidhauser, R., Weingartner, E., & Baltensperger, U. (2013). Effects of relative humidity on aerosol light scattering: Results from different European sites. *Atmospheric Chemistry and Physics*, *13*(21), 10,609–10,631. https://doi.org/10.5194/acp-13-10609-2013