Dissertation

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Leon Kuhn

Born in: Stuttgart, Germany

Oral examination: 28 January 2025

NitroNet – A deep-learning NO_2 profile retrieval
for the TROPOMI satellite instrument

Referees: Prof. Dr. Thomas Wagner
Prof. Dr. Norbert Frank

Abstract

Nitrogen dioxide (NO_2) is an important air pollutant, monitored globally by satellite instruments from space, in situ measurements at the surface, aircraft or balloons, and ground-based spectroscopic instruments. Satellite instruments, such as TROPOMI, provide NO_2 column densities (i.e. vertical concentration integrals) by spectral analysis of backscattered sunlight. However, the concentration profiles themselves cannot be obtained. This marks a considerable deficit, as they could be pivotal for studies on health effects of air pollution, air chemistry, and improved satellite retrievals.

This thesis presents the model "NitroNet", a new NO_2 profile retrieval based on an artificial neural network that predicts NO_2 profiles from TROPOMI NO_2 column densities and other ancillary variables. The network's training data is obtained from the regional chemistry and transport model WRF-Chem, operated on a central European domain for the month of May 2019. The contents of this thesis include the validation and optimization of the WRF-Chem model, followed by the training, hyperparameter optimization, and validation of the NitroNet neural network, where "validation" refers to the comparison of either model to satellite observations and measurements from in situ and ground-based spectroscopic instruments. Furthermore, the applicability of NitroNet on other spatio-temporal domains is successfully demonstrated.

Zusammenfassung

Stickstoffdioxid (NO_2) ist ein wichtiger Bestandteil von Luftverschmutzung. Die gängigen Messmethoden umfassen Satelliteninstrumente mit globaler Abdeckung, in-situ-Messungen an der Erdoberfläche, an Flugzeugen oder Ballons und bodenbasierte spektroskopische Messungen. Satelliteninstrumente wie z.B. TROPOMI können NO_2-Säulendichten (d.h. vertikal integrierte Konzentrationsprofile) mittels spektraler Analyse rückgestreuten Sonnenlichts ermitteln. Die Konzentrationsprofile selbst sind hingegen nicht verfügbar. Dies ist ein Manko, da NO_2-Konzentrationsprofile hilfreich für die Erforschung der gesundheitlichen Auswirkungen von Luftverschmutzung, der Atmosphärenchemie, sowie verbesserter Satelliten-Retrievals wären.

In dieser Arbeit wird das Modell „NitroNet“ vorgestellt, ein künstliches neuronales Netzwerk, welches NO_2-Konzentrationsprofile aus NO_2-Säulendichten von TROPOMI und weiteren Variablen bestimmen kann. Die nötigen Trainingsdaten wurden mit dem regionalen Chemie- und Transportmodell WRF-Chem für den Monat Mai 2019 in Mitteleuropa erzeugt. Der Inhalt der vorgelegten Arbeit umfasst die Validierung und Optimierung von WRF-Chem, sowie das Training, die Hyperparameter-Optimierung, und die Validierung von NitroNet. Mit „Validierung“ ist hier der Vergleich beider Modelle mit Satellitenbeobachtungen, in-situ-Messungen und spektroskopischen Messungen an der Erdoberfläche gemeint. Darüber hinaus wird die Anwendbarkeit von NitroNet in anderen geographischen Regionen und Jahreszeiten demonstriert.

Table of contents

List of Figures

List of Tables

List of code excerpts

Chapter 1

Introduction

The operation of satellite instruments, designed to retrieve information about the earth system from space, can be considered one of the most impactful extension to the fields of geoscience, atmospheric science, and remote sensing in the late 20^{th} century. The main power of satellites in sun-synchronous orbit lies within their capability to provide measurements with global coverage, usually with a daily overpass. Geostationary satellites, on the other hand, can provide multiple observations per day with regional coverage. Thereby, satellite instruments provide a plethora of data, the likes of which cannot be achieved with Earth-bound instruments.

The first operational passive remote sensing satellites were launched in the 1960s and 1970s, depending on the field of research. A significant milestone were the *Total Ozone Mapping Spectrometer* (TOMS) instruments, on board the Nimbus-7 satellite and other satellites (see McPeters et al., 1993). These first enabled a continuous and accurate monitoring of Earth's ozone layer, and played a seminal role in the investigation of the Antarctic ozone hole (see Stolarski et al., 1986). Throughout the 1990s and 2000s, the succeeding generations of spectroscopic satellite instruments were developed. The *Global Ozone Monitoring Experiment* instrument (GOME, see Burrows et al., 1999) was launched in 1995, and first enabled the detection of weak absorbers due to its high spectral resolution of ~ 0.2 nm and continuous spectral coverage in the ultraviolet (UV), visible, and near infrared (IR) spectral ranges. GOME was followed by the *Scanning Imaging Absorption Spectrometer for Atmospheric Chartography* (SCIAMACHY, see Burrows et al., 1995) in 2002, the *Ozone Monitoring Instrument* (OMI, see Levelt et al., 2006) in 2004, and the GOME-2 series (see Callies et al., 2000) in 2006.

Then, in 2017, the *Tropospheric Monitoring Instrument* (TROPOMI, see Veefkind et al., 2012) was launched on board the Sentinel-5P satellite mission. TROPOMI has a horizontal resolution of up to $3.5\ \text{km} \times 5.5\ \text{km}$, and is equipped with spectrometers in various wavelength ranges, including the visible and near-ultra violet. By spectral analysis of backscattered sunlight daily global maps of various trace gases, including NO_2, O_3, HCHO, SO_2, CO, CH_4,

aerosols, and cloud properties can be retrieved. These data are used to quantify chemical processes in the atmosphere, air pollution, the abundance of greenhouse gases, volcanic activity, emissions of specific trace gases, and for validation and data assimilation of atmospheric models.

Among the most important constituents of air pollution is nitrogen dioxide (NO_2). NO_2 is a short-lived toxic gas, emitted in the form of NO_x ($= NO + NO_2$) from anthropogenic sources, such as combustion in car motors, power plants, and industrial facilities, as well as natural sources, such as lightning, and soil bacteria. The additional abundance of volatile organic compounds from fuel combustion and chemical agents in the presence of sunlight has brought about the phenomenon of *summer smog*, which culminates in large amounts of tropospheric ozone. Summer smog was first recognized in the city of Los Angeles in the 1940s, where anomalously high ozone pollution was observed together with its immediate adverse effects on human health (see e.g. Haagen-Smit, 1952). Although efforts to monitor air pollution and reduce the emission of smog pollutants were successful in many parts of the world, and smog amplitudes in western countries have significantly decreased over the past decades (see Jiang et al., 2022), many regions of the world are still plagued by poor air quality. The global excess mortality from air pollution is estimated on the scale of $\sim$ 8.000.000 lives per year, although other pollutants such as particulate matter yield a major contribution to this number (see Lelieveld et al., 2015). Air pollution therefore remains a topic of high scientific, medical, and political urgency. Consequently, regulatory frameworks on behalf of the policymakers and an extensive measurement curriculum have been established.

Yet, our picture of the NO_2 distributions within our atmosphere remains incomplete: satellite instruments such as TROPOMI cannot provide vertical concentration profiles, but *column densities*, meaning the total amount of NO_2 along the vertical axis. These are augmented by in situ measurements at the surface, although so sparsely, that many regions of the Earth are not covered by them at all. Approaches to measuring NO_2 with vertical resolution do exist, e.g. with LIDAR, sondes, balloons, or aircraft, but are also operated too sparsely to yield meaningful insight into extended domains (see e.g. Sluis et al., 2010; Bourgeois et al., 2022; Lange et al., 2023; Riess et al., 2023; Volten et al., 2009; Berkhout et al., 2018; Su et al., 2021). Methods such as *cloud slicing* allow to derive NO_2 profiles from satellite observations with different atmospheric penetration depths over optically thick clouds, although at relatively low resolutions (e.g. seasonal means with $1° \times 1°$ horizontal resolution and 5 tropospheric layers, see Marais et al., 2021). Moreover, such profiles are also affected by systematic dependencies of both the NO_2 profiles and cloud altitudes on the meteorological conditions. A potential remedy for the lack of high-resolution NO_2 profile data with dense spatial coverage are *regional chemistry and transport* (RCT) models, such as the *Weather Research and Forecasting model coupled with Chemistry* (WRF-Chem, see Grell et al., 2005), which are used

to simulate trace gas distributions in the atmosphere numerically. RCT models are usually operated regionally on horizontal resolutions between 1 km × 1 km and 10 km × 10 km. In many regions of the world, the horizontal resolution of such simulations is bottlenecked by the resolution of the available emission inventories (e.g. 0.1° × 0.1° in the EDGARv5 emission inventory, see Crippa et al., 2020), which strains the models' capability to capture air quality dynamics accurately. Furthermore, global simulations at these resolutions are infeasible due to their immense computational demands. Lastly, RCT simulations often show significant disagreements to observational reference data, e.g. to NO_2 in situ measurements at the surface (see e.g. Terrenoire et al., 2015; Visser et al., 2019; Kuik et al., 2016; Kuik et al., 2018; Poraicu et al., 2023; Kuhn et al., 2024a).

This thesis presents my attempt to address this problem with a new retrieval algorithm for tropospheric NO_2 concentration profiles from TROPOMI satellite observations. The main satellite product, the tropospheric vertical column density (VCD), is defined as

$$V_t\left(c\right) = \int_{\text{Earth surface}}^{\text{tropopause}} c(z)\,\mathrm{d}z \tag{1.1}$$

where c denotes the NO_2 concentration and z the altitude. A profile retrieval asks for c, instead. This problem is ill-posed, because different profiles c can produce the same observation $V_t(c)$. Therefore, there exists no well-defined inverse V_t^{-1} in the sense that $c = V_t^{-1}\left(V_t\left(c\right)\right)$.

However, taking a step back, the notion of a single well-defined solution c can be discarded. Instead, it can be attempted to identify a "weak" solution in the sense of a maximum a posteriori estimator

$$\hat{c}(\boldsymbol{y}, z) = \underset{c(z)\in\mathbb{R}^+}{\arg\max}\left[p(c(z)|\boldsymbol{y})\right] \tag{1.2}$$

i.e. the most *likely* solution to the inverse problem, given a vector of informative constraints $\boldsymbol{y}$. Here, the constraints $\boldsymbol{y}$ may contain almost arbitrary variables, from atmospheric reanalyses or measurements, as long as they stand in a mutual relationship to $c(z)$, be it by causality, a highly complex functional relationship, or simply correlation. For example, the planetary boundary layer height, which describes up to which altitude the lower troposphere can be considered well-mixed, is available from continuously operational atmospheric reanalyses, and the shape of tropospheric trace gas profiles obviously depends on it. The tropospheric vertical column density V_t, which relates to the amplitude of trace gas profiles, is itself contained in $\boldsymbol{y}$.

However, $p(c(z)|\boldsymbol{y})$ is unknown, which makes it impossible to obtain an estimate of the maximum a posteriori within the framework of an analytical model. In fact, our ability to describe even just the *approximate* relationship between tropospheric NO_2 profiles and reasonable candidates for $\boldsymbol{y}$ in analytical terms is mostly limited to parametrizing the profile amplitude by V_t, and the approximate profile shape by the boundary layer height. Nonethe-

less, it is clear that many other variables, e.g. emission patterns, meteorological data, cloud information, and variables describing technical details of the satellite measurement, stand in *some* relationship to the colocated NO_2 profiles.

I therefore propose an NO_2 profile retrieval by means of *statistical learning*. This is made possible by running the computationally expensive RCT model WRF-Chem, which generates a large training data set that could not be obtained from measurements. This dataset holds a representation of the desired data relationship $p(c(z)|\boldsymbol{y})$ in the form of samples (training examples), linking the model-external $\boldsymbol{y}$ to the model-internal $c(z)$. This relationship is then captured by a machine learning model, here by means of an *artifical feed-forward neural network* (see e.g. Schmidhuber, 2015). Neural networks are ideal for this task, seeing that they possess the capacity to reconstruct non-linear relationships from highly complex processes (see e.g. Schmidhuber, 2015; Hornik et al., 1989). Once trained, the neural network becomes a surrogate of the RCT model, which takes $\boldsymbol{y}$ as input and predicts a matching concentration profile $c_{\mathrm{NN}}(\boldsymbol{y}, z)$. The neural network presented in this thesis is called "NitroNet".

This approach of training a neural network on model-based data has the following advantages over existing methods for obtaining NO_2 profiles:

- Non-satellite methods, such as ground-based or airborne spectroscopic measurements, sondes, etc., are either extremely sparse or not even in continuous operation. For example, the FRM_4DOAS project (see Fayt et al., 2021), a network of ground-based spectroscopic measurements used to retrieve profiles of various trace gases, operates merely three instruments in all of Germany. Meanwhile airborne measurements, as described e.g. by Riess et al. (2022) and Poraicu et al. (2023), are usually conducted over short time spans, e.g. during measurement campaigns. NitroNet, however, can provide NO_2 profiles with the same daily global spatio-temporal coverage as TROPOMI.

- Analytic NO_2 profile retrieval approaches for satellites (e.g. cloud slicing) require specific physical conditions to be met (e.g. dense cloud cover at varying altitudes), are affected by the dependencies of the NO_2 profiles and cloud altitudes on meteorological conditions, and result in spatio-temporal resolutions far lower than the satellite observations themselves (e.g. seasonal means with $1° \times 1°$ horizontal resolution as reported by Marais et al., 2021). NitroNet, on the other hand, operates on the same spatio-temporal resolution as TROPOMI (daily measurements with up to 3.5 km $\times$ 5.5 km horizontal resolution).

- RCT models may produce realistic NO_2 profiles, but they are very slow and suffer from various operational hurdles, such as a high sensitivity to certain input data (e.g. emission data), which may not be accurately available in some regions of the Earth. NitroNet is

orders of magnitude faster and, although requiring them as input, less sensitive to the aforementioned critical input data.

- Although it is possible to assimilate satellite observations into regional models, most of them (including WRF-Chem) lack an out-of-the-box implementation. On the other hand, constraining a neural network's predictions on satellite data (or data from any other source, given they are reliably available) is an automated process during training and requires no additional effort.

- RCT models are prone to systematic errors in localized sub-regions of their domain, which can be identified by validation against observational data. NitroNet utilizes a data curation approach, whereby such sub-regions are removed prior to training. This reduces the extent to which NitroNet adopts the systematic errors in the training data from the RCT model and results in demonstrably higher prediction quality.

On the other hand, caution is warranted. The regional model's representation of $p(c(z)|\boldsymbol{y})$ may deviate from the "true" $p(c(z)|\boldsymbol{y})$ and thus fail to represent the real world accurately. Moreover, the neural network's predictions $c_{\mathrm{NN}}(\boldsymbol{y}, z)$ may differ from the desired maximum a posteriori as expressed in eq. (1.2). The latter depends on numerous factors, e.g. the neural network's capacity, its training procedure (e.g. the choice of loss function), and the characteristics of $p(c(z)|\boldsymbol{y})$ (e.g. its number of modes). In other words, it must be assured that the regional model is a good representation of the real world, and that NitroNet accurately reproduces the results of the regional model. Portions of the presented work will address this by validation of WRF-Chem and NitroNet against independent measurements. Another related question is, how informative the selected constraints $\boldsymbol{y}$ are about the tropospheric NO_2 profile shape. A lack of crucial information in $\boldsymbol{y}$ would result in high neural network prediction errors, even if the neural network itself was designed appropriately. The relative contribution of the constraints to the network's predictive capability can be quantified by computing their Shapley scores (see e.g. Štrumbelj and Kononenko, 2013).

The thesis is structured as follows: Chapter 2 provides the necessary fundamentals. This covers the relevant aspects of NO_x chemistry, atmospheric dynamics, measurement methods, as well as introductions to RCT modelling and machine learning. Chapter 3 deals with the creation of the training dataset from the WRF-Chem RCT model. The model data are validated against observational reference data, and important model parameters are recalibrated. Chapter 4 describes the development and design of the NitroNet neural network. Then, NitroNet is intercompared to observational reference data and WRF-Chem predictions. The model's generalization capability is investigated in a regional and seasonal study. Chapter 5 summarizes the results, gives an outlook, and concludes the thesis.

Chapter 2

Fundamentals

2.1 Nitrogen dioxide in the troposphere

Nitrogen dioxide (NO_2) is an important marker of air pollution. It is mostly emitted into the troposphere as NO_x (= NO_2 + NO) from anthropogenic sources, e.g. combustion in car motors. As a highly toxic trace gas, it has been recognized for its negative impact on human health. Particularly in long-term exposure, even moderate levels of NO_2 can cause airway inflammation, asthma, and other respiratory symptoms in humans (World Health Organization, 2000; Latza et al., 2009; Khaniabadi et al., 2016; Huangfu and Atkinson, 2020). NO_2 also contributes to air pollution by conversion to other secondary air pollutants, most importantly ozone (O_3), nitric acid (HNO_3), peroxyacetyl nitrate (PAN), and particulate matter (PM). Particularly in warm, sunny weather in conjunction with photooxidants and volatile organic compounds (VOCs), NO_x plays a key role in the emergence of "summer smog". It is estimated, that air pollution leads to millions of premature deaths per year worldwide (Lelieveld et al., 2015). As such, it is not only a topic of scientific interest, but of political urgency. Since the 1950s, attempts to reduce air pollution by technical advancements (e.g. catalytic converters in cars) and governmental regulations (e.g. of industrial emissions) have been made. Subsequently, air pollution has indeed decreased in an ongoing trend (see, e.g. Fig. 2.1, which shows a time series of yearly-average NO_2 surface concentrations in Germany). Similar trends are observed in other countries. Nonetheless, violations of the WHO air quality guidelines are still a recurring problem in many places. For example, the European Environment Agency reported in 2023 that approximately 90 % of European citizens living in urban areas were exposed to NO_2 in amounts violating the 2021 WHO air quality guidelines (10 µg m^{-3} yearly average, see European Environment Agency, 2024; World Health Organization, 2021). On the other hand, the same report states that only 1% of the same populus were exposed to NO_2 pollution exceeding the EU air quality standards at the time (200 µg m^{-3} hourly averages with 3 permitted exceedances each year, 40 µg m^{-3} yearly average).

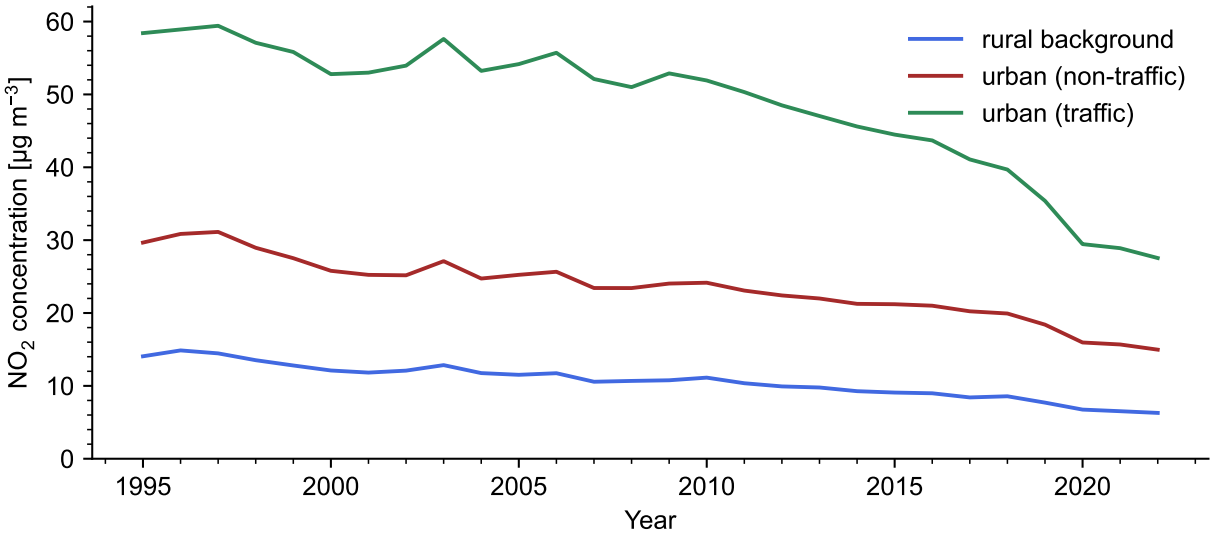

Figure 2.1: Average NO_2 surface concentrations from in situ measurements in Germany from 1995 – 2022. Yearly average values. Data taken from a technical report on air quality, published by the environmental agency of Germany (UBA, see Kessinger et al., 2023).

This section gives an overview of NO_x emissions (sect. 2.1.1), and relevant aspects of NO_x chemistry during day and night (sect. 2.1.2). It does not thematize stratospheric NO_x chemistry or further aspects of air pollution unrelated to NO_x. The text follows the explanations of Finlayson-Pitts and Pitts Jr. (2000) and Seinfeld and Pandis (2016).

2.1.1 Sources of NO_x

NO_2 is emitted into the atmosphere in the form of NO_x from high-temperature combustion in lightning events, biomass burning, and anthropogenic sources, as well as biogenic emissions from soil bacteria. It is therefore both a primary pollutant (from the NO_2 contained in NO_x), and a secondary pollutant (from the NO in NO_x, which is converted to NO_2). It is estimated, that globally approximately 72 Tg (Teragram) of NO_x (expressed as NO_2) are emitted every year (referring to the year 1992, see Finlayson-Pitts and Pitts Jr., 2000). Most cases of significant air pollution are dominated by anthropogenic emissions from road traffic, industrial processes, fuel burning for electrical utilities, other non-road engines, and aircraft. Such combustion processes emit with NO_2/NO_x ratios in the range of $5 - 40$ %, depending e.g. on temperature (Jimenez et al., 2000; Costantini et al., 2016; Wild et al., 2017; Richmond-Bryant et al., 2017). Another notable anthropogenic source of nitrogen is the use of agricultural fertilizers in a biannual cycle, which emits NO_x, NH_3, and N_2O. An exemplary overview of the anthropogenic NO_x emissions from different anthropogenic sectors is given in Table 2.1 (here: for May 2015 and 2018 in Germany). A map plot of the corresponding emission data from the EDGARv5 emission inventory is found in Fig. 2.2.

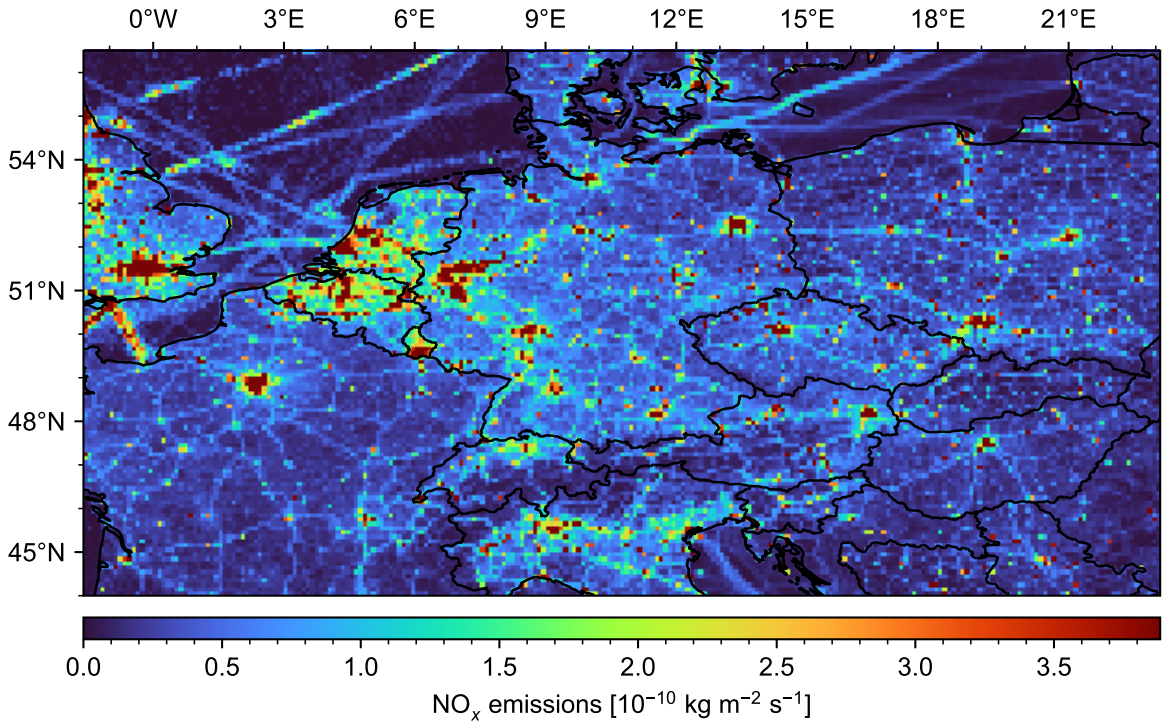

Figure 2.2: Anthropogenic NO_x emissions in Europe, May 2015. Monthly mean data with a resolution of 0.1° × 0.1°, taken from the EDGARv5 emission inventory (see Crippa et al., 2020). Shown here is the sum over all emission sectors listed in Table 2.1.

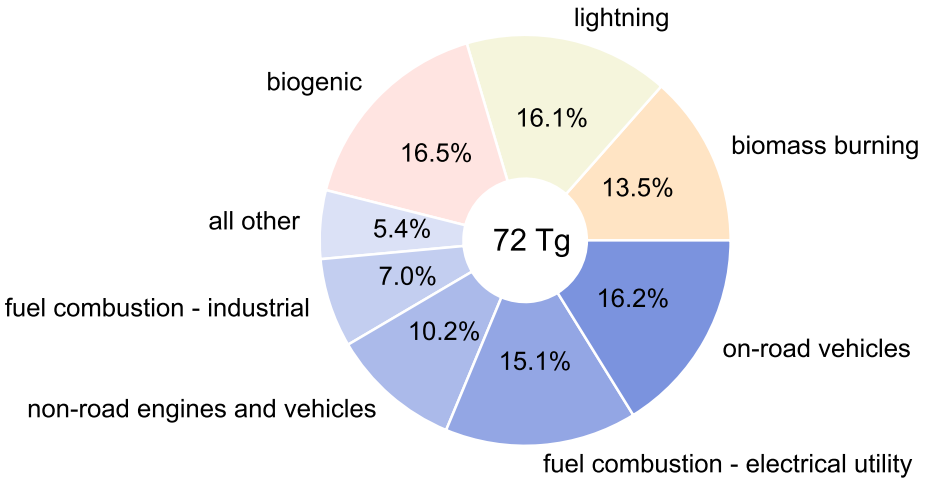

Figure 2.3: Pie chart of the anthropogenic and natural contributions to the global yearly NO_x emissions. All contributions from anthropogenic sources are drawn in shades of blue. The distinction between anthropogenic and natural sources is based on an estimation of the year 1992 (Müller, 1992). The partitioning of the anthropogenic sources is based on an estimation of the Environmental Protection Agency of the year 1996 and refers to the United States (U.S. Environmental Protection Agency, 1997). The estimation of lightning NO_x was taken from Finlayson-Pitts and Pitts Jr. (2000).

emission sector	UBA-E [%]	EDGARv5 [%]
traffic (no resuspension)	43.8	38.7
power industry	18.5	15.0
agricultural soils	10.1	4.5
energy for buildings	7.3	6.6
manufacturing industry	7.2	15.8
non-metallic minerals production	2.5	0.0
production of chemicals	2.4	0.3
shipping	2.1	2.1
iron and steel production	1.8	< 0.1
oil refineries and transformation industry	1.7	1.8
aviation landing and take-off	1.2	1.8
railways, pipelines, and off-road transport	0.9	2.3
production of food, pulp, and paper	0.3	0.3
manure management	0.1	1.1
fuel exploitation	0.1	0.0
solid waste incineration	0.1	0.1
non-ferrous metal production	0	< 0.1
non-energy use of fuels	< 0.1	0.0
agricultural waste burning	N/A	0.3
fossil fuel fires	N/A	< 0.1
aviation climbing and descent	N/A	5.9
aviation cruise	N/A	3.4

Table 2.1: Overview of the anthropogenic NO_x emissions in Germany. Each entry denotes the contribution of one emission sector from two independent emission inventories (Emission inventory of the German Umweltbundesamt in May 2018 ("UBA-E"), see Hausmann et al., 2020, and EDGARv5 in May 2015, see Crippa et al., 2020). N/A means "not available".

2.1.2 Chemistry of NO_x

Non-organic NO_x photochemistry

Upon emission nitrogen oxides are subject to the photochemistry of the troposphere. NO can undergo thermal oxidation:

$$2\,\mathrm{NO} + \mathrm{O_2} \longrightarrow 2\,\mathrm{NO_2} \tag{2.1}$$

Note that eq. (2.1) is of second order in [NO] and only relevant at high partial NO pressures (e.g. in plumes emitted from power plant stacks). Under more moderate conditions, it is

rather insignificant. The presence of NO_x, O_3, and light leads to a steady state governed by the following reactions

$$NO + O_3 \longrightarrow NO_2 + O_2 \tag{2.2}$$

$$O(^3P) + O_2 \xrightarrow{M} O_3 \tag{2.3}$$

$$NO_2 + h\nu(\lambda < 420 \text{ nm}) \longrightarrow NO + O(^3P) \tag{2.4}$$

from which the *Leighton relationship* (see Leighton, 2012) follows:

$$\frac{[O_3][NO]}{[NO_2]} = \frac{J_{NO_2}}{k_{2.2}} \tag{2.5}$$

where:

J_{NO_2} : photolysis rate of NO_2

$k_{2.2}$: reaction rate of eq. (2.2)

Here, $O(^3P)$ denotes the low energy triplet state of atomic oxygen. Both J_{NO_2} and $k_{2.2}$ are not constant; J_{NO_2} depends on the actinic flux (which obviously follows a strong diurnal cycle), and $k_{2.2}$ is temperature dependent (Atkinson et al., 2004). The reaction of $O(^3P)$ with the abundantly available O_2 of the atmosphere can be regarded as instantaneous.

Organic NO_x photochemistry

The key mechanism of summer smog is a complex scheme of reactions involving NO_x, atmospheric oxidants, CO, volatile organic compounds (VOCs), and water vapor. These reactions are sequences, in which an organic molecule is oxidized, which in turn oxidizes NO to NO_2, and eventually restores the original oxidant. NO_2 photodissociates in reaction (2.4) during daytime, which restores NO and leaves O for the formation of O_3. Only through this mechanism the high levels of NO_2 and subsequently O_3 observed in summer smog can be explained.

The relevant oxidants of summer smog are O_3 and the "odd hydrogen radicals" OH, HO_2, and RO_2. The OH radical is the most important oxidant of the atmosphere. Because it is the dominant sink for many trace gases, it is also referred to as the *detergent of the atmosphere*. For example, the reaction of NO_2 with OH to HNO_3 is a dominant atmospheric NO_x sink, limiting its lifetime to $\sim 2-6$ h during summer, and to ~ 20 h in winter (Liu et al., 2016; Shah et al., 2020; Lange et al., 2022). OH is produced in photolytic processes, such as

$$O_3 + h\nu(\lambda < 340 \text{ nm}) \longrightarrow O(^1D) + O_2 \tag{2.6}$$

$$O(^1D) + H_2O \longrightarrow 2\,OH \tag{2.7}$$

$$HONO + h\nu(\lambda < 370\ nm) \longrightarrow OH + NO \tag{2.8}$$

$$HNO_3 + h\nu \longrightarrow OH + NO_2 \tag{2.9}$$

$$HOOH + h\nu(\lambda < 370\ nm) \longrightarrow 2\,OH \tag{2.10}$$

and within the cycle of smog chain reactions via the hydroperoxy radical (HO_2),

$$HO_2 + NO \longrightarrow OH + NO_2 \tag{2.11}$$

whereby NO is converted to NO_2 specifically without loss of ozone. The same is possible with an alkylperoxy radical (RO_2) instead:

$$RO_2 + NO \longrightarrow RO + NO_2 \tag{2.12}$$

This additional oxidation channel leads to the eventual surplus of NO_x (and thereby, O_3) during summer smog. A main source of daytime HO_2 is aldehyde photolysis (mainly formaldehyde, i.e. HCHO):

$$RCHO + h\nu(\lambda < 420\ nm) \longrightarrow R + HCO \tag{2.13}$$

$$H + O_2 \xrightarrow{M} HO_2 \tag{2.14}$$

$$HCO + O_2 \longrightarrow HO_2 + CO \tag{2.15}$$

where H in eq. (2.14) is available i.e. if RCHO = HCHO. There exist further production mechanisms of HO_2, but they are overall similar in the sense that they revolve around the oxidation of organic molecules. An example of a non-organic HO_2 source is the decomposition of peroxynitric acid,

$$HO_2NO_2 \longleftrightarrow HO_2 + NO_2 \tag{2.16}$$

RO_2 is produced by the oxidation of alkyl radicals, which in turn are produced by the oxidation of organics. For example, OH can abstract hydrogen from an alkane:

$$OH + RH \longrightarrow R + H_2O \tag{2.17}$$

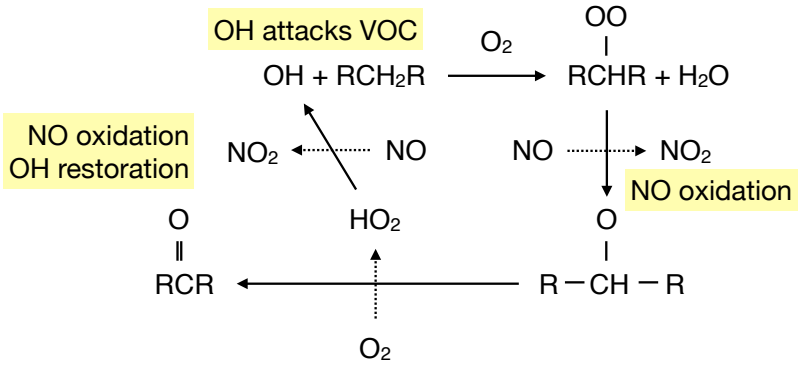

Figure 2.4: Typical VOC oxidation sequence of summer smog. Here, an alkane (RCH_2R) is oxidized by OH. The original OH molecule is eventually restored, and two NO molecules are oxidized to NO_2. Adapted from Finlayson-Pitts and Pitts Jr. (2000).

with similar processes for other oxidants. The alkyl radical then forms the alkylperoxy radical:

$$\mathrm{R} + \mathrm{O_2} \xrightarrow{\mathrm{M}} \mathrm{RO_2} \tag{2.18}$$

With this knowledge at hand, a typical summer smog reaction sequence as shown in Fig. 2.4 can be understood. Here, OH (the oxidant) attacks RCH_2R (the VOC), eventually restoring OH from HO_2, and producing two molecules of NO_2 on the way. With the original oxidant restored, the cycle may repeat, but will eventually be terminated by other reactions, e.g.

$$\mathrm{OH} + \mathrm{NO_2} \xrightarrow{\mathrm{M}} \mathrm{HNO_3} \tag{2.19}$$

$$\mathrm{HO_2} + \mathrm{HO_2} \xrightarrow{\mathrm{M,H_2O}} \mathrm{H_2O_2} + \mathrm{O_2} \tag{2.20}$$

$$\mathrm{RO_2} + \mathrm{RO_2} \longrightarrow \mathrm{R_2O_2} + \mathrm{O_2} \tag{2.21}$$

$$\mathrm{RO_2} + \mathrm{HO_2} \longrightarrow \mathrm{ROOH} + \mathrm{O_2} \tag{2.22}$$

with eq. (2.19), followed by dry or wet deposition, being the dominant NO_x sink.

Depending on whether peroxide formation (eq. 2.20 – 2.22) or the formation of nitric acid (eq. 2.19) is the dominant radical source, two regimes may emerge: If peroxide formation dominates, the *NO_x-sensitive regime* is established. This is the case, when the VOC concentration is high compared to the NO_x concentration. Then, the O_3 formation is ultimately determined (and limited) by the oxidation of NO, but not by oxidation of the abundant VOCs. The O_3 production rate grows with the NO_x concentration, but is mostly insensitive to the VOC concentration. In the opposite case, if nitric acid formation dominates, the OH con-

centration shrinks with growing NO_2 concentration. The ozone production depends on OH as the oxidizer of the VOCs. Increasing the NO_x concentration will no longer lead to faster O_3 production (and may even reduce it), but an increase in VOCs will. This is called the *VOC-sensitive regime*.

Although the main products of summer smog are NO_2 and ultimately O_3, there are also other relevant air pollutants related to NO_x, notably peroxyacetyl nitrate (PAN, a powerful lachrymator and long-lived NO_x reservoir) formed via

$$CH_3CO_3 + NO_2 \longrightarrow CH_3CO_3NO_2 \tag{2.23}$$

and particulate matter, e.g. from oxidation of NO_2 to HNO_3 (see eq. 2.19) and subsequent formation of ammonium nitrate (a salt):

$$HNO_3 + NH_3 \longrightarrow NH_4NO_3 \tag{2.24}$$

Nighttime NO_x chemistry

The lack of photolysis during nighttime changes NO_x chemistry fundamentally. The nitrate radical (NO_3), an important nighttime oxidant, and nitrogen pentoxide (N_2O_5) are formed via

$$NO_2 + O_3 \longrightarrow NO_3 + O_2 \tag{2.25}$$

$$NO_3 + NO_2 \longleftrightarrow N_2O_5 \tag{2.26}$$

Both photodissociate rapidly during daytime. NO_3 and N_2O_5 act as NO_2 reservoirs, and as HNO_3 sources, via

$$NO_3 + RH \longrightarrow HNO_3 + R \tag{2.27}$$

$$N_2O_{5(g)} + H_2O \xrightarrow{\text{surfaces, aerosols}} 2\,HNO_3 \tag{2.28}$$

Nighttime chemistry of NO_x is not highly important for this thesis, because the used satellite measurements are conducted at daytime. Nonetheless, it has some relevance for the validation of regional chemistry and transport models, which is addressed in Chapter 3.

2.2 Vertical structure of the atmosphere

This section gives a coarse overview of the vertical structure of the atmosphere. The barometric height formula, as well as the dry and moist adiabatic lapse rates are introduced. The

distinction between the atmospheric layers is an important prerequisite for the identification of different atmospheric transport mechanisms, which in turn are essential for the formation of NO_2 profiles. Furthermore, many of the input variables used for the NitroNet model are only available within, and thus specific to, the troposphere. Lastly, approximate pressure and temperature profiles are required later to convert NO_2 profiles between different vertical coordinate systems. The explanations given here follow Roedel and Wagner (2017) and Jacob (1999).

2.2.1 Stratosphere, tropopause, and troposphere

Earth's atmosphere is coarsely divided into the troposphere ($\lesssim 10-15$ km), the tropopause ($\sim 10-15$ km), and the stratosphere ($\sim 10-50$ km). Beyond the stratosphere, there are the mesosphere and the thermosphere, which are not relevant here.

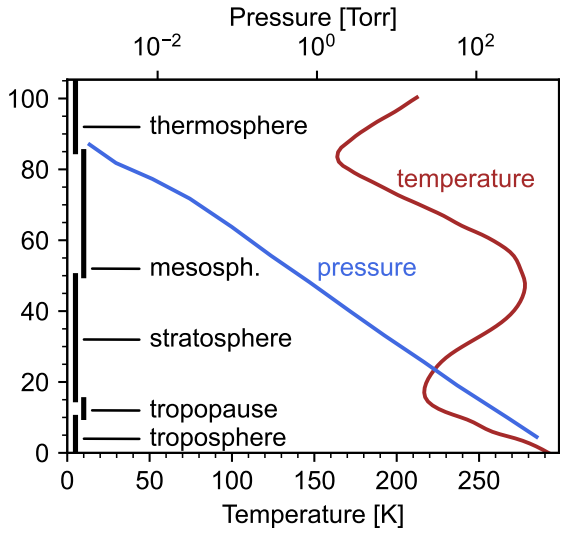

Figure 2.5: Different layers of the atmosphere and typical pressure and temperature profiles. Adapted from Finlayson-Pitts and Pitts Jr. (2000).

Troposphere and stratosphere can be distinguished by their temperature gradients. When sunlight reaches the Earth (*actinic radiation*), a large fraction of it is absorbed by the surface, which subsequently heats itself and the air parcels immediately above. Through thermal updraft (= warm air rising due to lower density), tropospheric air can be transported upwards on the time scale of days or less while cooling down quasi-adiabatically. As a result, a (mostly) linear temperature falloff is established. Of course, the transport of air is subject to meteorological phenomena (e.g. *inversions*), which make the true vertical structure of the atmosphere more complex.

In the stratosphere the temperature gradient is flipped, meaning that temperature increases with altitude. This is due to the absorption of light by ozone (O_3), produced via the *Chapman cycle*:

$$O_2 + h\nu(\lambda < 242\ \text{nm}) \longrightarrow 2\,O \tag{2.29}$$

$$O + O_2 \xrightarrow{M} O_3 \tag{2.30}$$

$$\mathrm{O} + \mathrm{O_3} \longrightarrow 2\,\mathrm{O_2} \tag{2.31}$$

$$\mathrm{O_3} + h\nu(\lambda \lesssim 336\ \mathrm{nm}) \longrightarrow \mathrm{O}(^1\mathrm{D}) + \mathrm{O_2} \tag{2.32}$$

The formation of stratospheric ozone has important implications for the radiative balance of the Earth: Ozone absorbs strongly at wavelengths between $200 - 310$ nm (the UV-C region of the electromagnetic spectrum). Photodissociation of O_3 as in eq. (2.32) requires energy equivalent to light with a wavelength of 310 nm. All surplus energy absorbed from light of shorter wavelengths will be expelled in the form of heat. Additionally, eq. (2.30) is exothermic. Subsequently, the stratosphere heats up and inverts the temperature gradient of the troposphere. Furthermore, the absorption of UV-C in the stratosphere limits the photochemical processes of the troposphere to energies corresponding to wavelengths of > 290 nm, and shields humans, animals, and plants from most of the hazardous impacts of UV radiation, making it an essential prerequisite for human life (see e.g. Hussein, 2005). Due to the stratospheric thermal inversion, a stable layering is formed. Convection is suppressed and mixing is far slower than in the troposphere. In consequence, the stratosphere is also mostly devoid of meteorological effects (e.g. the formation of clouds, or precipitation), with a few rare exceptions (e.g. the formation of polar stratospheric clouds). In other words, while the tropospheric temperature gradient is determined by convection (in the form of air circulation), that of the stratosphere is determined by its radiative balance.

2.2.2 Derivation of the vertical pressure profile (barometric height formula)

The barometric height formula describes pressure as a function of altitude above ground. The pressure exerted onto a cross-sectional area below a column of air reads

$$\mathrm{d}p = -\rho g\,\mathrm{d}z = -p \cdot \frac{Mg}{RT}\,\mathrm{d}z \tag{2.33}$$

where:

p : pressure

ρ : density of air

g : gravity acceleration of Earth (≈ 9.81 m s^{-2})

z : altitude above ground

M : molar mass ($\approx 28,97$ g mol^{-1} for dry air)

R : universal gas constant (≈ 8.314 J mol^{-1} K^{-1})

T : temperature

The second part of the equation is obtained from the ideal gas equation ($\rho = \frac{Mp}{RT}$). If temperature is assumed constant (a vast simplification), integration yields

$$p(z) = p_0 \exp\left(-\frac{Mg}{RT}z\right) := p_0 \exp\left(-z/z_0\right) \tag{2.34}$$

where:

p_0 : surface pressure (= 1013.25 mbar at standard conditions)

z_0 : the "scale height", defined as $z_0 = \frac{RT}{Mg}$

2.2.3 Atmospheric lapse rate and temperature profiles

As depicted in Fig. 2.5, the atmospheric temperature depends strongly on altitude. The barometric height formula can be augmented by a vertical temperature gradient (the *lapse rate*), derived from thermodynamic principles. This is common textbook knowledge (see e.g. Roedel and Wagner, 2017), hence the corresponding derivations were moved to the Appendices A.1 and A.2. The most important aspects of the lapse rate and vertical temperature profiles are summarized here:

- The *dry adiabatic lapse rate* Γ equates to $\Gamma = \dfrac{\mathrm{d}T}{\mathrm{d}z} \approx -0,00981$ K m^{-1} and describes the vertical temperature gradient of dry air.

- For the case of moist air *without* condensation, the moist adiabatic lapse rate equates to

$$\frac{\mathrm{d}T}{\mathrm{d}z} = -\frac{g}{s \cdot c_{p,\mathrm{water}} + (1-s) \cdot c_{p,\mathrm{air}}} \tag{2.35}$$

  where:

$c_{p,\text{water}}$: specific heat of water vapor ($\approx 1,00$ J K^{-1} g^{-1})

$c_{p,\text{air}}$: specific heat of dry air ($\approx 1,86$ J K^{-1} g^{-1})

s : mass ratio of water vapor to total air

which usually differs insignificantly from that of dry air (e.g. by less than 1 % for $s = 0.01$).

- If the water content of the rising air exceeds the saturation level, and condensation occurs, the lapse rate must account for the release of latent heat (L). It then reads

$$\frac{\mathrm{d}T}{\mathrm{d}z} = -\frac{pMg}{RTL\frac{\rho_{\text{w,sat}}}{\mathrm{d}T} + pC_p}$$

where:

$\rho_{\text{w,sat}}$: absolute moisture at the condensation limit

C_p : specific molar heat at constant pressure

With condensation, the (absolute) lapse rate can be significantly smaller. For example, the *international standard atmosphere* assumes a lapse rate of $\frac{\mathrm{d}T}{\mathrm{d}z} = -0,0065$ K m^{-1} (see International Civil Aviation Organization, 1993).

- Integration of the barometric height formula from eq. (2.34) using a fixed lapse rate yields

$$p(z) = p_0 \cdot \left(1 + \frac{\frac{\mathrm{d}T}{\mathrm{d}z}z}{T_0}\right)^{-\frac{Mg}{R\frac{\mathrm{d}T}{\mathrm{d}z}}}$$

where the subscript 0 refers to temperature and pressure at the surface. Using the values from the international standard atmosphere, namely

$$T_0 = 288.15 \text{ K}, \quad p_0 = 1013.25 \text{ hPa}, \quad \frac{\mathrm{d}T}{\mathrm{d}z} = -0.0065 \text{ K m}^{-1}$$

one obtains

$$p(z) = 1013.25 \text{ hPa} \cdot \left(1 - \frac{0,0065 \text{ K m}^{-1}}{288.15 \text{ K}} \cdot z\right)^{5.255} \tag{2.36}$$

and

$$z(p) = \frac{288.15\ \text{K}}{0,0065\ \text{K m}^{-1}} \cdot \left(1 - \left(\frac{p}{1013.25\ \text{hPa}}\right)^{\frac{1}{5.255}}\right) \tag{2.37}$$

2.2.4 The free troposphere, the planetary boundary layer, and its sub-layers

The troposphere can be divided into further sub-layers, each governed by different forces. This section leads to the concept of the *planetary boundary layer* (PBL), which is relevant to this thesis in multiple regards. For now, the PBL is described qualitatively from considerations of atmospheric (fluid) dynamics in horizontal direction. The relevance of the PBL for the formation of NO_2 profiles is discussed afterwards in the context of vertical mixing within the PBL (see sect. 2.2.5). Like before, this section aims to deliver a concise summary. Detailed derivations for the equations given here are found in Appendix A.3. As depicted in Fig. 2.6, the troposphere can be separated into the following further sublayers:

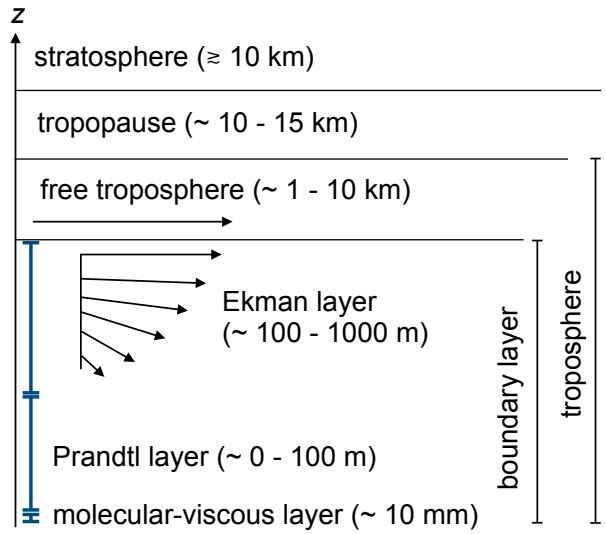

Figure 2.6: Atmospheric layers up to the stratosphere.

- The free troposphere ($\sim 1 - 10$ km). This layer is characterized by a force equilibrium between the *Coriolis force* and the pressure gradient, i.e.

$$\boldsymbol{F}_c = 2\rho \cdot (\boldsymbol{v} \times \boldsymbol{\Omega}) = -\boldsymbol{\nabla}_h p = \boldsymbol{F}_p$$

where:

ρ	: density
p	: pressure
$\boldsymbol{v}$	: horizontal wind speed vector
$\boldsymbol{\Omega}$	: Earth's angular velocity
$\boldsymbol{\nabla}_h = (\partial_x, \partial_y, 0)$	: the horizontal gradient operator

This results in *geostrophic horizontal flow*, described by the velocity equations

$$v_x = -\frac{1}{\rho f}\partial_y p \qquad \text{and} \qquad v_y = \frac{1}{\rho f}\partial_x p \tag{2.38}$$

where f denotes the *Coriolis parameter* (see Appendix A.3).

- The *Ekman layer* ($\sim 100 - 1000$ m) which, in contrast to the free troposphere, exhibits friction. The friction force is expressed as the derivative $\boldsymbol{F}_f = \frac{\mathrm{d}}{\mathrm{d}z}\boldsymbol{\tau}$ of a *shear stress vector* $\boldsymbol{\tau} = (\tau_{xz}, \tau_{yz}, 0)^{\mathrm{T}}$, whose entries are defined as

$$\tau_{xz} = -(K + \nu)\rho\frac{\partial v_x}{\partial z}$$

where K denotes the turbulent diffusion coefficient and ν the kinematic viscosity. τ_{yz} is defined in analogy. The force equilibrium $\boldsymbol{F}_c + \boldsymbol{F}_p + \boldsymbol{F}_f = 0$ results in an inclination of the wind trajectory towards regions of lower pressure. The strength of this inclination grows with the relative influence of friction, i.e. with the distance to the free tropospheric interface. This altitude dependent inclination is called the *Ekman spiral*. The planetary boundary layer (PBL) is often defined as the region between Earth's surface and the top of the Ekman layer.

- The *Prandtl layer* ($\sim 0 - 100$ m), characterized by exceptionally strong friction forces in the force equilibrium $\boldsymbol{F}_c + \boldsymbol{F}_p + \boldsymbol{F}_f = 0$. In the stationary state, this regime can be identified by small relative shear stress gradients, i.e. cases where

$$\Delta z \cdot \frac{\mathrm{d}\tau}{\mathrm{d}z} \ll \boldsymbol{\tau}$$

holds for some distance Δz. Here, Δz defines the vertical extent of the Prandtl layer.

- The *molecular-viscous* layer (~ 10 mm), where friction by thermal diffusion dominates turbulent friction, i.e. $\nu \gg K$.

2.2.5 Vertical mixing in the planetary boundary layer

The considerations in the previous section revolved around horizontal flow, although the introduction of friction in fluid motion already implied some form of vertical exchange by thermal and turbulent diffusion. Depending on the atmosphere's layering (stable or unstable), vertical exchange can initiate vertical mixing, which significantly influences the shape of atmospheric trace gas profiles. In the following, a qualitative overview of vertical mixing in the PBL is given.

Vertical updraft, buoyancy, and atmospheric stability

A notable contribution to the vertical transport of the PBL is vertical (up)draft of air parcels due to buoyancy. Buoyancy describes the vertical motion of an object (here: an air parcel) of density ρ and volume V, immersed in a fluid of density ρ_f. In such a scenario, two vertical forces are exerted onto the immersed object: the gravity force $F_g = \rho V g$ and the pressure gradient force $F_p = -\rho_f V g$. The buoyant force reads

$$F_b = F_g + F_p = (\rho - \rho_f) V g \tag{2.39}$$

In this notation a positive force points downards. Normalized w.r.t. volume the equation reads $F_b = (\rho - \rho_f) g$. If $\rho_f > \rho$, the immersed object is forced upwards, which is known as *Archimedes' principle*.

In the context of vertical motion it is worthwhile to revisit the role of the atmospheric lapse rate $\frac{\mathrm{d}}{\mathrm{d}z} T_{\mathrm{atm}}$. The dry and wet adiabatic lapse rates described in sect. 2.2 were derived under assumption of adiabatic conditions. Any atmosphere left to evolve adiabatically from an arbitrary initial state is expected to converge towards an equilibrium state with neutral buoyancy (where, again, the atmospheric temperature profile follows the adiabatic lapse rate). In reality, however, this atmospheric equilibrium is disturbed by external sources and sinks of thermal heat, such as condensation of water vapor and absorption of radiation. As a consequence, the true atmospheric lapse rate $\frac{\mathrm{d}}{\mathrm{d}z} T_{\mathrm{atm}}$ may differ from the adiabatic lapse rate Γ, which leads to the notion of *atmospheric stability*. Consider a tropospheric air parcel at altitude z, being pushed upwards to $z + \mathrm{d}z$. Because $p(z + \mathrm{d}z) < p(z)$, the air parcel expands (here: assumed adiabatically) and subsequently cools. According to eq. (2.39), one might expect the air parcel to sink back down to its original altitude on account of buoyancy. However, the temperature of the surrounding atmosphere is also expected to have changed from z to $z + \mathrm{d}z$. The decisive question is, whether the adiabatic cooling rate of the air parcel (Γ) is smaller or larger than the temperature gradient of the surrounding atmosphere ($\frac{\mathrm{d}}{\mathrm{d}z} T_{\mathrm{atm}}$):

- If $\frac{\mathrm{d}}{\mathrm{d}z} T_{\mathrm{atm}} > \Gamma$, the air parcel at $z + \mathrm{d}z$ is *colder* and its density *higher* than the surrounding atmosphere. The air parcel sinks back to its original position, and vertical motion is suppressed. This atmospheric condition is called *stable*.

- If $\frac{\mathrm{d}}{\mathrm{d}z} T_{\mathrm{atm}} < \Gamma$, the air parcel at $z + \mathrm{d}z$ is *warmer* and its density *lower* than the surrounding atmosphere. Any initial "nudge" in vertical direction is subsequently amplified by buoyancy. This atmospheric condition is called *unstable*, and the atmosphere is called *convective*.

- If $\frac{\mathrm{d}}{\mathrm{d}z} T_{\mathrm{atm}} = \Gamma$, vertical motion is neither suppressed, nor amplified. The atmosphere is

called *neutral*.

- If $\frac{\mathrm{d}}{\mathrm{d}z}T_{\mathrm{atm}} > 0$ (a special case of the stable atmosphere), temperature increases with altitude, which yields particularly stable conditions. This is also referred to as a *temperature inversion*. For example, such an inversion is observed in the stratosphere due to the absorption of solar UV radiation.

An important realization is that both upwards and downwards displacements are equally amplified in an unstable atmosphere. Therefore, instability causes bidirectional vertical mixing, as opposed to merely unidirectional transport.

Diurnal mixing cycle in the planetary boundary layer

The previous considerations are sufficient to explain the typical mixing processes in the lower troposphere, as well as their diurnal dependence. The main driver for atmospheric mixing is radiative forcing by sunlight. Earth's surface absorbs the energy of incident sunlight, heats up, and in turn heats the air immediately above. This uptake of heat violates the adiabatic condition $\mathrm{d}Q = 0$. Subsequently, an unstable atmospheric layering ($\frac{\mathrm{d}}{\mathrm{d}z}T_{\mathrm{atm}} < \Gamma$) is formed. This, together with wind shear as a result of air dragging over the rigid surface, leads to turbulent motion. In consequence, the *convective mixed layer* is formed, whose constituents are generally assumed to be well-mixed within timescales of < 1 h. Near the surface (up to the lowest ~ 10 % of the convective mixed layer), another sublayer (the *surface layer*) characterized by very strong wind shear is formed, which can yield an even higher contribution to turbulence than the buoyant flux (see e.g. Stull, 1988). The convective mixed layer builds up from sunrise until approximately noontime, and it is capped by an inversion (also called the *entrainment zone* during daytime) at typically $1-2$ km altitude. The depth of the mixed boundary layer is called *planetary boundary layer height* (PBLH). In extreme environments (e.g. in arctic or tropical regions), the PBLH can be as low as ~ 100 m or as high as ~ 4000 m. The PBLH depends e.g. on surface type (due to the influence of surface roughness on wind shear, and surface material on radiative uptake capacity) and the actinic flux. At nighttime the convective mixing mechanism collapses. As the radiative forcing by sunlight depletes, thermal cooling sets in near the surface, resulting in a stable nocturnal boundary layer due to the radiation inversion. The wind shear and its contribution to vertical mixing may persist. Above the stable nocturnal layer a residual layer remains, which contains the atmospheric constituents that were well-mixed during daytime, but are no longer subject to further mixing, due to the stable nocturnal layer below and the free troposphere above. At sunrise the cycle repeats. Figure 2.7a shows an overview of the described process.

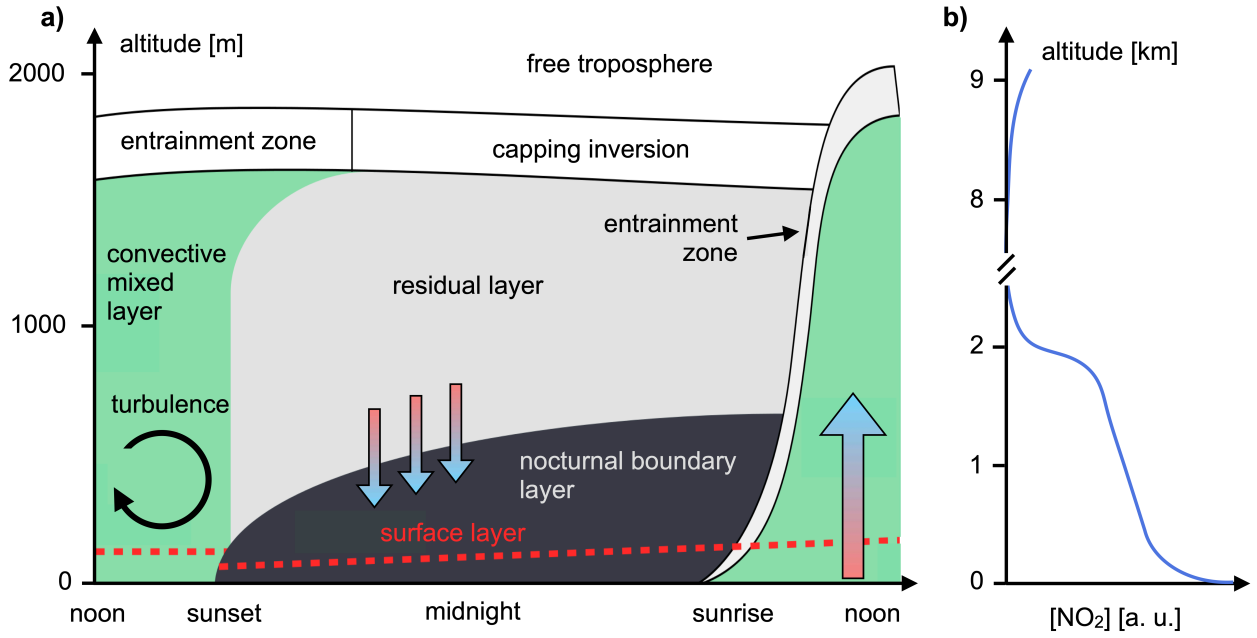

Figure 2.7: Overview of the planetary boundary layer. a) Diurnal cycle of the convective planetary boundary layer. Adapted from Stull (1988). **b)** Typical vertical NO_2 profile within the troposphere in arbitrary units, see e.g. Douros et al. (2023).

2.2.6 Typical shape of daytime tropospheric NO_2 profiles

With this knowledge at hand, the typical shape of daytime tropospheric NO_2 profiles can be explained. Towards the surface, where most of the NO_2 emissions are injected, the profiles have a strong exponential tail. Within the well-mixed PBL, trace gas profiles often show a moderate linear or exponential fall-off towards higher altitudes (see e.g. Frieß et al., 2019; Kumar et al., 2020). At the top of the PBL steep gradients occur, usually described as a sudden concentration cutoff. Beyond the PBL, in the free troposphere, concentrations remain at a correspondingly low level. Approaching the stratosphere, NO_2 concentrations may rise again due to aircraft emissions, lightning, and possibly the photolysis of stratospheric nitrous oxide (N_2O, see e.g. Liley et al., 2000).

2.3 Measurement methods for NO_2

Various methods for measuring NO_2 exist, each with their respective advantages and disadvantages. The main method discussed here is *differential optical absorption spectroscopy* (DOAS), which is of fundamental importance for ground- and satellite-based remote sensing of NO_2 and other weak absorbers. In this section the DOAS principle is explained, following the work of Richter and Wagner (2011). Then, the NO_2 retrieval algorithm of the satellite instrument TROPOMI is summarized, following the product user manual (PUM, see Eskes

et al., 2022) and the Algorithm Theoretical Basis Document (ATBD, see van Geffen et al., 2022b). Lastly, different methods for NO_2 in situ measurements at the surface are discussed with particular attention to the molybdenum chemiluminescence method and its deficits.

2.3.1 Differential optical absorption spectroscopy (DOAS)

Differential optical absorption spectroscopy (DOAS) is a method for the spectral analysis of light which allows for the quantification of trace gases in the atmosphere (see e.g. Noxon, 1975; Platt et al., 1979). The idea behind DOAS is to utilize knowledge of the gases' *absorption cross sections* $\sigma(\lambda)$, which describe their wavelength-dependent light absorption strength. The resulting spectral absorption features can be measured with spectrometers, and the contributing trace gases can be identified by association with the known absorption cross sections. The DOAS method can be applied to direct and scattered light, either from natural sources (e.g. the sun, "passive DOAS") or artificial sources (e.g. a halogen lamp in a laboratory, "active DOAS").

The absorption of light is described by the Lambert-Beer law, which can be extended for atmospheric spectroscopic observations as

$$I(\lambda) = A(\lambda) I_0(\lambda) \cdot \exp\left(-\sum_k \sigma_k(\lambda) S_k\right) \tag{2.40}$$

where:

- λ : wavelength
- A : efficiency factor (see below)
- I : intensity spectrum after absorption
- I_0 : initial intensity spectrum
- σ_k : absorption cross section of absorber k (units: cm^2 $\mathrm{molec.}^{-1}$)
- S_k : column density of absorber k (units: molec. cm^{-2})

The unitless efficiency factor A is required in satellite-based measurements and ground-based measurements of scattered sunlight to account for the fact that I is a spectrum of scattered and reflected light, and depends e.g. on surface albedo and the amount of clouds in the light path. The *column density* S of an absorber is defined as

$$S = \int_L c(s)\,\mathrm{d}s \tag{2.41}$$

i.e. the integral of the absorber's concentration c along the light path L. In the general case this

light path is slant, and the column density is called *slant column density* (SCD). Additionally, light is attenuated by elastic and inelastic scattering. *Rayleigh scattering* (elastic, $\propto \lambda^{-4}$) refers to scattering on particles much smaller than the wavelength of the light (e.g. air molecules). *Mie scattering* (elastic, $\propto \lambda^{-\alpha}$, where α is called the Ångström coefficient) describes scattering on particles of about the size of the light's wavelength or above (e.g. raindrops). Inelastic scattering occurs in the form of *Raman* scattering, which changes the wavelength of the scattered light. If the sun is used as the light source, Raman scattering causes the Fraunhofer lines (strong absorption features due to elements in the sun's ionosphere) in the spectrum of scattered sunlight to be filled up. This is known as the *Ring effect* (Grainger and Ring, 1962). The combined influence of elastic and inelastic scattering can be accounted for by augmenting eq. (2.40) to

$$I(\lambda) = A(\lambda) I_0(\lambda) \cdot \exp\left(-\sum_k \sigma_k(\lambda) S_k + \epsilon_R(\lambda) + \epsilon_M(\lambda) + c_R R(\lambda) \right) \tag{2.42}$$

where:

$\epsilon_R(\lambda)$: Rayleigh extinction coefficient

$\epsilon_M(\lambda)$: Mie extinction coefficient

c_R : Ring coefficient, describes the strength of the Ring effect

$R(\lambda)$: Ring spectrum (see Solomon et al., 1987)

The trace gases' absorption cross sections can be separated into a broadband portion $\sigma^*(\lambda)$ and a narrowband portion $\sigma'(\lambda)$, i.e. $\sigma_k(\lambda) = \sigma_k^*(\lambda) + \sigma_k'(\lambda)$. Then, the contributions of the efficiency factor, the broadband absorption, and Rayleigh/Mie scattering can be described by a polynomial $P(\lambda)$, usually of order 2 – 5. Rearranging eq. (2.42) yields the *optical depth*

$$\tau(\lambda) := \ln\left(\frac{I_0(\lambda)}{I(\lambda)} \right) = \sum_k \sigma_k'(\lambda) S_k + c_R R(\lambda) + P(\lambda) \tag{2.43}$$

Assuming that I_0 and I are known, the column densities S_k can be obtained from eq. (2.43) by means of least-squares optimization. Here, the absorption cross sections are fitted to the measured optical depth, while varying the column densities, the Ring coefficient, and the polynomial coefficients as free fit parameters. I_0 should represent a light spectrum identical to I with exception of the attenuation terms covered in eq. (2.43). For satellite measurements usually a direct sun spectrum (without atmospheric absorption) is used as I_0. Alternatively, a spectrum of backscattered sunlight can also be used, taken over regions with negligible absorption of the absorber of interest.

Figure 2.8 shows a sketch of the typical satellite measuring geometry and a visualization of the DOAS fit procedure. Note that Fig. 2.8b refers to the the fit procedure of the TROPOMI satellite retrieval (described in more detail below in sect. 2.3.4), which slightly deviates from eq. (2.43) because it implements an intensity fit, i.e. the free fit parameters are fitted to the measured intensity instead of the optical depth. The procedure is mathematically identical and described in van Geffen et al. (2020).

Instrument response function

The high-resolution cross sections taken from literature must be convolved with the instrument response function (IRF) $H(\lambda)$ of the spectrometer in order to account for differences in spectral resolution. The convolution is defined as

$$(I * H)(\lambda) = \int I(\lambda')H(\lambda - \lambda')\,\mathrm{d}\lambda' \tag{2.44}$$

and analogously applied to I_0. The instrument response of spectrometers can change over time, and may become impossible to measure (i.e. on a satellite after launch, see Munro et al., 2016). In this case there exist workarounds to determine the IRF, e.g. by modelling it as a pseudo-absorber (see e.g. Beirle et al., 2017).

For more detailed elaborations on advanced aspects of DOAS the reader is referred to Wagner et al. (2001), Puķīte et al. (2010), Platt and Stutz (2008), Richter and Wagner (2011), and van Geffen et al. (2022b).

2.3.2 The air mass factor

In the general case, spectroscopic measurements in the atmosphere receive light from a *slant* (as opposed to strictly vertical) light path. The corresponding column densities are referred to as *slant column densities* (SCDs). The prefix "d" is used to denote a *differential* SCD (dSCD). In many situations, the *vertical column density* (VCD), meaning the vertically integrated concentration, is the more informative quantity. For example, SCDs depend on the solar zenith angle and the associated elongation of light paths, while VCDs can be interpreted independently of the sun geometry. The *air mass factor* (AMF, A) is defined as the ratio of SCD (S) and VCD (V):

$$A = \frac{S}{V} \tag{2.45}$$

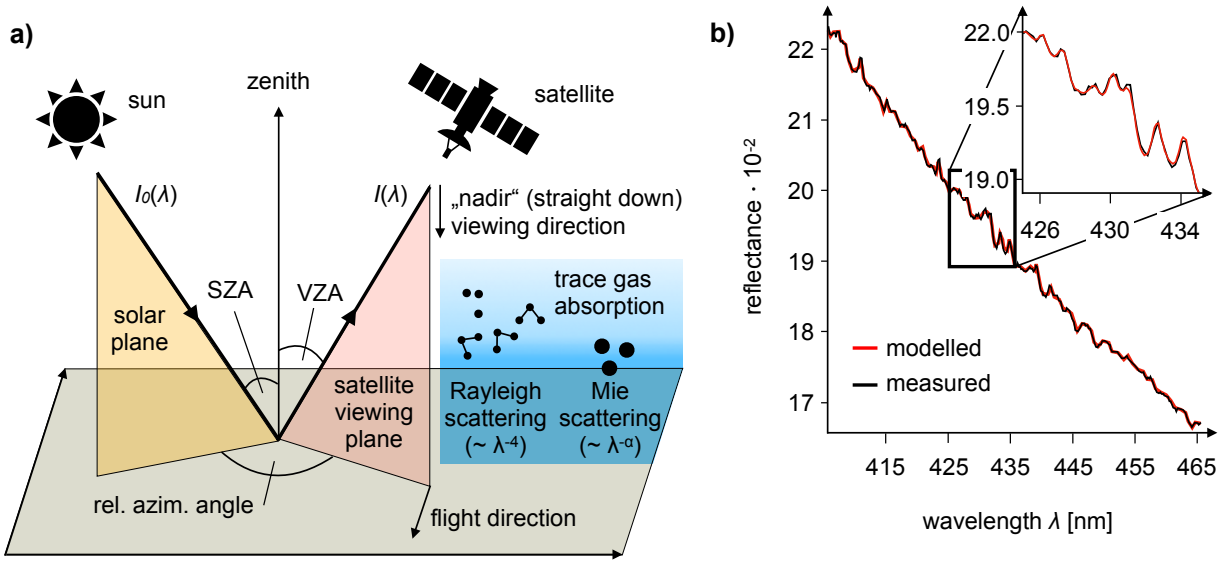

Figure 2.8: Application of the DOAS method in satellite measurements of atmospheric trace gases. a) Overview of the typical measurement geometry. SZA and VZA refer to the *solar zenith angle* and *viewing zenith angle*. The satellite viewing plane is orthogonal to the direction of flight. **b)** Exemplary intensity fit in the TROPOMI NO_2 retrieval. The black line shows a measured reflectance spectrum, recorded on 4 July 2018 in a polluted region of Rotterdam, Netherlands. The red line shows the corresponding modelled reflectance spectrum. Data taken and re-plotted from van Geffen et al. (2022b).

In order to convert SCDs to VCDs, the AMF must be determined. In the satellite measurement geometry, the air mass factor can be approximated for stratospheric absorbers as

$$A \approx \frac{1}{\cos(\mathrm{SZA})} + \frac{1}{\cos(\mathrm{VZA})} \quad \text{valid for} \quad \mathrm{SZA} < 80° \tag{2.46}$$

where SZA and VZA denote the *solar zenith angle* and *viewing zenith angle*, respectively. The approximation may extend to parts of the troposphere at larger wavelengths. In reality, particularly in satellite measurements, the AMF depends on many more factors, such as wavelength, surface albedo and pressure, trace gas profiles, clouds, and aerosols. In such cases, advanced retrieval algorithms can resort to the use of radiative transfer models (such as McArtim, see Deutschmann et al., 2011). The computation of AMFs within the TROPOMI NO_2 retrieval is explained in detail in sect. 2.3.4.

2.3.3 The TROPOMI satellite instrument

The TROPOMI satellite instrument is a hyperspectral imaging spectrometer on board of the Sentinel-5 Precursor (S-5P) satellite. S-5P was launched in October 2017 and is operated in a low-earth polar orbit at roughly 824 km altitude. TROPOMI uses four detectors, split

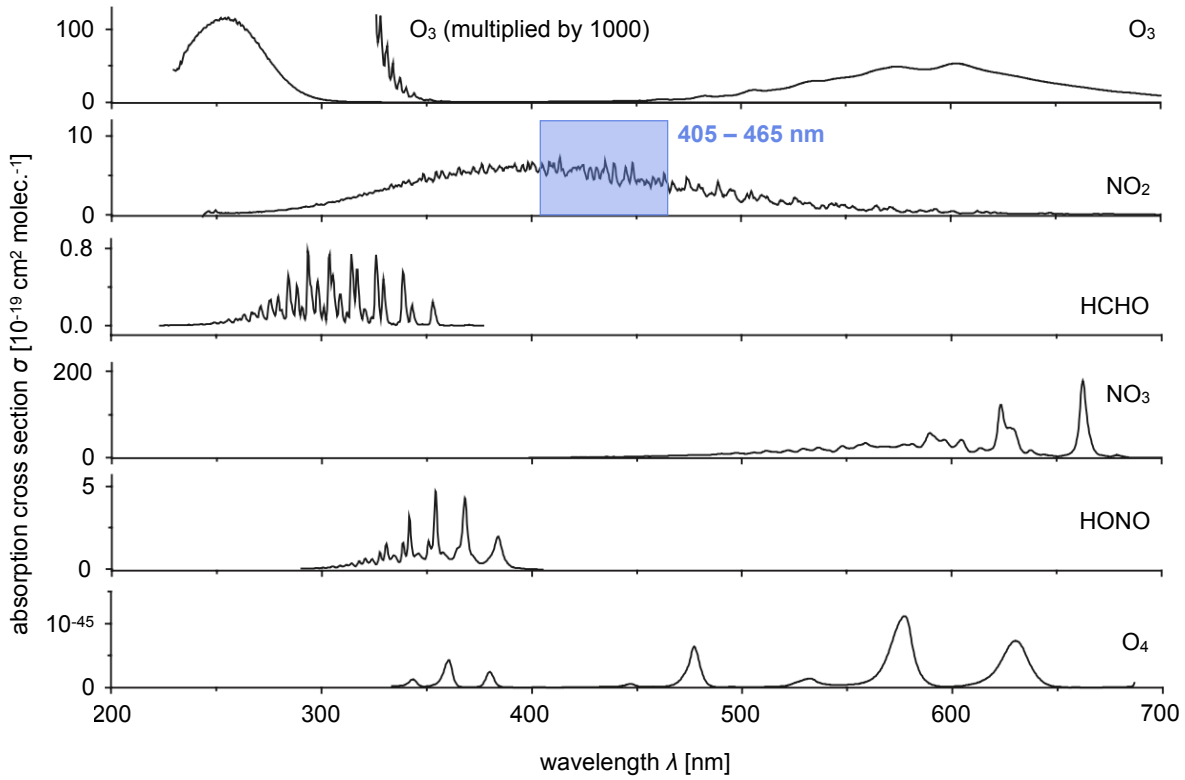

Figure 2.9: Absorption cross sections of selected atmospheric trace gases. Adapted from Platt and Stutz (2008). The blue region from 405 – 465 nm marks the fit range of the TROPOMI NO_2 retrieval. The absorption cross section of O_4 is given in units of cm^5 $molec.^{-2}$ because O_4 is a O_2-O_2 collision complex, and thus the O_4 concentration is expressed as the squared O_2 concentration.

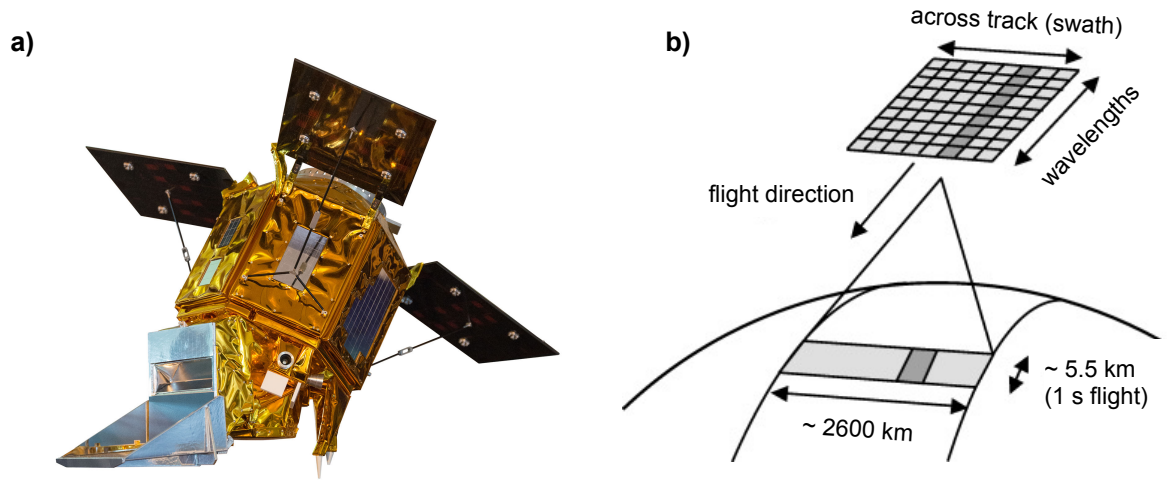

Figure 2.10: TROPOMI onboard the Sentinel-5P satellite. a) Picture of Sentinel-5P, carrying the TROPOMI instrument. Taken from Wikipedia contributor SkywalkerPL (2023). **b)** Measurement principle of the TROPOMI instrument. Adapted from Veefkind et al. (2012).

detector	UV[(1)]		UV-vis[(2)]		NIR[(3)]		SWIR[(4)]
band	1	2	3	4	5	6	7
spectral range [nm]	$270-300$	$300-320$	$310-405$	$405-500$	$675-725$	$725-775$	$2305-2385$
spectral resolution [nm]	1.0	0.5	0.55	0.55	0.5	0.5	0.25
spectral sampling [nm]	0.065	0.065	0.2	0.2	0.1	0.1	< 0.1

Table 2.2: Overview of TROPOMI's detectors and spectral bands. The abbreviations in the table's headers refer to the following spectral regions:
(1) ultraviolet
(2) ultraviolet-visible
(3) near infrared
(4) shortwave infrared

into seven spectral bands. An overview is given in Table 2.2. For NO_2 measurements the ultraviolet-visible (UV-vis) detector is used. Further descriptions are given in Veefkind et al. (2012). TROPOMI scans Earth's surface in a push-broom manner (see Fig. 2.10b). In this procedure, the earthshine radiance of a 2600 km $\times$ 5.5 km swath (a "stripe" of Earth's surface) is projected onto a 2-dimensional detector. One detector dimension is used to resolve the measurement across the track, and the other is used for spectral resolution (obtained by a grating). Note, that previous to 6 August 2019, the instrument was operated with a swath length of 7 km instead of 5.5 km. The resulting pixel size (often referred to as the "horizontal resolution" of the instrument) depends on viewing geometry: The across-track resolution reaches from 3.5 km in *nadir* geometry (VZA = 0), to up to 14.4 km km at the swath edges. By flying in a low Earth sun-synchronous orbit, daily coverage of the Earth with a total of 14 – 15 daily orbits is achieved. The orbital repetition cycle spans 16 days. The satellite equator overpass occurs at $\sim$ 13:30 mean local solar time.

Altogether, the TROPOMI satellite instrument is suitable to retrieve daily global maps of trace gas column densities with a horizontal resolution on the scale of ~ 20 km^2. This marks a serious improvement over preceding satellite instruments, whose horizontal resolution was far lower (e.g. OMI (24 km $\times$ 13 km), SCIAMACHY (60 km $\times$ 30 km), GOME-2 (80 km $\times$ 40 km), and GOME (320 km $\times$ 40 km)). The drastically enhanced spatial resolution of TROPOMI allows to resolve trace gas distributions on an urban to sub-urban scale, on which many recent studies were based, e.g. on NO_x point source emissions (e.g. Beirle et al., 2021), volcanic plume composition (e.g. Warnach, 2022), NO_2 plumes from individual ships (e.g. Georgoulias et al., 2020; Riess et al., 2022), and many more.

2.3.4 The TROPOMI NO_2 retrieval

The TROPOMI retrieval algorithm for NO_2 (and other trace gases) relies on the DOAS method, but includes many more operational steps than just a DOAS fit. The main data

product used in this thesis is the tropospheric NO_2 VCD (v02.04.00), which further requires the separation of stratospheric and tropospheric SCDs and the conversion of tropospheric SCDs to tropospheric VCDs. The full retrieval process is documented in the PUM (Eskes et al., 2022) and the ATBD (van Geffen et al., 2022b), of which the most important aspects are summarized in the following. Throughout this chapter, TROPOMI's NO_2 retrieval is referred to as "the retrieval".

Separation of stratospheric and tropospheric NO_2

By application of the DOAS method to the light spectra measured by TROPOMI, a total NO_2 slant column $S = S_s + S_t$ is obtained (here and in the following subscript "s" and "t" indicate the stratospheric and tropospheric portion of a variable), see also van Geffen et al. (2020). This total SCD is separated into a stratospheric and a tropospheric part. In this procedure, the global chemistry and transport model TM5-MP (see Williams et al., 2017) is used to simulate the coupled troposphere-stratosphere distribution of NO_2. The global model assimilates TROPOMI's SCDs over areas known to contain little or no tropospheric NO_2. The stratospheric vertical column density V_s can then be obtained as the integral from the model's tropopause layer to its highest layer.

Computation of the air mass factor and vertical column densities

The conversion between slant and vertical column densities requires knowledge of the AMF. The AMF depends on the vertical trace gas profile, and can be written as

$$A = \frac{\sum_l A_l V_l c_l}{\sum_l V_l} \tag{2.47}$$

where:

- A_l : the "box AMF", i.e. the AMF of a single model layer
- V_l : the "box VCD", i.e. the column density of a single model layer
- c_l : a temperature correction term, see van Geffen et al. (2022b)
- l : the layer index

Here, "model" refers to TM5-MP. The altitude-dependent box AMFs are taken from a pre-calculated look-up table, see Lorente et al. (2017). For this procedure, the Doubling-Adding KNMI radiative transfer model (de Haan et al., 1987) is operated in an atmosphere with hypothetical NO_2 profiles. Then, the ratio between the excess NO_2 slant column δS and the

vertical column δV_l added to each layer is computed. This corresponds to the box AMFs, i.e.

$$A_l = \delta S / \delta V_l \tag{2.48}$$

The altitude-dependent AMFs are stored in a 6-dimensional look-up table, depending on the solar zenith angle, the viewing zenith angle, the relative azimuth angle, the surface albedo and pressure, and the midlevel atmospheric pressure. Combined with eq. (2.47) and the stratospheric VCD V_s obtained from vertical summation of the box VCDs from the TM5-MP model, this yields the stratospheric slant column:

$$S_s = V_s \cdot A_s = \sum_{l=l_{\mathrm{tp}}+1}^{N} V_l \cdot \frac{\sum_{l=l_{\mathrm{tp}}+1}^{N} A_l V_l c_l}{\sum_{l=l_{\mathrm{tp}}+1}^{N} V_l} = \sum_{l=l_{\mathrm{tp}}+1}^{N} A_l V_l c_l \tag{2.49}$$

where:

l_{tp} : the tropopause layer of the TM5-MP model

N : the total number of layers in the TM5-MP model (currently: $N_l = 34$)

Here, the summation limits are chosen such that only stratospheric layers are considered. Then, the slant and vertical tropospheric columns are computed as

$$S_t = S - S_s \tag{2.50}$$

$$V_t = S_t / A_t \tag{2.51}$$

Each VCD in the TROPOMI retrieval has an associated quality score ("qa-value") $0 \leq f_{\mathrm{QA}} \leq 1$, which describes the quality of the retrieval. The PUM recommends to use a lower threshold of $f_{\mathrm{QA}} > 0.75$, which also removes scenes with a cloud fraction of 0.5 or higher.

Averaging kernels, comparison to model data, and alternative satellite products

The *averaging kernels* (AKs) of the satellite retrieval are used to describe the instrument's sensitivity to different vertical layers of the atmosphere. The total and tropospheric AKs are defined as

$$K_l = A_l / A \qquad K_{l,t} = \begin{cases} K_l \cdot A / A_t & l \leq l_{\mathrm{tp}} \\ 0 & l > l_{\mathrm{tp}} \end{cases} \tag{2.52}$$

The AKs become relevant in the validation of NO_2 VCDs obtained from e.g. RCT simulations with TROPOMI's tropospheric NO_2 VCDs, where it is required to account for the possibly reduced sensitivity of the satellite measurement by replacing the TM5-MP a priori NO_2 profiles used in the retrieval with those from the RCT simulation. This produces new tropospheric NO_2 VCDs and air mass factors via the equations

$$V_t' = \frac{A}{A'} V_t \tag{2.53}$$

$$A' = A \cdot \frac{\sum_{l=0}^{l_{\mathrm{tp}}} K_{l,t} V_l}{\sum_{l=0}^{l_{\mathrm{tp}}} V_l} \tag{2.54}$$

where:

V_t' : the new tropospheric VCD

A' : the new air mass factor

V_l : the box VCD of layer l (taken from the RCT simulation)

Here, "new" refers to the quantities obtained upon exchange of the a priori profiles. The main benefit of replacing the TM5-MP a priori profiles is the following: The TM5-MP model operates globally with a horizontal resolution of $1° \times 1°$, corresponding to approximately 70 km $\times$ 110 km (longitude $\times$ latitude) in European latitude regions. This resolution is orders of magnitude below that of the satellite measurement, and often leads to averaging over regions with both high and low pollution levels (e.g. cities and nearby rural regions). A representative example is shown in Fig. 2.11, where a single TM5-MP model cell covers an area including several German cities (e.g. Mannheim, where strong point sources in the form of a coal power plant (Großkraftwerk Mannheim) and a chemical production site of the BASF company are located), and far less polluted agricultural regions in-between. Obviously, the average NO_2 profile over this large region is by no means representative for the individual pixels of the far better resolved satellite measurements. Consequently, the a priori profiles of the retrieval should be replaced by high-resolution alternatives whenever possible. Alternative profiles are usually obtained from RCT models, running on a higher horizontal resolution than TM5-MP. Based on higher resolved a priori profiles, alternative regional NO_2 satellite products have emerged, such as described by Liu et al. (2021) with $0.3° \times 0.2°$ horizontal resolution, or Douros et al. (2023) with $0.1° \times 0.1°$ horizontal resolution. The exchange of a priori profiles generally results in increased tropospheric NO_2 VCDs by approximately 15 %. The effect is strongest in polluted regions.

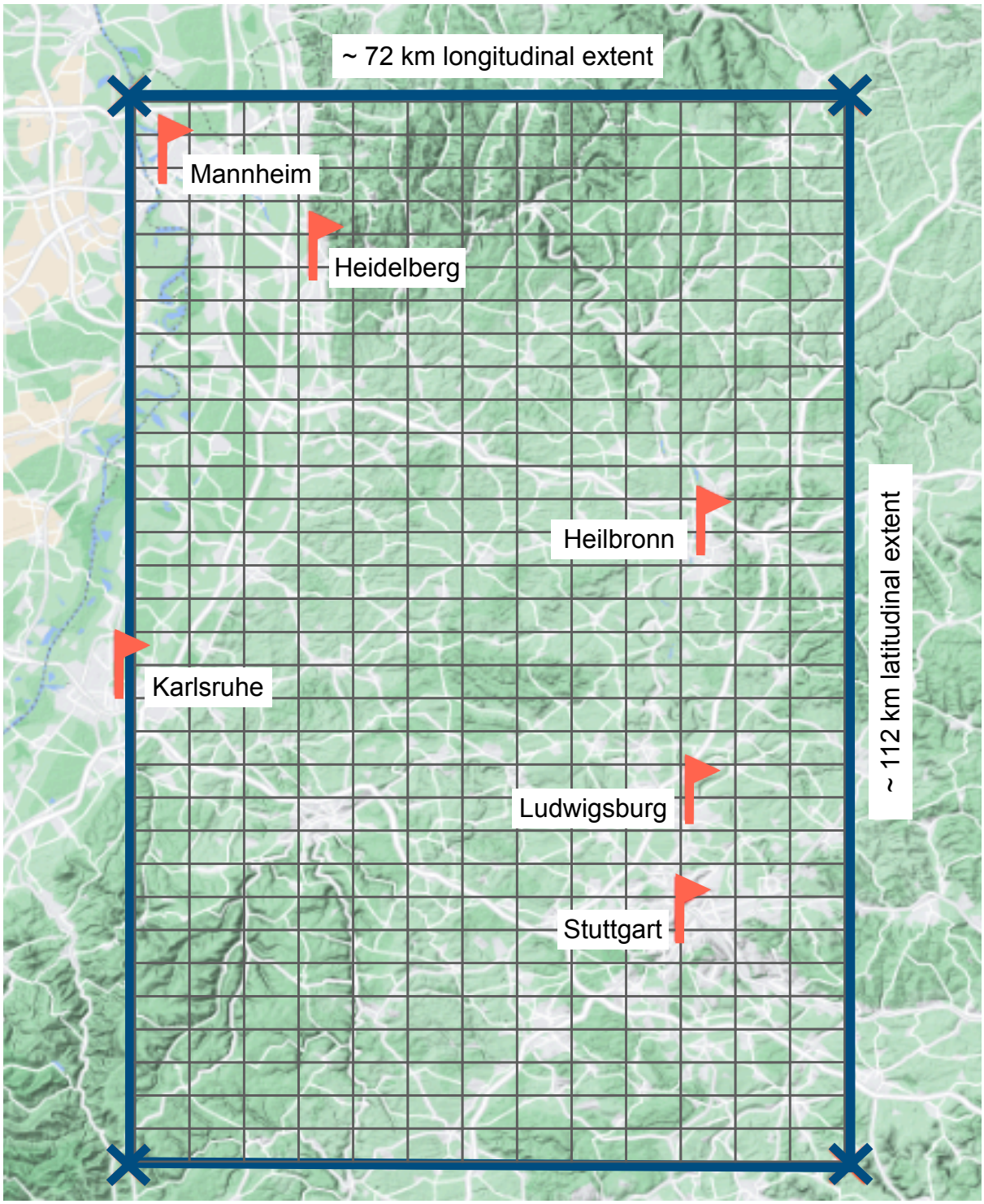

Figure 2.11: Illustration of the TM5-MP horizontal model resolution over a region in south-west Germany. The large red rectangle represents the resolution of a single TM5-MP grid cell. The gray lattice represents the resolution of the TROPOMI measurements (here: assuming a resolution of 5.5 km $\times$ 3.5 km, e.g. close to nadir viewing geometry). Maps Data: © 2024 GeoBasis-DE/BKG (© 2009), Google.

Overview of the different TROPOMI NO_2 product versions

The first version of the retrieval, v01.00.00, went into operation on 28 June 2018. Since then significant improvements to the retrieval code were achieved. The current processor version is v02.04.00. The most relevant changes occurred in v01.04.00, and v02.02.00. In v01.04.00, changes to the cloud retrieval algorithm were made, which resulted in a significant decrease in assumed cloud pressure. As a result the tropospheric NO_2 VCD increased by up to 50 % over polluted regions. Processor version v02.02.00 saw changes to the used irradiance data, further updates of the cloud pressure estimation introduced in v01.04.00, and a correction to the surface albedo in cloud-free scenes (resulting in approximately 15 % increased NO_2 VCDs over polluted regions). Overall, the transition from processor versions v01.x to v02.02.00 were associated with a 10 % to 40 % increase in tropospheric columns, depending on season and pollution levels (see van Geffen et al., 2022a).

Validation measurements and uncertainties of the NO_2 retrieval

To analyze tropospheric NO_2 VCDs from satellite measurements is a complex endeavor. The mission requirements ask for < 50 % total bias, and < 0.7 Pmolec. cm^{-2} dispersion. While the bias requirement is considered satisfied, the dispersion requirement is partially exceeded (see Compernolle, 2023). The retrieval aspects described above leave room for uncertainties, particularly in the stratosphere-troposphere separation, the computation of air mass factors (related to the low resolution of the TM5-MP model), and the assessment of cloud coverage and pressure. The overall uncertainty of the retrieved NO_2 VCDs is on the order of $30-60$ %, of which around $20-50$ % can be attributed to the uncertainty of the tropospheric AMF (Liu et al., 2021). TROPOMI's tropospheric NO_2 VCDs were validated in multiple studies, of which a selection is summarized in the following:

- Griffin et al. (2019) validate the v01.01.00 data product against aircraft measurements, Pandora instruments (ground-based spectrometers, which measure NO_2 VCDs), and tropospheric NO_2 columns from the Ozone Monitoring Instrument (OMI), one of TROPOMI's predecessors. An overall bias of -15 % to -30 % was determined. Upon use of higher-resolved a priori profiles, the bias was reduced to 0 % to -25 %.

- Judd et al. (2020) present a validation study of the v01.02.00 data product against airborne spectrometers and Pandora instruments. A bias of -19 % to -33 % was determined in the period from June to September 2018 and partially attributed to the resolution of the TM5-MP a priori profiles. Upon exchange with profiles from a higher resolved RCT model, the bias was reduced to -9 % to -19 %.

- Tack et al. (2021) evaluate TROPOMI v01.03.01 data against airborne observations, resulting in a bias of −14 %. Exchange of a priori profiles reduced the bias to −1 %.

- Ialongo et al. (2020) conduct a validation study on total NO_2 columns of the v01.02.02 data product against a single Pandora instrument in Helsinki, Finland. TROPOMI showed overestimations for small total columns, but underestimations for large total columns. The authors attribute the deviations to a possible overestimation of the stratospheric column, and the low resolution of the TM5-MP model.

- Liu et al. (2021) present an improved (inofficial) TROPOMI NO_2 data product, implementing e.g. the Stratospheric Estimation Algorithm from Mainz (STREAM) for the estimation of stratospheric NO_2. The new data product was evaluated against ground-based spectroscopic measurements. While a previous data product of the authors showed a bias of −55.3 %, the improved product's bias was reduced to −34.7 %. Exchange of a priori profiles further reduced the bias to −20 %.

- van Geffen et al. (2022a) validate the v02.02.00 data product against ground-based spectroscopic measurements of tropospheric, stratospheric and total NO_2 columns. The authors mention a tropospheric column bias of up to −51 % in polluted regions for the processor versions v01.02.00 and v01.03.00. Due to several retrieval improvements leading to v02.02.00, the average tropospheric column bias was reduced to −23 %.

- Lange et al. (2023) conduct a study on the v01.03.02 and v02.03.01 data products against airborne, ground-based and mobile car DOAS measurements. For v01.03.02, similar biases as in previous studies, of up to −50 % in polluted regions, were found. Interestingly, the v02.03.01 data show an overall positive bias of +20 %, and exchange of a priori profiles was found to have no significant effect on the bias of the v01.03.02 data. This finding is qualitatively different to the previous validation studies, but may be linked to the relatively small study region comprised of only $\sim$ 100 TROPOMI pixels.

The results of these studies can be summarized as follows: TROPOMI's tropospheric NO_2 retrieval suffers from low biases on the scale of approx. −20 % to −30 %, depending on time and location of the validation studies. Exchange of a priori profiles can reduce the bias by approximately 15 %. Furthermore, significant improvements were achieved in the transition to v02.x.

2.3.5 Multi-Axis DOAS (MAX-DOAS)

In Multi-Axis DOAS (MAX-DOAS, see Hönninger et al., 2004), the DOAS principle is extended from ground-based measurements at a single fixed elevation angle to measurements, which

cycle through different elevation angles. The goal is to reconstruct the chemical composition of different layers of Earth's atmosphere by retrieval of trace gas *profiles* (as opposed to single differential column densities from standard DOAS). A typical elevation angle sequence could be $1°, 2°, 3°, 4°, 5°, 6°, 8°, 15°, 30°, 90°$. Measurements at very low elevation angles are often obscured by buildings, tree canopies, or mountains. A selection of publications which have utilized MAX-DOAS for the assessment of tropospheric trace gas profiles are Hönninger et al. (2004), Frieß et al. (2006), Kumar et al. (2020), Chan et al. (2020), and Hendrick et al. (2014).

What adds considerable complexity to MAX-DOAS evaluations are the required inversion algorithms. In contrast to standard DOAS, MAX-DOAS aims to reconstruct trace gas profiles, i.e. trace gas concentrations as a function of altitude. The conversion of dSCD sequences from different elevation angles to concentration profiles is non-trivial and oftentimes associated with considerable computational effort. An intercomparison study of MAX-DOAS retrieval algorithms by Tirpitz et al. (2021) revealed relative retrieval uncertainties between 3 % and 70 %. Over the last two decades, two classes of MAX-DOAS retrievals have established: Parametrized algorithms (such as the *Mainz Profile Algorithm*, MAPA, see Beirle et al., 2019), which typically find solutions in pre-calculated look-up tables with a Monte-Carlo method, and *optimal-estimation* (OE) based algorithms (such as the *Mexican MAX-DOAS Fit*, MMF, see Friedrich et al., 2019), which make use of online forward modelling in the framework of Bayesian statistics. Two exemplary algorithms, which are used for MAX-DOAS retrieval later in this thesis, are outlined in the following.

Mainz profile algorithm (MAPA)

The *Mainz profile algorithm* (MAPA, see Beirle et al., 2019) is a parametrized MAX-DOAS retrieval. Within MAPA, trace gas profiles are described by three parameters:

1. the integrated column c (corresponding to the VCD for trace gases, and the *aerosol optical depth* for aerosols)

2. the layer height h, which can be understood as a representation of the PBLH

3. a shape parameter $s \in (0, 2)$

A shape parameter of $s = 1$ corresponds to a simple box profile, i.e.

$$p(z) = \begin{cases} c/h & z \leq h \\ 0 & z > h \end{cases} \tag{2.55}$$

where z denotes altitude and $p(z)$ the corresponding concentration values. For a shape parameter of $s < 1$, the fraction s of the total column c is modelled as a box profile and the remaining fraction $1 - s$ as an exponential tail:

$$p(z) = \begin{cases} s \cdot c/h & z \leq h \\ \exp\left(-\frac{z-h}{h} \cdot \frac{s}{1-s}\right) \cdot s \cdot c/h & z > h \end{cases} \tag{2.56}$$

A shape parameter $s > 1$ is used to express an elevated layer, starting at altitude h_1, with vertical extent h_2, reaching up to h:

$$p(z) = \begin{cases} 0 & z < h_1 \\ c/h_2 & h_1 < z \leq h \\ 0 & z > h \end{cases} \tag{2.57}$$

with

$$h_1 = (s-1) \cdot h \tag{2.58}$$

$$h_2 = (2-s) \cdot h \tag{2.59}$$

$$h = h_1 + h_2 \tag{2.60}$$

Some typical resulting profiles are displayed in Fig. 2.12. Given a sequence of m elevation angles with corresponding dSCDs (and a few other inputs, e.g. the presumed O_4 VCD), MAPA returns the parameter values c, h, and s of a trace gas (or aerosol) profile in optimal corresponding agreement. As described in sect. 2.3.2, the vertical column is linked to dSCD via dSCD = VCD $\cdot A$. Here, A denotes the differential air mass factor (dAMF). Optimal values of c are obtained using a linear fit of the form

$$c = \frac{\boldsymbol{S}_x \cdot \boldsymbol{A}}{\boldsymbol{A} \cdot \boldsymbol{A}} \tag{2.61}$$

where:

$\boldsymbol{S}_x$: the measured dSCD sequence

$\boldsymbol{A}$: the vector of dAMFs

i.e. the optimal value for c is determined by an ensemble of c-estimates at single elevation angles, weighted by the respective dAMF. The dAMFs have been computed beforehand using

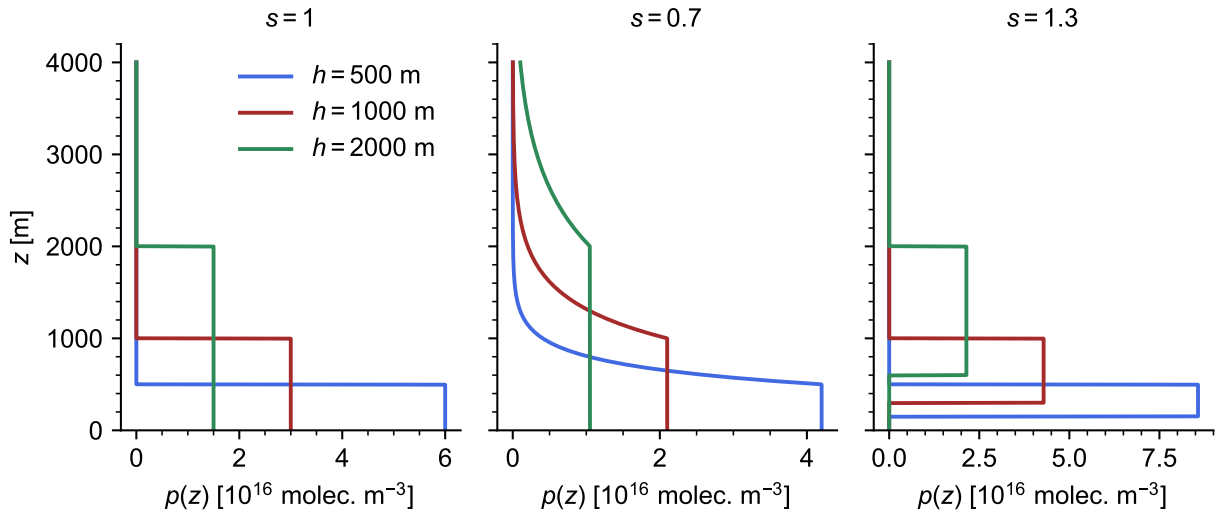

Figure 2.12: Profile shapes of the MAPA retrieval. Different values of h (500 m, 1000 m, 2000 m) are colour coded. Each subplot shows a different choice of the shape parameter s (1, 0.7, 1.3). All plots use $c = 3 \cdot 10^{15}$ molec. cm^{-2}. Adapted from Beirle et al. (2019).

the McArtim radiative transfer model (see Deutschmann et al., 2011), and stored in a look-up table. This look-up table links the remaining profile parameters to corresponding dSCD sequences. Optimal values of h and s are determined in a Monte-Carlo approach, i.e. by drawing random samples from the parameter space and dismissing parameter combinations, whose dSCD sequence (from the look-up table) disagrees with the measured dSCD sequence. Agreement is quantified by the RMSE:

$$R = \sqrt{\frac{(\boldsymbol{S}_y - \boldsymbol{S}_x)^2}{m}} \tag{2.62}$$

where $\boldsymbol{S}_y$ denotes the dSCD sequence from the look-up table. Either the single best match, or an ensemble of up to n parameter triples with $R/R_{\mathrm{bm}} < F$ is identified as the solution, where R_{bm} denotes the RMSE of the best match, and F a user-chosen threshold (standard value: $F = 1.3$). From such an ensemble, the weighted mean, along with further statistical diagnostics, can be computed.

Finally, the role of the O_4 VCD and its scaling factors shall be briefly discussed: The trace gas profiles obtained from MAPA depend on the dAMFs, which depend on the aerosol profiles. The O_4 dAMFs are derived from the measured O_4 dSCDs by division by the O_4 VCD. The O_4 VCD can either be provided by the user, or computed from temperature and pressure profiles. As summarized in Wagner et al. (2019) (including the peer review of that article, which is openly accessible), measurements of O_4 often show larger absorption than radiative transfer simulations predict. The effect is disputed (see e.g. Ortega et al., 2016),

and a comprehensive explanation is still missing. Nonetheless, O_4 scaling factors of $f < 1$ are often applied in order to compensate for the effect. MAPA provides the option to use a fixed O_4 scaling factor, e.g. $f = 0.8$. Alternatively, the "optimal" scaling factor can be determined in a fit routine in analogy to eq. (2.61).

MAPA provides a flagging scheme to identify possibly erroneous retrievals. Each profile can be either unflagged, or flagged with "warning" or "error".

Mexican MAX-DOAS fit (MMF), and optimal estimation (OE)

The *Mexican MAX-DOAS fit* (MMF, see Friedrich et al., 2019) is an *optimal estimation* (OE, see Rodgers, 2000) based MAX-DOAS retrieval algorithm. Here the retrieval problem is formulated as follows:

A measurement vector $\boldsymbol{y}$ is produced by a forward process

$$\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{b}) + \boldsymbol{\epsilon} \tag{2.63}$$

where:

- $\boldsymbol{x}$: the atmospheric state vector
- $\boldsymbol{b}$: further parameters of the forward process
- $\boldsymbol{\epsilon}$: the measurement error

The goal is to obtain the state vector $\boldsymbol{x}$ (the trace gas profile) from the observations $\boldsymbol{y}$ (the dSCD sequence). The inverse problem stated in eq. (2.63) is usually ill-posed, meaning that no well-defined inverse F^{-1} exists, for which $\boldsymbol{x} = F^{-1}(\boldsymbol{y})$. In the OE framework, the optimal solution is obtained by minimization of the cost function

$$\chi^2(\boldsymbol{x}) = (\boldsymbol{y} - F(\boldsymbol{x}, \boldsymbol{b}))^{\mathrm{T}} \mathbf{S}_\epsilon^{-1} (\boldsymbol{y} - F(\boldsymbol{x}, \boldsymbol{b})) + (\boldsymbol{x} - \boldsymbol{x}_a)^{\mathrm{T}} \mathbf{S}_a^{-1} (\boldsymbol{x} - \boldsymbol{x}_a) \tag{2.64}$$

where:

- $\mathbf{S}_\epsilon$: the measurement covariance matrix
- $\boldsymbol{x}_a$: the a priori state vector
- $\mathbf{S}_a$: the a priori covariance matrix

If the measurements are independent of each other, $\mathbf{S}_\epsilon$ is a diagonal matrix, whose entries denote the squared measurement errors. The a priori constraints contain the "best knowledge" of the atmosphere prior to conducting the measurement. For example, in the case of the

MMF retrieval, the a priori information is taken from a WRF-Chem simulation over Mexico in 2011. The optimal solution ("maximum a posteriori", MAP) is the one, that minimizes the χ^2 problem defined by eq. (2.64):

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \chi^2(\boldsymbol{x}) \tag{2.65}$$

To find $\hat{\boldsymbol{x}}$ is not trivial. Note, that F represents a costly radiative transfer code, which should be run as few times as possible. In the case of MMF, $\hat{\boldsymbol{x}}$ is obtained using an iterative Gauss-Newton scheme, of the form

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_a + \left(\mathbf{J}_i^{\mathrm{T}}\mathbf{S}_\epsilon^{-1}\mathbf{J}_i + \mathbf{S}_a^{-1}\right)^{-1}\mathbf{J}_i^{\mathrm{T}}\mathbf{S}_\epsilon \cdot \left[(\boldsymbol{y} - F(\boldsymbol{x}_i)) - \mathbf{J}_i\left(\boldsymbol{x}_a - \boldsymbol{x}_i\right)\right] \tag{2.66}$$

where:

i : the iteration index

$\mathbf{J}_i$: the Jacobian of the forward process, defined as $(\mathbf{J}_i)_{jk} = \frac{\partial F(\boldsymbol{x}_{i,j})}{\partial \boldsymbol{x}_{i,k}}$

In contrast to MAPA, MMF can return averaging kernels, defined as

$$\mathbf{K} = \frac{\partial \hat{\boldsymbol{x}}}{\partial \boldsymbol{x}} \tag{2.67}$$

i.e. the sensitivity of the inversion result to the true atmospheric state. In contrast to the TROPOMI retrieval (see sect. 2.3.4), where the AKs were represented as a vector, they are represented as a $m \times m$ matrix here, where m denotes the number of atmospheric layers assumed in the retrieval. The i-th row of $\mathbf{K}$ denotes the sensitivity of the retrieved trace gas concentration in layer i with respect to all m layers. An ideal measurement is characterized by $\mathbf{K} = \mathbb{1}$, where $\mathbb{1}$ denotes the unity matrix. The solution $\hat{\boldsymbol{x}}$ can be understood as a weighted mixture of the a priori vector $\boldsymbol{x}_a$ and the true state vector $\boldsymbol{x}$, as in

$$\hat{\boldsymbol{x}} = \boldsymbol{x}_a + \mathbf{K}(\boldsymbol{x} - \boldsymbol{x}_a) = \mathbf{K}\boldsymbol{x} + (\mathbb{1} - \mathbf{K})\boldsymbol{x}_a \tag{2.68}$$

When comparing NO_2 profiles from an RCT model to MAX-DOAS measurements, the AKs must be applied to the model profiles via

$$\hat{\boldsymbol{x}}_m = \mathbf{K}\boldsymbol{x}_m + (\mathbb{1} - \mathbf{K})\boldsymbol{x}_a \tag{2.69}$$

where:

$\hat{\boldsymbol{x}}_m$: the model profiles with AKs applied

$\boldsymbol{x}_m$: the original model profiles

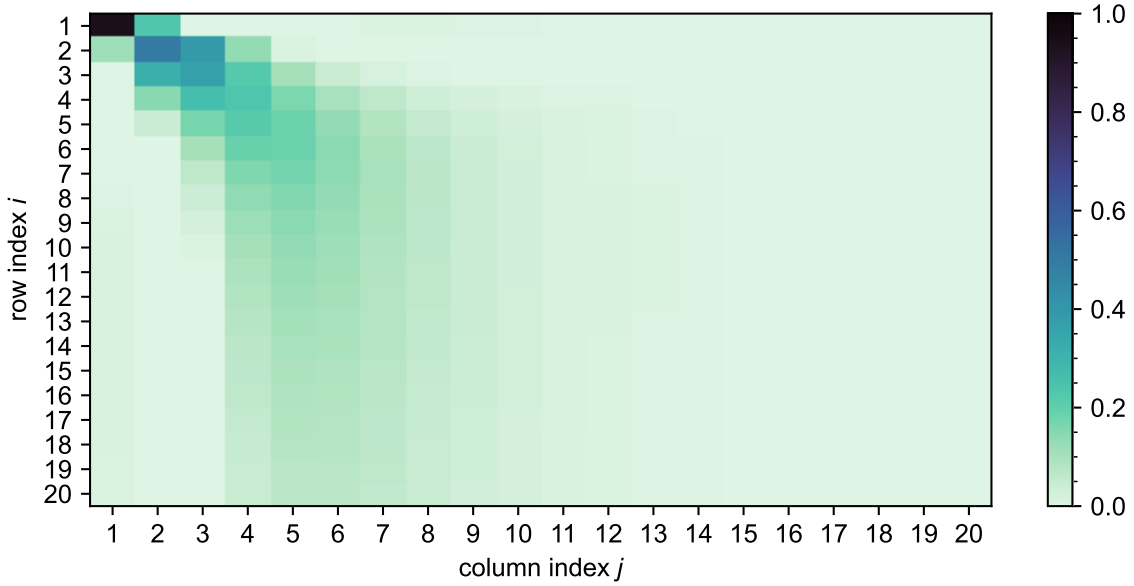

Figure 2.13: Example of a MAX-DOAS averaging kernel matrix. Shown here are values from a MAX-DOAS instrument in Heidelberg, Germany, averaged for the month of May 2022. Each entry denotes the retrieval sensitivity of one layer (indexed by *i*) to the concentrations in all other layers (indexed by *j*). The rows and columns are ordered such that the index 1 represents the lowest layer of the retrieval and index 20 the highest. The layer extent is 200 m. Data taken from FRM_4DOAS (Fayt et al., 2021).

The averaging kernels can be used to assess the sensitivity of the measurement. A typical observation is that MAX-DOAS measurements show significant sensitivity to the lowest $1-2$ km. Further above, the AKs quickly drop off. An example is shown in Fig. 2.13. For further reading, the reader is referred to Friedrich et al. (2019), Rodgers (2000), Frieß et al. (2019), and Bösch (2019), which also list several OE-based alternatives to MMF.

The FRM_4DOAS project

The *Fiducial Reference Measurements for Ground-Based DOAS Air-Quality Observations* (FRM_4DOAS, see Fayt et al., 2021) is a collaborative project, which aims at collecting measurements from MAX-DOAS instruments installed in different locations across the globe. The spectral data are then evaluated in a standardized fashion, using the MAPA and MMF retrieval algorithms. The retrievals are conducted on a vertical lattice with 200 m spacing, reaching from the surface up to 4 km altitude. An overview of the instrument locations used in this thesis is found in Table 2.3. Note, that some stations are not continuously available (e.g. data from Mainz, Germany are not available for May 2022). Individual stations can operate multiple spectrometers, e.g. for different azimuthal viewing directions.

location	lat. [°N]	lon. [°E]	operated by
Bremen, Germany	53.10	8.85	IUP-UB[(1)]
Mainz, Germany	49.99	8.23	MPIC[(2)]
Heidelberg, Germany	49.42	8.67	IUP-HD[(3)]
San Pietro Capofiume, Italy	44.65	11.62	ISAC-CNR[(4)]
Uccle, Belgium	50.80	4.36	BIRA-IASB[(5)]
De Bilt, Netherlands	52.10	5.18	KNMI[(6)]
Cabauw, Netherlands	51.97	4.93	KNMI[(6)]

Table 2.3: Overview of the geolocations and operators of the seven FRM$_4$DOAS MAX-DOAS instruments used in this thesis.

(1) Institute for Environmental Physics, University of Bremen, Germany.
(2) Max-Planck-Institute for Chemistry, Mainz, Germany.
(3) Institute for Environmental Physics, University of Heidelberg, Germany.
(4) Institute of Atmospheric Sciences and Climate, Bologna, Italy.
(5) Belgian Institute for Space Aeronomy, Uccle, Belgium.
(6) Royal Netherlands Meteorological Institute, De Bilt, Netherlands.

2.3.6 In situ measurements

The last class of measurements relevant for this thesis are the *in situ measurements*. These are measurements of trace gases, conducted directly at the point of interest (as opposed to remote measurements). Unlike DOAS, which returns column densities, in situ measurements yield actual trace gas *concentrations*. The in situ measurements referenced in this thesis are exclusively conducted at surface level within Europe, usually at around 4 m altitude. Some instrument characteristics discussed in the following (particularly in sect. 2.3.6) may not extend to other parts of the world.

"Umweltbundesamt" (UBA), the European Environment Agency (EEA), and the AirBase dataset

The European Union (EU) has issued the *Ambient Air Quality Directive* (2008/50/EC, see The European Parliament and The European Council, 2008). It defines yearly and hourly concentration threshold values for NO_2 (yearly: 40 µg m^{-3}, hourly: 200 µg m^{-3}). Member states of the EU can be legally prosecuted for exceeding these limits, and have implemented corresponding federal laws. It is the duty of federal agencies to ensure abidance to these laws. In Germany the responsible agency is called "Umweltbundesamt" (en: "Federal Agency for the Environment", UBA). The UBA directs the federal states to conduct in situ measurements for the surveillance of ambient air quality, collects their data, and forwards them to the EU. The "European Environment Agency" (EEA) uses these data to assemble a joint dataset

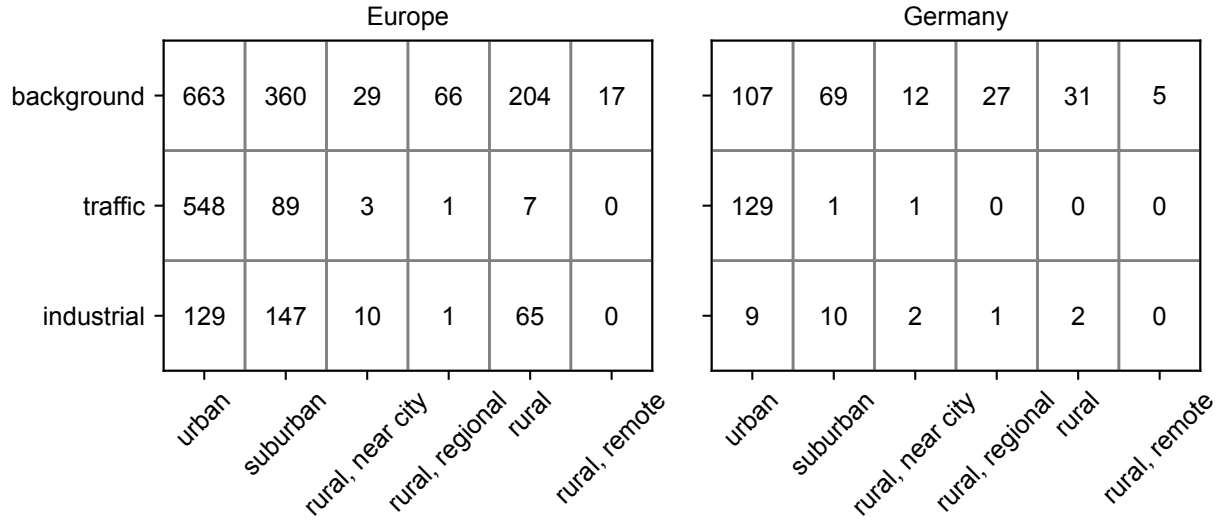

Figure 2.14: Classification matrix of the AirBase instruments. Each entry denotes the number of instruments corresponding to the instrument type (background, traffic industrial) and environment (urban, suburban, etc.). The left-side matrix refers to all AirBase instruments in Europe, and the right-side matrix exclusively to those in Germany.

called *AirBase*. Besides NO_2, AirBase contains measurements of NO, O_3, and CO, although not all stations can measure all gases. In 2022, the AirBase dataset contained NO_2 data of 2339 instruments, classified either as "background" (1339), "traffic" (648), or "industrial" (352). In Germany, a total of 406 instruments were in operation (background: 251, traffic: 131, industrial: 24). Stations are classified as traffic or industrial, if the emissions of either traffic or industrial facilities dominate the measurement, and the station is not representative of a larger surrounding area. In return, classification as a background station does not imply that the measurements are unaffected by traffic or industrial emissions. Besides this ternary classification, a further sub-classification exists, based on the stations' vicinity to urban regions. The subclasses are called "urban", "suburban", "rural, near city", "rural, regional", "rural", and "rural, remote". A classification matrix for the AirBase stations in Europe and Germany is found in Fig. 2.14.

The molybdenum-chemiluminescence (Mo-CL) measurement method

There exist different measurement techniques for in situ measurements of NO_2. These include:

- Cavity-enhanced DOAS, where the DOAS principle is applied to light which is reflected back and forth between two mirrors in a cavity. This increases the light path up to several kilometers and allows to use the DOAS method for in situ measurements. See e.g. Horbanski et al. (2019).

- Fluorescence methods, where the NO_2 molecules of the ambient air are excited, e.g.

using a laser, and quantified, based on the light flux emitted from the transition back to the ground state. See e.g. Javed et al. (2019).

- Cavity attenuated phase shift spectroscopy (CAPS), where NO_2 is quantified by sending square-wave modulated LED light into a cavity with ambient air and measuring the phase shift, which occurs during its traversal. See e.g. Ge et al. (2013).
- The molybdenum-based chemiluminescence method (Mo-CL), which will be discussed in detail in the following.

According to the AirBase metadata, approximately 98 % of the European in situ instruments use the *chemiluminescence method*. This method requires conversion of NO_2 to NO using a converter, usually by means of a heated molybdenum cartridge ("molebydenum-based chemiluminescence", abbreviated as Mo-CL). This method is associated with strong cross-sensitivities to other nitrogen compounds, resulting in significant biases during the NO_2 measurements.

The measurement principle of the Mo-CL method can be outlined as follows. Consider the reaction of NO and O_3:

$$\mathrm{NO} + \mathrm{O_3} \longrightarrow \mathrm{NO_2} + \mathrm{O_2} \tag{2.70}$$

At typical ambient temperatures a fraction of ~ 10 % of the produced NO_2 enters an excited state, from which it can return to the ground state either by emission of a photon or by collision with other molecules (*quenching*):

$$\mathrm{NO} + \mathrm{O_3} \longrightarrow \mathrm{NO_2}^* + \mathrm{O_2} \tag{2.71}$$

$$\mathrm{NO_2}^* \longrightarrow \mathrm{NO_2} + h\nu \tag{2.72}$$

$$\mathrm{NO_2}^* + \mathrm{M} \longrightarrow \mathrm{NO_2} + \mathrm{M} \tag{2.73}$$

In order to minimize the influence of quenching, Mo-CL instruments are usually operated at low pressures of below 100 mbar. Based on eq. (2.72), the ambient NO concentration can be quantified by measuring the light flux resulting from the luminescent decay.

In order to measure NO_2, it must first be converted to NO. This is possible by means of a molybdenum (Mo) cartridge, heated to ~ 300 °C:

$$\mathrm{Mo} + 3\,\mathrm{NO_2} \longrightarrow \mathrm{MoO_3} + 3\,\mathrm{NO} \tag{2.74}$$

as described in Villena et al. (2012). From thereon, a Mo-CL measurement is a simple two-step process: First, the NO concentration is measured via eqs. (2.71) and (2.72). Then the

NO_2 concentration is measured by reducing it to NO via eq. (2.74) and measuring the NO concentration again. Subtracting the second measurement from the first yields the amount of NO, which was obtained from the reduction of NO_2 via eq. (2.74):

$$[\mathrm{NO}]_1 = f_{\mathrm{NO}}(I_1) \tag{2.75}$$

$$[\mathrm{NO}]_2 = f_{\mathrm{NO}}(I_2) \tag{2.76}$$

$$[\mathrm{NO_2}] = f_{\mathrm{NO_2}}(\Delta[\mathrm{NO}]) = f_{\mathrm{NO_2}}([\mathrm{NO}]_2 - [\mathrm{NO}]_1) \tag{2.77}$$

where:

$[\mathrm{NO}]$: the NO concentration

$[\mathrm{NO_2}]$: the NO_2 concentration

I : the instrument's measurement signal (light flux from luminescent decay)

f_{NO} : the instrument's NO calibration function

$f_{\mathrm{NO_2}}$: the instrument's NO_2 calibration function

$\Delta\mathrm{NO}$: the NO difference between the second and first measurement

$_{1,2}$: denote the first and second measurement

The instrument's calibration functions f_{NO} and $f_{\mathrm{NO_2}}$ are determined in a laboratory beforehand. A detailed description of commercial Mo-CL instruments is given by Fontijn et al. (1970).

For over a decade, research studies on Mo-CL measurements have revealed concerning disagreements to alternative measurement methods, manifesting in the form of a positive NO_2 bias. This is supposedly linked to non-NO_x reactive nitrogen species (such as PAN, HNO_3, higher alkyl nitrates, etc.), commonly summarized as NO_z, which interfere with the measurement by reduction to NO prior to the second measurement. As a result, they create a positive NO_2 false-signal by contributing to ΔNO. The effect is expected to be stronger in less polluted regions, where the contribution of "true" NO_2 to ΔNO is comparably small. An overview of the literature on this topic is given in the following:

- Dunlea et al. (2007) present data from a field campaign in Mexico City. Mo-CL measurements are compared to DOAS measurements. The Mo-CL measurements were found to be biased by +22 % on average, and up to +50 % during afternoon hours. The bias is shown to correlate with NO_z.
- Steinbacher et al. (2007) compare Mo-CL to photolytic luminescence measurements in rural environments of Switzerland and identify biases of +17 % to +57 %, depending on the elevation of the measurement site.

- Lamsal et al. (2008) compare Mo-CL, photolytic chemiluminescence, and DOAS measurements from Mexico City. A Mo-CL bias of up to +100 % is identified. Correction factors for Mo-CL measurements, ranging from 0.4 during winter to 0.8 during summer, are proposed based on an assessment of the relative impact of different NO_z compounds to the Mo-CL bias.

- Villena et al. (2012) compare Mo-CL and DOAS measurements from a field campaign in Santiago de Chile. Here, a Mo-CL bias of up to +300 % was identified. A converse effect (i.e. a negative bias) is identified for photolytic chemiluminescence measurements during measurements taken in a road traffic tunnel.

- Ge et al. (2013) compare Mo-CL measurements to CAPS measurements in Beijing, China. Mo-CL biases on the scale of +20 % were observed and showed a clear correlation to NO_z.

- Reed et al. (2016) delve into the hypothesis of an unknown "oxidant X", which is supposedly able to convert NO to NO_2. The study was motivated by anomalously high NO_2 concentrations, reported from photolytic chemiluminescence measurements in low-NO_x environments. The authors explain, that "oxidant X" is *not* required in order to explain the measurement results. Instead, they deliver an alternative explanation, based on the thermal decomposition of NO_z (particularly PAN) to NO_2 inside the in situ instruments, even if no Mo cartridge is used.

- Jung et al. (2017) investigate the seasonal variations of the Mo-CL bias at a suburban site in Korea. The biases range from +16.6 % in winter to +28.9 % in spring. Interestingly, the study finds practically no diurnal dependence of the bias, which Lamsal et al. (2008) and Villena et al. (2012) clearly identified.

- Poraicu et al. (2023) report on an RCT simulation with the WRF-Chem model, which was used to derive Mo-CL biases of approximately +25 % based on modelled NO_2 and NO_z concentrations over a Belgian domain.

Figure 2.15 shows exemplary measurements from Lamsal et al. (2008) and Villena et al. (2012), with a clear diurnal cycle in direct comparison to DOAS measurements. To summarize, there appears to be clear consensus on the existence of the Mo-CL bias. The strong correlation to NO_z and O_3 support the claim that photooxidants other than NO_2 are its main cause. On the other hand, the reported bias varies significantly between the referenced studies, ranging from +20 % up to +300 %. For completeness on this matter it should be mentioned, that despite the qualitative unanimity among the cited literature, the UBA has made contrary statements on the existence of the Mo-CL bias in a personal communication, see Appendix B.1.

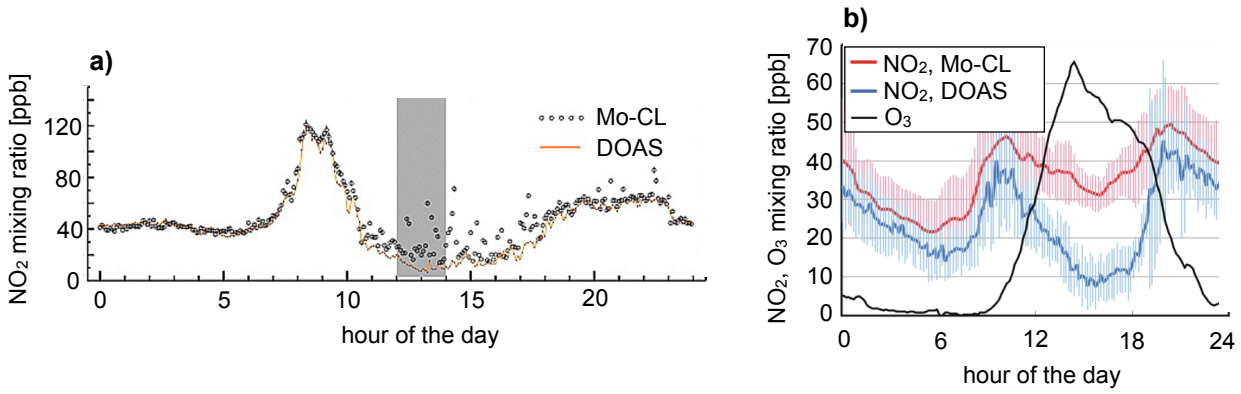

Figure 2.15: NO_2 mixing ratios obtained from Mo-CL and DOAS measurements. a) Adapted from Lamsal et al. (2008). **b)** Adapted from Villena et al. (2012).

Correction method for the Mo-CL bias

Lamsal et al. (2008) present an empirical recipe for the correction of the Mo-CL bias based on the volume mixing ratios (VMRs) of the most relevant NO_z gases, being PAN, HNO_3, and the alkyl nitrates. The bias is corrected with a multiplicative factor F, computed as

$$F = \frac{[NO_2{}^*]}{[NO_2]} = 1 + \frac{0.95 \cdot [PAN] + 0.35 \cdot [HNO_3] + \sum[AN]}{[NO_2]} \tag{2.78}$$

where $\sum[AN]$ refers to the sum of alkyl nitrates, and the square brackets denote the VMRs of the respective gases. Thereby, bias-corrected Mo-CL measurement results are obtained as

$$[NO_2] = [NO_2{}^*]/F \tag{2.79}$$

$$[NO_x] = [NO] + [NO_2{}^*]/F \tag{2.80}$$

However, the required VMRs of PAN, HNO_3, and the alkyl nitrates are usually unknown, because they are not being measured by the situ instruments. Nonetheless, F can be computed on the basis of colocated modelled VMRs, e.g. from a RCT simulation. This approach is used in Chapters 3 and 4. It should be noted, that the chemical mechanisms of RCT models (e.g. the "Model for OZone and Related chemical Tracers" (MOZART), see Emmons et al., 2010) often do not implement alkyl nitrate chemistry, i.e. they assume $\sum[AN] = 0$.

2.4 Regional chemistry and transport (RCT) modelling

Regional chemistry and transport (RCT) modelling is an important discipline in the field of atmospheric research. RCT models are implementations of atmospheric mechanisms into

computer code, based on textbook knowledge and empirical parametrizations. This includes processes such as transport and mixing, dry and wet deposition, radiative transfer, microphysics, photochemistry, and emissions. The main goal of running RCT simulations is to obtain atmospheric variables or trace gas distributions with dense spatial and temporal coverage (e.g. with a horizontal resolution of 3 km × 3 km and hourly output). This is particularly helpful for the systematic investigation of trace gases, which cannot realistically be measured in a spatially continuous manner in three dimensions. However, RCT models come at a considerable computational expense. For example, a simulation of one month over a domain ranging from 1 °W to 23 °E and 43 °N to 56 °N with a horizontal resolution of 3 km × 3 km takes approximately one week on ∼ 800 CPUs (or 134400 core-hours).

In the following, the important theoretical aspects of RCT modelling are discussed, with a strong focus on the state-of-the-art *Weather Research and Forecast model with chemistry* (WRF-Chem, see Grell et al., 2005). The technical details of our own WRF-Chem simulation runs are presented in Chapter 3. The NO_2 profiles of this WRF-Chem simulation are used to train the NitroNet neural network.

2.4.1 Model classifications

Chemistry and transport models are often distinguished by the following criteria:

- Depending on whether the model includes feedback between atmospheric chemistry and dynamics, it is referred to as an "online" (with feedback) or "offline" (without feedback) coupled model. WRF-Chem is an online coupled model, and therefore allows meteorology/transport and chemistry to affect and interact with each other directly. In comparison to offline coupled models, online coupled models are often considered more realistic.

- Models are further distinguished by their overall spatial resolution and coverage. Global models, such as TM5-MP, operate on a reduced horizontal resolution (e.g. $1° \times 1°$) in order to compensate for the computational demands related to global coverage. Models like WRF-Chem, which are designed to operate on significantly extended domains (e.g. entire countries or continents), but not globally, are called "regional models". They typically operate on horizontal resolutions of 1 km × 1 km to 10 km × 10 km. The highest resolved models are called "local" models, and operate on resolutions of ∼ 100 m × 100 m or better (see e.g. Karl, 2018).

- The reference system, in which the transport equations of the model are solved, can be either global/fixed, or local/moving. By this criterion, models are further divided into *Eulerian* and *Lagrangian* models. WRF-Chem is an Eulerian model.

2.4.2 Chemical mechanisms, physical schemes, and parametrizations

RCT models can be seen as modular systems, in the sense that they allow the user to exchange the computational codes responsible for different physical mechanisms (e.g. boundary layer modelling, radiative transfer, microphysics, etc.). These computational units are called *schemes*. Different schemes may be based on different physical assumptions and parametrizations and therefore have different computational demands. Modelling of photochemistry is similarly flexible, by means of the *Kinetic PreProcessor* software (KPP, see Lin et al., 2023). KPP allows to define *chemical mechanisms* via chemical reactions along with their kinetic reaction constants (or photolysis rates, in case of photolysis reactions), which KPP then translates into a system of ordinary differential equations (ODEs), that WRF-Chem solves using a *Rosenbrock solver*.

2.4.3 Vertical coordinates

RCT models typically use vertical pressure coordinates. Historically, these were developed in order to achieve terrain-following vertical grids in numerical simulations (see Phillips, 1957, where the first coordinate of this kind, called the σ-coordinate, was introduced). The σ-coordinate is defined as

$$\sigma(p) = p/p_s \tag{2.81}$$

where:

p : pressure

p_s : surface pressure

WRF-Chem uses the η-coordinate, a successor to the σ-coordinate, defined as

$$\eta(p) = \frac{p(z_s) - p_t}{p_0 - p_t} \tag{2.82}$$

where:

$p(z_s)$: pressure at the nearest model terrain level z_s

p_t : pressure at the model top

p_0 : the standard atmosphere mean sea level pressure (= 1013 hPa)

As such, the η-coordinate is another form of the σ-coordinate. By definition, $0 \leq \eta \leq 1$, where 0 defines the model top, and 1 the model bottom. The TM5-MP model in the TROPOMI

retrieval uses the hybrid σ-coordinate, which defines pressure via

$$p(\sigma_a, \sigma_b) = \sigma_a + \sigma_b \cdot p_s \tag{2.83}$$

where σ_a and σ_b represent the dimensionless vertical coordinates of the system. For interpolation between vertical pressure coordinate systems the NCAR Command Language (NCL, see Brown et al., 2019) can be used. In particular, NCL supplies a `sigma2hybrid` function to interpolate WRF-Chem simulation output to the pressure levels of the TROPOMI retrieval. This is necessary when validating NO_2 profiles from a WRF-Chem simulation against TROPOMI measurements. In that pursuit, the simulated tropospheric NO_2 VCD is computed as

$$V_t = \sum_{l=0}^{l_{\text{tp}}} V_l = \sum_{l=0}^{l_{\text{tp}}} c_l h_l \tag{2.84}$$

where:

l_{tp} : the tropopause layer of in the TROPOMI retrieval

V_l : the modelled NO_2 box VCD in layer l

c_l : the modelled NO_2 concentration in layer l

h_l : the vertical layer extent of layer l

The layer extent h_l is obtained from numerical integration of the barometric height formula in its differential form (see eq. 2.33). The required temperature profile is taken from the simulation. The surface pressure p_0 is taken from the TROPOMI retrieval.

2.4.4 Data assimilation, nudging, and reanalysis

Data assimilation (DA) describes the practice of using observations, i.e. from satellite instruments or in situ measurements, to improve the accuracy of a RCT simulation during runtime. Such methods are often considered a necessity, in order to prevent the model from drifting away from the true state (e.g. due to errors in the initial and boundary conditions, physical processes not taken into consideration, or parametrizations being overly simplified). In practice the atmospheric state can be well forecast for a few days, whereafter forecasts increasingly diverge from the true state. The term *meteorological reanalysis* refers to datasets, which were obtained from short-range weather simulations with observational data assimilation, intended to be gap-free and fully consistent (e.g. the ERA5 dataset, see Hersbach et al., 2020).

WRF-Chem uses a four-dimensional data assimilation method called *nudging* or *Newtonian relaxation*, see Stauffer and Seaman (1990). Nudging relaxes the model trajectory towards

either individual observations (e.g. from meteorological measurement stations), or an *analysis* (i.e. a set of observations, calculated at every grid point of the model). Furthermore, an advanced approach to nudging called *spectral nudging* exists (see e.g. Kida et al., 1991; Omrani et al., 2012). Spectral nudging has the advantage of being selective to spatial scales. Thereby a model can be nudged with respect to the large-scale atmospheric circulations of an analysis, while maintaining full computational autonomy on smaller scales.

2.4.5 Emission inventories, speciation, temporal and vertical emission profiles

Anthropogenic emission inventories

Emission data are among the most important input to a RCT simulation, particularly for short-lived gases from highly variable anthropogenic sources, such as NO_2. Anthropogenic emission datasets are referred to as *emission inventories*, and may vary in horizontal resolution, spatial coverage, speciation, and emission classification. Furthermore, emission inventories can represent either monthly or yearly emission averages. Emission inventories are essentially 2-dimensional maps (resolved in longitude and latitude), which describe the emission strength of different *emission sectors*. Emission sector names are standardized by the *Selected Nomenclature for Air Pollution* (SNAP, see European Environment Agency, 2023), which identifies each sector by a number. For example, the SNAP number "3" corresponds to emissions from industrial combustion. Notable emission inventories are EDGARv5 (monthly means, global coverage, $0.1° \times 0.1°$ (lat $\times$ lon), see Crippa et al., 2020), CAMS-REG-v4.2 (yearly means, European coverage, $0.1° \times 0.05°$, see Kuenen et al., 2022), and HTAP_v3 (monthly means, global coverage, mixed resolution depending on region, see Crippa et al., 2023). The uncertainty of such inventories is estimated between 20 % to 50 % (see e.g. Solazzo et al., 2021, who refer to greenhouse gas emissions specifically, or Kuenen et al., 2022). Additionally there exist regional emission inventories, often issued by federal agencies such as the UBA (see sect. 2.3.6), for example the high-resolution emission inventory for Germany "UBA-E" (yearly means, $0.01° \times 0.01°$, see Hausmann et al., 2020). Global and regional emission inventories can be combined in order to maximize coverage and resolution using emission pre-processors, such as HERMESv3 (Guevara et al., 2019).

Temporal emission profiles

Currently available emission inventories typically have either a yearly or monthly resolution. Many emission sectors, however, are characterized by systematic temporal variations on different time scales. For example, emissions from road traffic are expected to show characteristic peaks during the late morning and afternoon rush hours, and agricultural emissions typically

occur in specific months of the year, shortly after the use of fertilizers. In order to capture these temporal patterns, the coarsely resolved emission data are scaled to an hourly resolution using presumed hourly, daily, and monthly scaling factors, called *temporal profiles*. The emission E_X (units: kg m^{-2} s^{-1}) of a species X is then given as

$$E_X(m,d,h) = \sum_k \hat{E}_{X,k} \cdot p_{\text{monthly},k}(m) \cdot p_{\text{daily},k}(d) \cdot p_{\text{hourly},k}(h) \tag{2.85}$$

where:

k : the emission sector

$\hat{E}_{X,k}$: the emission of species X from sector k, as given in the inventory

$p_{\text{monthly},k}$: the monthly temporal profile of sector k

$p_{\text{daily},k}$: the daily temporal profile of sector k

$p_{\text{hourly},k}$: the hourly temporal profile of sector k

E_X and $\hat{E}_{X,k}$ depend on longitude and latitude, whose notation is omitted here. The monthly, daily, and hourly temporal profiles are normalized to 12 (annual cycle), 7 (weekly cycle), and 24 (diurnal cycle). Because different emission sectors follow vastly different temporal patterns, the temporal profiles are defined for each sector individually. The profiles are intended to represent average emission patterns on a regional scale, but may be inaccurate on local scales.

Recommendations for temporal emission profiles are given by Builtjes et al. (2002), Crippa et al. (2020), and Kumar et al. (2021). Such publications use statistical data (e.g. reports from power plants on their daily energy production, or car counts on highways) to derive harmonized temporal profiles.

Speciation of emissions

Emission inventories report NO_x and VOCs as *lumped species*. This means, that the partitioning of NO_x into NO and NO_2 must be defined by the user via so-called *speciation profiles*. For speciated emissions, eq. (2.85) changes to

$$E_X(m,d,h) = \sum_k \hat{E}_{X,\text{lump},k} \cdot p_{\text{monthly},k}(m) \cdot p_{\text{daily},k}(d) \cdot p_{\text{hourly},k}(h) \cdot p_{\text{spec},k}(X) \tag{2.86}$$

where:

$\hat{E}_{X,\text{lump},k}$: the lump containing X of sector k

$p_{\text{spec},k}$: the speciation profile of sector k

For example, the lump of $X = NO_2$ is NO_x. For unlumped species, $\hat{E}_{X,\mathrm{lump},k} = \hat{E}_{X,k}$ and $p_{\mathrm{spec},k}(X) = 1$. A reasonable speciation of NO_x could be 87.5 % NO, and 12.5 % NO_2, based on NO_2/NO_x ratios in the range of $5 - 40$ %, as reported for combustion processes (Jimenez et al., 2000; Costantini et al., 2016; Wild et al., 2017; Richmond-Bryant et al., 2017). For VOCs, speciation profile recommendations are given by Huang et al. (2017).

Vertical emission profiles

Vertical emission profiles distribute the emissions of different sectors along the vertical axis, in direct analogy to the temporal profiles. While the majority of NO_x emissions occurs at the surface (road traffic, agricultural soils, and to some extent residential heating), other emission sectors emit exclusively at higher altitudes. This is the case for combustion processes in the energy industry, production processes in the manufacturing industry, waste disposal, and the extraction of fossil fuels. Equation (2.86) can be extended by a vertical emission profile:

$$E_X(m,d,h,z) = \sum_k \hat{E}_{X,\mathrm{lump},k} \cdot p_{\mathrm{monthly},k}(m) \cdot p_{\mathrm{daily},k}(d) \cdot p_{\mathrm{hourly},k}(h) \cdot p_{\mathrm{spec},k}(X) \cdot p_{\mathrm{vertical},k}(z) \tag{2.87}$$

where:

$p_{\mathrm{vertical},k}$: the vertical profile of sector k

z : altitude

An important contribution to the vertical displacement of emissions comes from the plume-rise effect: When hot gasses from combustion processes are emitted into ambient air, they can rise up by hundreds of meters due to thermal updraft. Figure 2.16 shows a plume rise scenario from a power plant, recorded with an NO_2 imaging instrument (see Kuhn et al., 2022). The NO_2 plume is emitted from a stack of approximately 200 m height (bottom right corner) and advected towards the left image corner. Within the first 300 m downwind, the plume rises to an altitude of $400 - 500$ m, where it remains throughout the field of view. Depending on local meteorology, the effect can be far stronger than shown in Fig. 2.16. Because the spatial resolution of RCT simulations might not always be sufficient to compute plume rise explicitly, it is often already accounted for in the vertical emission profiles. For example, in case of a point source such as shown in Fig. 2.16 vertical emission profiles might place the emissions at the *effective* emission height of $400 - 500$ m, instead of the stack height at 200 m. Recommendations for vertical profiles are given by Bieser et al. (2011) and Pozzer et al. (2009). In the study of Pozzer et al. (2009), surface NO_x concentrations in polluted regions were found to change by up to 100 % upon use of vertical emission profiles.

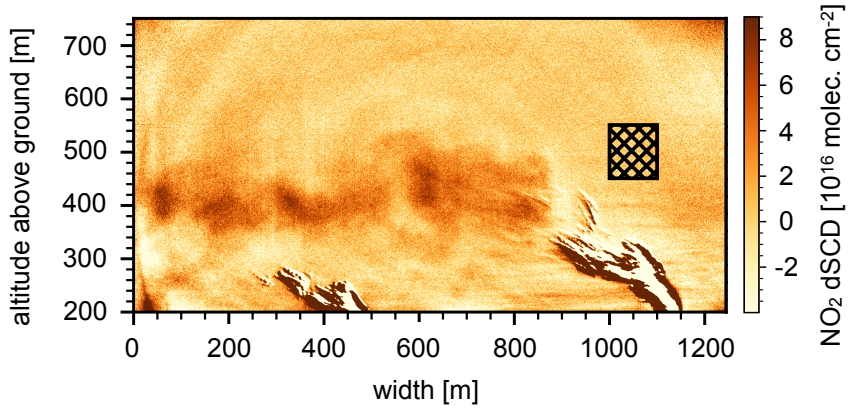

Figure 2.16: Plume rise from a power plant stack. Recorded with the gas correlation spectroscopy based NO_2 camera, see Kuhn et al. (2022). The structure in the bottom right (at $\sim$ 1100 m width) resembles the condensed phase of the plume, which unlike the un-condensed part starting at $\sim$ 900 m width, cannot be accurately quantified by the instrument. The measurement was taken at the Großkraftwerk Mannheim on 26 April 2021. The patterned patch in the center-right side defines a background region used for the evaluation of the instrument's raw data and can be ignored.

Other emission data

The preceding explanations have referred to anthropogenic emissions. In many parts of the world (e.g. central Europe), anthropogenic emissions can be assumed to be the most important emission group with respect to NO_x. Nonetheless, some emissions into the atmosphere are of non-anthropogenic origin and must be accounted for in RCT simulations. This includes mainly:

- Emissions from biomass burning. Although a large fraction of biomass burning is indeed of anthropogenic origin, there are natural contributions, e.g. from forest fires. A popular biomass burning emission inventory is the *Fire Inventory from NCAR* (FINN, see Wiedinmyer et al., 2011) at a horizontal resolution of 1 km $\times$ 1 km and temporal resolution of 24 h. Vertical and temporal emission profiles are included in FINN, in the form of a diurnal variation that peaks in the early afternoon (1 pm local time), and a vertical plume-rise parametrization according to Freitas et al. (2007).

- Biogenic emissions, which can be computed online from plant-cover data with the *Model of Emissions of Gases and Aerosols from Nature* (MEGAN, see Guenther et al., 2006). Emission models like MEGAN partly resort to parametrizations (e.g. via the "leaf area index" for plant emissions). MEGAN is also used to compute biogenic soil NO_x emissions, while anthropogenic soil emissions (e.g. from agricultural fertilizers) are included in the anthropogenic emission inventories, see Table 2.1.

- Other natural NO_x emissions, which are parametrized in the RCT model itself. For

example, lightning events during thunderstorms are parametrized with a lightning flash rate combined with a presumed NO_x emission per flash. The available schemes in WRF-Chem are described in the user guide, see University Corporation for Atmospheric Research (2024).

2.4.6 Uncertainties in RCT simulations

Uncertainties in RCT simulations arise from uncertainties in the input data (meteorological data, chemical initial and boundary conditions, emission data), and the choice of physical schemes and parametrizations. The impact of either type of uncertainty can be assessed by repeating the simulation multiple times, while varying the input data or the model's schemes and parametrizations. For example, the *Copernicus Atmosphere Monitoring Service* (CAMS) model investigates the latter by an ensemble approach, where many RCT models are operated in unison, from which statistical diagnostics on the ensemble uncertainty can be obtained (see e.g. Eskes et al., 2024). For the simulations presented in this thesis, no such assessments were made due to the associated computational demands.

2.4.7 A review of NO_2 RCT simulations in scientific literature

In closing this section, an overview of existing peer-reviewed RCT model evaluation papers with focus on NO_2 is given. Note, that the RCT simulations mentioned here were run with different horizontal resolutions, physical parametrizations, and chemical mechanisms. As such they cannot be directly compared to each other. Nonetheless, they reveal reoccurring patterns of disagreement between simulations and observational datasets. All of the studies mentioned below have used AirBase/UBA data from background stations, or similar data from measurements in China.

- Terrenoire et al. (2015) present a model evaluation of the CHIMERE RCT model with the MELCHIOR chemical mechanism over Europe, and a horizontal resolution of $0.125° \times 0.0625°$. Emission data was patched together from different regional emission inventories. Strong underestimations of the modelled NO_2 surface concentrations, namely -33.9 % for rural background stations and -53.6 % for urban background stations, were identified and attributed to underestimations in the emission data.

- Petetin et al. (2015) evaluate the CHIMERE RCT model over the region of Paris against surface observations and aircraft measurements of black carbon and NO_2. Three emission inventories (EMEP, TNO, and TNO-MP) were used, and found to overestimate the NO_x emission by up to $+30$ %. The simulated NO_x concentrations showed qualitative differences to background measurements, with overestimations in the afternoon

and during the night. However, the study only used data from a single background measurement location. The uncertainty of the NO_x simulation results was estimated around 35 %, of which 19 % were attributed to uncertainties in vertical mixing.

- Visser et al. (2019) report on a simulation using WRF-Chem v.3.7.1 with the *Carbon Bond Mechanism Z* (CBM-Z) over central Europe, and a horizontal resolution of 20 km × 20 km. The noontime NO_2 surface concentrations and VCDs were compared to AirBase and OMI measurements, and found to be low-biased by −38.5 % and −15 %, respectively. The authors identify an underestimation of soil emissions in their emission inventory (TNO-MACC-III, short for "Monitoring Atmospheric Composition and Climate" by the Netherlands Organisation for Applied Scientific Research) as a possible explanation.

- Kuik et al. (2016) present WRF-Chem v.3.7.1 simulations with the *Regional Acid Deposition Model 2* (RADM2), the same emission inventory as Visser et al. (2019) over the region of Berlin, Germany, and a horizontal resolution of up to 1 km × 1 km. The authors find underestimations of surface NO_2 by more than −50 % during daytime and overestimations of +60 % at night-time. The study reveals, that increasing the spatial resolution (including downscaling of emission data) of the model from 15 km × 15 km to 1 km × 1 km slightly improves agreement, but not to a satisfying degree. In a follow-up publication (Kuik et al., 2018) the authors attribute the disagreements to underestimations in the emission data.

- Poraicu et al. (2023) show a WRF-Chem v.4.1.2 simulation with the CBM-Z mechanism over a Belgian domain, and a horizontal resolution of up to 1 km × 1 km. Emissions were patched together from local emission inventories (1 km × 1 km) and coarser regional inventories (0.1° × 0.1°). Poraicu et al. (2023) are the only authors in this selection of articles, who have adopted the Mo-CL bias correction according to Lamsal et al. (2008). Nonetheless, the NO_2 surface concentrations are underestimated by −40 % at noon, and up to +77 % during nighttime. The tropospheric NO_2 VCD, however, only shows a small bias of −4 % in comparison to TROPOMI measurements.

- Kumar et al. (2021) present a simulation using the MECO(n) model system over a German domain using the *Mainz Isoprene (chemical) Mechanism* (MIM 1), and a horizontal resolution of 2.2 km × 2.2 km. The simulated NO_2 surface concentrations show a small bias of −7 %. However, this statement refers to monthly-mean values, where daytime underestimations and nighttime overestimations might cancel out.

- Mar et al. (2016) study the influence of the chemical mechanism on modelled O_3 and NO_2 by comparison of the MOZART and RADM2 mechanisms in WRF-Chem v.3.5.1

on a European domain and a resolution of 45 km × 45 km. On average the NO_2 concentrations from MOZART were ~ 2 µg m^{-3} larger than those from RADM2, but up to ~ 20 µg m^{-3} larger for O_3. The difference in simulated NO_2 is almost negligible, and cannot sufficiently describe the biases observed in this selection of articles.

- Knote et al. (2015) conduct a box-model study on the differences between various chemical mechanisms with respect to NO_2, NO_x, O_3, and other trace gases. They reveal, that e.g. NO_x can vary by up to 25 % and NO_3 by up to 100 %, depending on the choice of chemical mechanism.

- Du et al. (2020) demonstrate, that tuning the vertical mixing parametrization of different boundary layer schemes in WRF-Chem drastically reduces the model bias for particulate matter at nighttime during summer months. This article deserves mentioning here, because it gave the decisive hint towards investigating the mixing scheme of WRF-Chem, discussed in more detail in Chapter 3.

Altogether, despite differences in model domains, resolution and chemical mechanisms, the contemporary literature comes to the clear consensus, that WRF-Chem (and other RCT simulations) underestimate NO_2 surface concentrations by $20 - 50$ %. A few studies claim, that the biases can be attributed to inaccuracies in the emission data. The investigation of diurnal cycles reveals underestimations at daytime and even stronger overestimations at nighttime. Except for Poraicu et al. (2023) none of the studies have attempted to correct for the Mo-CL bias. Comparisons to tropospheric NO_2 VCDs from satellite measurements occur more rarely in literature and show generally better agreement. However, those often yield only a single measurement of the vertically integrated concentration per day, and thereby do not cover the diurnal cycle. Furthermore, because the VCDs represent vertically integrated tropospheric loads of NO_2, they are less indicative of faulty profile shapes near the surface.

2.5 Fundamentals of machine learning with artificial neural networks

The term *machine learning* (ML) describes the field of study concerned with designing computational models and algorithms that solve complex problems by extracting relevant information from large datasets comprised of training examples. The power of machine learning lies within its potential to do so without full explicit instructions stemming from prior analytical understanding of the problem itself. A prominent type of ML model are the *artificial neural networks*. In this thesis a neural network is used to solve a regression task, i.e. it predicts a quantity of interest $\boldsymbol{y}$ (the NO_2 concentration at a given altitude) given an input vector

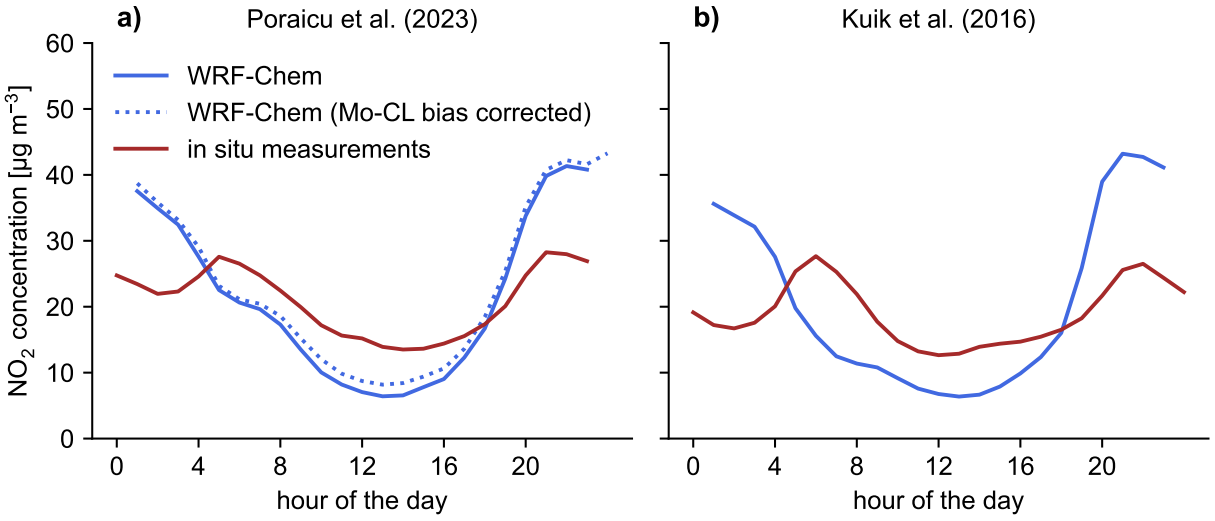

Figure 2.17: Diurnal cycles of surface NO_2 from recently published RCT simulations. a) Simulation on a Belgian domain. Data taken and re-plotted from Poraicu et al. (2023). **b)** Simulation on a domain over Berlin, Germany. Data taken and re-plotted from Kuik et al. (2016). Both simulations have horizontal resolutions of up to 1 km $\times$ 1 km and were evaluated against background in situ measurements.

$\boldsymbol{x}$ (containing satellite observations and ancillary variables). In the following an overview of the necessary ML fundamentals is given. Note that the introductory Chapter 1 described the research problem of this thesis in the notation conventions of inverse problems (where e.g. $\boldsymbol{y}$ denotes the constraints of a posterior probability density). The slight changes in notation from hereon is unavoidable in order to comply with the modern conventions that have established in the field of machine learning.

2.5.1 Artificial feed-forward neural networks

Artificial neural networks (NNs) are arguably the most capable ML models known today. A comprehensive summary is found in Schmidhuber (2015). Their history reaches back to the 1940s (see McCulloch and Pitts, 1943), when the idea of designing a computational model inspired by the structure of the human brain was first proposed. It was known at that time that the brain operates by transmission of electric signals between nodes ordered in layers. To this day most neural networks are designed on the basis of this topology.

The building blocks of neural networks are the *artificial neurons*, which should be understood as a minimal computational unit (see Fig. 2.18a). They take an input vector $\boldsymbol{x}$ and compute the term

$$y = \phi\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + b\right) \tag{2.88}$$

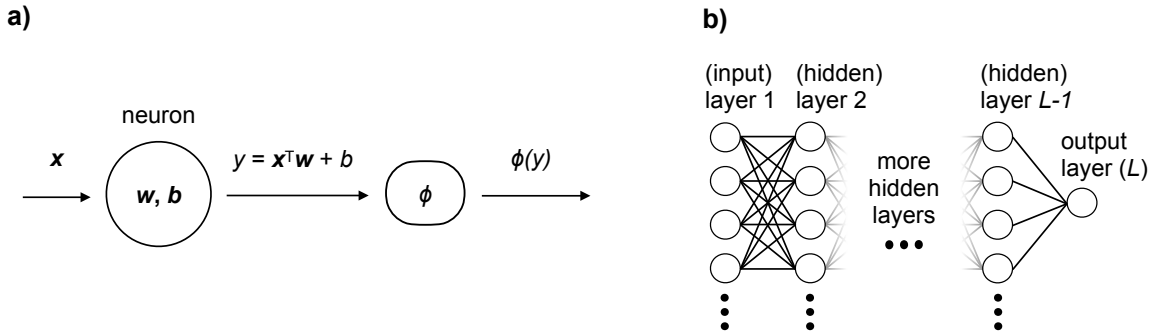

Figure 2.18: Artificial neurons and feed-forward neural networks. a) Computational process inside a single artificial neuron. **b)** Multiple neurons in a feed-forward neural network. The information (or signal) flows from the left to the right through several hidden layers. The final layer returns the neural network's prediction. In general, the last layer can have multiple output neurons. The NitroNet model has only one, as shown here.

where:

- $\boldsymbol{x}$: the input vector
- $\boldsymbol{w}$: the neuron's weight vector
- b : the neuron's bias
- ϕ : the neuron's activation function ($\phi : \mathbb{R} \to \mathbb{R}$)

The rationale behind this computation is, that biological neurons work in a similar way: By computing a sum of multiple input sources (e.g. neurons of the preceding layer) they condense the received information. The incorporation of weights allows the neuron to differentiate between input sources of higher and lower importance. Notice, that prior to the application of the activation function ϕ, the neuron's computation is merely an affine transformation of $\boldsymbol{x}$. Because the affine transformations form an algebraic group, a network of neurons expressed as a chain of affines could be reduced to a single neuron. For this reason an activation function is required, that makes the neuron's output non-linear. Thereby, complex non-linear models can be built by chaining neurons. A usual choice for ϕ is the *rectified linear unit*, defined as

$$\mathrm{ReLU}(x) = \max(x, 0) \tag{2.89}$$

However, there are many alternatives, often slight variations of ReLU, such as the *Parametric*

Rectified Linear Unit (PReLU, see He et al., 2015) defined as

$$\mathrm{PReLU}(x) = \begin{cases} ax & x \leq 0 \\ x & x > 0 \end{cases}, \tag{2.90}$$

where a denotes a "trainable" parameter (see below).

In the canonical topology neurons are stacked in layers, and each neuron is connected to all neurons of the following layer (*dense layer*). Such networks are called *feed-forward neural networks* (see Fig. 2.18b). The first layer is called *input layer*, and the subsequent layers are called *hidden layers*. The term *deep neural network* refers to neural networks which contain a complex inner structure on account of their many hidden layers. The last layer is called *output layer* and can contain multiple neurons. The neural network of NitroNet, however, has only a single output neuron. Equation (2.88) can be extended from a single neuron to an entire layer of neurons, written as

$$\boldsymbol{y} = \phi\left(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}\right) \tag{2.91}$$

where:

$\boldsymbol{x}$: the input vector

$\mathbf{W}$: the layer's weight matrix

$\boldsymbol{b}$: the layer's bias vector

Here, ϕ is applied element-wise. The input vectors are usually stacked in a matrix (the *feature matrix* $\mathbf{X}$), so that each row represents one input vector (one *instance*). The entries of the input vectors are termed *features*. A big advantage of neural networks is that their computations can be parallelized on graphics processing units (GPUs), which are orders of magnitude faster than serial computer processors (CPUs).

Feed-forward neural networks can be used to approximate highly non-linear functions and processes. In fact there exists mathematical proof that neural networks are *universal function approximators*, meaning that they can approximate arbitrary "well-behaved" functions, see e.g. Hornik et al. (1989). The neural networks obtain this capability from their enormous internal complexity and many degrees of freedom from their weight matrices and bias vectors. For example, a neural network with 5 layers of 100 neurons each has

$$\underbrace{5 \cdot 100^2}_{\text{weights}} + \underbrace{5 \cdot 100}_{\text{biases}} = 50500 \tag{2.92}$$

free parameters. As a downside, it is this level of complexity that turns neural networks into

black box models, meaning that insight into how they process/interpret their input data is very limited.

2.5.2 Other neural networks

There exist many other types of neural networks. Notable examples are *convolutional NNs* (CNNs, used in image processing, or more broadly, processing data with spatial context, see LeCun et al., 2015), *recurrent NNs* (RNNs, used for processing sequential data, e.g. time series, see LeCun et al., 2015), and *physics-informed NNs* (PINNs, which incorporate prior physical knowledge, see Raissi et al., 2019). The *invertible NNs* (INNs) are a relatively new addition to the family of neural networks, an overview of which is given in Ardizzone et al. (2019). They are designed to solve inverse problems in a framework of Bayesian statistics, and were originally planned to be used in this thesis. However, the combined results of Chapters 3 and 4 demonstrate that feed-forward NNs are sufficient for the task of NO_2 profile prediction, while being much easier to train.

2.5.3 Training of neural networks

To train a neural network means to optimize its trainable parameters (i.e. its weights and biases, and possibly a few others, such as the a-parameter of the PReLU function) in an iterative process from a large dataset $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$ of exemplary input instances ($\boldsymbol{x}_i$) and desired outputs ($\boldsymbol{y}_i$, called *targets*). This is conceptually very similar to fitting a non-ML model (e.g. a polynomial) to data. However, the neural networks' vast complexity often requires highly optimized numerical routines in order to find a satisfying solution.

Training and test splits

Prior to training the available data are split randomly into two disjoint subsets: a *training set* and a *test set*. A third set, the *validation set*, will be introduced shortly in sect. 2.5.5. The training set is used to optimize the neural network's trainable parameters. After training, the neural network is evaluated on the independent test set. A typical procedure would be to use 90 % of all data for training, and 10 % for testing. Good performance on both subsets indicates that the neural network has extracted ("learned") functional relationships of general validity which extend to previously unseen instances ("generalization"). On the contrary, a neural network that performs well on the training set but poorly on the test set has adapted to information exclusive to the training set (e.g. its noise) and is of no use. This is called *overfitting* and indicates that the neural network posesses too many degrees of freedom, or requires more abundant/varied training data. A common remedy against overfitting is to make use of *regularizations* (see sect. 2.5.4).

The maximum likelihood principle

Neural network training can be approached from statistical principles. Consider the neural network as a function $\mathrm{NN}_{\boldsymbol{\theta}}$, so that given an input x_i we may write

$$\hat{y}_i = \mathrm{NN}_{\boldsymbol{\theta}}(x_i) \tag{2.93}$$

where $\boldsymbol{\theta}$ denotes the tunable parameters of the network. The training set consists of N pairs $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$. x_i and y_i are in a mutual relationship of the form

$$y_i = f(x_i) + \epsilon_i \tag{2.94}$$

where:

f : the "true" forward function

ϵ_i : the noise term, here distributed as independent, identical Gaussians: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Note that x_i and y_i can be vector quantities, but for the ease of notation they are treated as scalars here. The goal of training the neural network is to find a set of optimal parameters $\hat{\boldsymbol{\theta}}$ so that the neural network best approximates f. In the ideal case,

$$\hat{y}_i = \mathrm{NN}_{\hat{\boldsymbol{\theta}}}(x_i) = f(x_i) \tag{2.95}$$

$$\epsilon_i = y_i - \hat{y}_i \sim \mathcal{N}(0, \sigma^2) \tag{2.96}$$

holds. In other words, the probability of the neural network to reproduce the observations y_i can be written as

$$p(y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \hat{y}_i\right)^2\right) \tag{2.97}$$

where the dependence on $\boldsymbol{\theta}$ is contained in $\hat{y}_i$. If the noise terms are independent of each other, the joint probability density factorizes to

$$p_{\boldsymbol{\theta}}(y) := p(y|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i, \hat{y}_i) \tag{2.98}$$

The *maximum likelihood estimator* $\hat{\boldsymbol{\theta}}$ maximizes this probability, i.e.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left[p_{\boldsymbol{\theta}}(y) \right] \tag{2.99}$$

$$= \arg\max_{\boldsymbol{\theta}} \left[\ln p_{\boldsymbol{\theta}}(y) \right] \tag{2.100}$$

$$= \arg\min_{\boldsymbol{\theta}} \left[-\ln p_{\boldsymbol{\theta}}(y) \right] \tag{2.101}$$

$$= \arg\min_{\boldsymbol{\theta}} \left[-\ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(y_i - \hat{y}_i \right)^2 \right) \right] \tag{2.102}$$

$$= \arg\min_{\boldsymbol{\theta}} \left[\underbrace{-\sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{independent of } \boldsymbol{\theta}} + \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 \right] \tag{2.103}$$

$$= \arg\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 \right] \tag{2.104}$$

Altogether, the optimal network parameters $\hat{\boldsymbol{\theta}}$ are those which minimize the term

$$\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{N} \left(y_i - \mathrm{NN}_{\boldsymbol{\theta}}(x_i) \right)^2 \tag{2.105}$$

i.e. the sum of squared errors. Thereby the training of a neural network is framed as an optimization problem. Note, however, that the sum of squared errors in eq. (2.105) is only obtained under the assumption of independent y_i, independent errors ϵ_i, and a Gaussian error distribution with fixed variance σ^2.

Obtaining ideal network parameters by minimizing the sum of squared errors could also be justified by a more general rationale: a neural network performs best, if its parameters are chosen such that its predictions show the smallest possible difference to the targets. In practice this is expressed by means of a *loss function* $\mathcal{L}(\hat{y}, y)$, which describes the deviations between the network's prediction and the "ground truth" (the targets in the training set). Typical loss functions for regression tasks are the mean squared error (MSE, as justified by the maximum likelihood considerations above), the L_1 norm defined as $L_1(\hat{y}, y) = |\hat{y} - y|$, and slightly altered versions thereof. The minimization of a neural network's loss is a numerical problem which requires the gradient of the loss function with respect to the network's parameters. This gradient is obtained using the *backpropagation* method, which essentially applies the chain rule of calculus to the neural network, and is explained in detail in Appendix A.4.

Optimizers

With the loss gradient known, training a neural network is a matter of traversing its parameter space. The goal is to reach the global minimum of the loss landscape by means of gradient descent in an iterative process. The standard method is called *stochastic gradient descent* (SGD), which optimizes $\boldsymbol{\theta}$ based on the average gradient obtained from b training examples:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{1}{b} \sum_{i=1}^{b} \nabla_{\boldsymbol{\theta}} C_i \tag{2.106}$$

where:

- t : iteration step
- η : the step size
- $\nabla_{\boldsymbol{\theta}} C_i$: the gradient based on a single network prediction
- b : the *batch size*, meaning the number of training examples averaged over

Different batches are used at every iteration step. A full iteration over all instances of the training set is called one *training epoch*. The step size η is also called *learning rate*, whose choice is critical: If η is chosen too small, the training will be slow and possibly stagnate in a local loss minimum. On the other hand, if η is chosen too large, the training might diverge. This problem can be mitigated by the use of modern optimizers, which improve on SGD. The most common optimizer is *Adam* (for "Adaptive Moment Estimation", see Kingma and Ba, 2017), which uses an *adaptive* learning rate for each parameter by computing

$$\boldsymbol{m}_{t+1} = \beta_1 \boldsymbol{m}_t + (1 - \beta_1) \nabla_{\boldsymbol{\theta}_t} C \tag{2.107}$$

$$\boldsymbol{s}_{t+1} = \beta_2 \boldsymbol{s}_t + (1 - \beta_2) \nabla_{\boldsymbol{\theta}_t} C \otimes \nabla_{\boldsymbol{\theta}_t} C \tag{2.108}$$

$$\hat{\boldsymbol{m}}_{t+1} = \frac{\boldsymbol{m}_{t+1}}{1 - \beta_1^{t+1}} \tag{2.109}$$

$$\hat{\boldsymbol{s}}_{t+1} = \frac{\boldsymbol{s}_{t+1}}{1 - \beta_2^{t+1}} \tag{2.110}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \hat{\boldsymbol{m}}_{t+1} \oslash \sqrt{\hat{\boldsymbol{s}}_{t+1} + \epsilon} \tag{2.111}$$

$$\tag{2.112}$$

where:

$\boldsymbol{m}$: the *first moment* estimate, initialized as 0

$\boldsymbol{s}$: the *second moment* estimate, initialized as 0

$\hat{\boldsymbol{m}}$: the bias-corrected first moment estimate

$\hat{\boldsymbol{s}}$: the bias-corrected second moment estimate

β_1, β_2 : exponential decay rates of the momentum estimates

ϵ : a small number to provide numerical stability, e.g. $\epsilon = 10^{-8}$

$\otimes, \oslash$: element-wise multiplication and division

Usual choices for the decay parameters are $\beta_1 = 0.9$, $\beta_2 = 0.999$. Other variants of Adam exist, which deploy further optimizations, i.e. *Nesterov Adam* (NAdam, see Dozat, 2017) and *AdamW* (see Loshchilov and Hutter, 2019).

2.5.4 Regularization

The high dimensionality of the loss landscape poses a major hurdle in the pursuit of finding its global minimum by means of gradient methods. In particular the training process can be noisy or slow, and it can stagnate, diverge, converge to sub-optimal local minima, or overfit. There exist a plethora of *regularization* methods, intended to suppress such irregularities. In the development of NitroNet, the following methods were tested:

- *Batch normalization* ("batch norm"), where the inputs to each network layer are re-centered and re-scaled. Batch norm allows for higher learning rates, thus speeding up the training process. See Ioffe and Szegedy (2015).

- *Dropout*, which means randomly disabling individual neurons of the network with a probability p_{dropout}. Dropout demonstrably reduces overfitting, by forcing the neural network to find a solution, that depends on the ensemble of neurons rather than only a few of them. The dropout mechanism is deactivated at runtime. See Srivastava et al. (2014).

- *Data transformations*, where the input features undergo bijective transformations, which bring them to a mutual scale and normalize long-tailed distributions. The regularizing effect of this procedure corresponds to the findings of e.g. Ioffe and Szegedy (2015). Furthermore, data transformations prevent individual features (e.g. with an exceptionally broad distribution of values) from dominating the neural network's loss. The transformations are fitted to the training data, and are then applied to the instances at runtime.

Typical transformations are simple shifting and scaling, the Yeo-Johnson transformation (see Yeo and Johnson, 2000), and the quantile transformation (see Pedregosa et al., 2011). Data transformations can also be applied to the training targets.

Batch norm and dropout were tested, because neural networks with a large number of input variables and comparably few training examples are generally more prone to overfitting. As shown later in Chapter 4, this turned out not to be the case for NitroNet, which ultimately performed best without batch norm and dropout. Data transformations were considered based on an analysis of the probability distributions of the input features, some of which were found to be strongly skewed or long-tailed. Note, that other components of neural networks are known to have additional regularizing side-effects, e.g. some activation functions (see e.g. Misra, 2020).

2.5.5 Hyperparameter optimization

The majority of parameters in a neural network are trainable parameters, summarized in $\boldsymbol{\theta}$. These are the parameters optimized during the training process. The term *hyperparameter* refers to network parameters, which are fixed during training and cannot be optimized via backpropagation. Examples of hyperparameters are the number of layers and neurons, the learning rate, the loss function, the activation function, the optimizer, the choice of regularization methods and transformations, etc.

Hyperparameters are commonly optimized in a "brute force" approach. This means, that many neural network variants with different hyperparameters are trained and compared by some metric (e.g. loss at the end of training), from which an ideal combination of hyperparameters can be determined. The usual methods are *grid search*, where hyperparameters from a manually selected set of candidates are tested, and *random search*, where hyperparameters are sampled randomly from a subset of the hyperparameter space, see Bergstra and Bengio (2012). There exist further more complex Bayesian methods for the efficient exploration of the hyperparameter space, see e.g. Wu et al. (2019). Besides the training set and the test set, the process of hyperparameter optimization necessitates a third split, called the *validation set*. The neural network variants are trained on the training set, and compared to each other on the validation set, from which the ideal hyperparameters are determined. Then, the final neural network is evaluated on the test set. The use of a validation set "in-between" ensures, that the hyperparameters are not overfitted to the test set, which is supposed to remain entirely independent for the final model evaluation.

2.5.6 Computation of Shapley scores, feature relevance

An important aspect of neural network analysis is to compute its *feature relevance*. This means to find estimates of how important each input feature is for the overall prediction quality. The standard method for feature relevance analysis is based on the computation of the *Shapley scores*, named after the game theorist Lloyd Shapley (see Shapley, 1951). The Shapley scores describe, how much a single "player" (here: input feature) contributes to the "success" (here: prediction quality) of a "coalition" (here: subsets of input features) within a "game" (here: the task of making predictions). The Shapley score of the i-th feature x_i is defined as

$$R_i = \sum_{S \subseteq P \backslash \{x_i\}} \underbrace{\frac{|S|!\,(|P| - |S| - 1)!}{|P|!}}_{\text{weighting factor}} \left(f\left(S \cup \{x_i\}, \boldsymbol{y}, \hat{\boldsymbol{y}}\right) - f\left(S, \boldsymbol{y}, \hat{\boldsymbol{y}}\right) \right) \tag{2.113}$$

where:

P	: the set of all features
S	: the coalition
$f(S, \boldsymbol{y}, \hat{\boldsymbol{y}})$	: a function of choice, which acts as a measure for the performance of coalition S
$\boldsymbol{y}$	: the ground truth
$\hat{\boldsymbol{y}}$	: the neural network's predictions

Note, that the scaling factor is sometimes defined differently, depending on the context. Equation (2.113) shall be explained by a simple example: Suppose a neural network with 4 features (x_1, x_2, x_3, x_4). In order to probe the importance of feature x_1, a coalition $S \subseteq P \backslash \{x_i\}$ is picked, e.g. $\{x_2, x_3\}$. The neural network is then evaluated twice: once using the features $S = \{x_2, x_3\}$, and once using the features $S \cup \{x_i\} = \{x_1, x_2, x_3\}$. The performance of the neural network is described by some function f. A suitable choice for f could be a normed root mean squared error (RMSE), e.g.

$$f(S, \boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\mathrm{RMSE}(S, \boldsymbol{y}, \hat{\boldsymbol{y}}) - \mathrm{RMSE}(\emptyset, \boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{RMSE}(P, \boldsymbol{y}, \hat{\boldsymbol{y}}) - \mathrm{RMSE}(\emptyset, \boldsymbol{y}, \hat{\boldsymbol{y}})} \tag{2.114}$$

where $\emptyset$ denotes the empty set. With this definition, f maps a fully informed neural network (using all input features) to a value of 1, and an entirely uninformed neural network (using no input features) to a value of 0. The difference $f\left(S \cup \{x_1\}, \boldsymbol{y}, \hat{\boldsymbol{y}}\right) - f\left(S, \boldsymbol{y}, \hat{\boldsymbol{y}}\right)$ measures by how much the neural network's accuracy is reduced when using only the input features $S = \{x_2, x_3\}$ as opposed to $S \cup \{x_1\} = \{x_1, x_2, x_3\}$. Omission of a feature (e.g. x_1 in the preceding example) is simulated by shuffling its corresponding column in the feature matrix

X. The Shapley score R_1 of the feature x_1 is obtained by computing the weighted sum over all coalitions $S \subseteq P \backslash \{x_1\}$. The procedure is then repeated for the remaining features.

Unfortunately, the summation over a power set in eq. (2.113) poses a significant computational hurdle. For example, in order to obtain all Shapley scores for a neural network with 30 features, meaning $|P| = 30$, a total of $|P| \cdot 2^{|P|-1} = 30 \cdot 2^{29}$ (approx. 16 billion) summands must be computed, each of which requires a full evaluation of the neural network on the validation or test data. The computational burden can be reduced by

- grouping features, e.g. in the case of NitroNet, wind speeds at different altitudes. A group of features is treated like a single feature in eq. (2.113).
- reducing the full sum over $S \subseteq P \backslash \{x_i\}$ to a sum over a representative subset. This can be achieved in an iterative approach, like so: All R_i are initialized as zero. A random feature index $i \in 1, ..., |P|$ is sampled randomly. For the corresponding feature x_i, a random coalition $S \subseteq P \backslash \{x_i\}$ is sampled randomly. The corresponding summand according to eq. (2.113) is computed, and added to R_i. The iteration ends, once the feature relevances (defined here as $F_i = R_i / \sum_{j=1}^{|P|} R_j$) have converged for all i.

Both approaches were combined in the feature relevance analysis of NitroNet.

2.5.7 Uncertainties in machine learning

Training a neural network is associated with different types of uncertainties and errors. Consider a neural network, whose predictions $\hat{\boldsymbol{y}} = \mathrm{NN}(\boldsymbol{x})$ are supposed to approximate a "true" function $\boldsymbol{y} = g(\boldsymbol{x})$. The main types of uncertainty can be classified as follows:

- *Model misspecification error*: The chosen model family $\mathcal{F}$ does not contain the true mechanism g, meaning that

$$\arg\min_{\mathrm{NN} \in \mathcal{F}} \mathcal{L}(\mathrm{NN}, g) \neq g \tag{2.115}$$

  Here, $\mathcal{L}(\mathrm{NN}, g)$ means the loss between the predictions of the neural network NN, and the true function g, evaluated on the same input $\boldsymbol{x}$; here, and in the following, this simplified notation is used deliberately for easier readability. Model misspecification errors can occur, e.g. if a machine learning model lacks complexity (e.g. linear models acting on non-linear problems), but is rarely the case with modern neural networks.

- *Epistemic error*: Knowledge of the true mechanism g is limited to a *finite* training set, hence

$$\arg\min_{\mathrm{NN} \in \mathcal{F}} \mathcal{L}(\mathrm{NN}, \mathrm{TS}) \neq \arg\min_{\mathrm{NN} \in \mathcal{F}} \mathcal{L}(\mathrm{NN}, g) \tag{2.116}$$

where TS denotes the training set.

- *Optimization error*: The optimization procedure may be incapable of finding the global loss minimum, meaning that

$$\underset{\mathrm{NN}\in\mathcal{F}}{\text{algorithmic min}}\,\mathcal{L}(\mathrm{NN},\mathrm{TS}) \neq \underset{\mathrm{NN}\in\mathcal{F}}{\arg\min}\,\mathcal{L}(\mathrm{NN},\mathrm{TS}) \tag{2.117}$$

Further uncertainties exist in the probabilistic case, e.g. in the context of ill-posed inverse problems. Then, there exists no unambiguous mapping $\boldsymbol{y} = g(\boldsymbol{x})$, that a neural network could reproduce; the "true" process can at best be described by a conditional probability distribution $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x})$. This gives rise to two more errors:

- *Aleatoric error*: The targets $\boldsymbol{y}$ (or the mechanism g) and the features $\boldsymbol{x}$ are noisy. For example, the process of photon counting in spectral measurements follows a Poisson distribution and is not deterministic.
- *Ambiguity error*: The features $\boldsymbol{x}$ do not contain the required information to fully recover $\boldsymbol{y}$. For example, different tropospheric trace gas profiles can produce the same tropospheric column density. Therefore, given a tropospheric column density alone, the corresponding trace gas profile cannot be recovered without ambiguity.

The probabilistic case can be resolved by probabilistic machine learning models, e.g. the invertible neural networks. Feed-forward neural networks, however, are deterministic, which means that $p(\boldsymbol{y}|\boldsymbol{x})$ is effectively reduced to a single value. This can be the maximum a posteriori, a weighted mean, or the median, depending on the characteristics of the distribution and hyperparameters such as the loss function.

Monte-Carlo input uncertainty propagation

The uncertainties explained up to this point refer to the question of how well a neural network can be trained to reproduce a true mechanism under given circumstances. Even a "perfectly" trained neural network, i.e. an exact copy of the mechanism g, suffers from an additional kind of uncertainty at runtime. This is the uncertainty of the input variables $\boldsymbol{x}$, which propagates through the neural network. The satellite data used in this thesis, for example, have an intrinsic measurement uncertainty. The a priori assumptions and model uncertainties, that go into the satellite retrieval, further add to this uncertainty. Its propagation through a neural network can be quantified with a Monte-Carlo approach as described by Anderson (1976).

If the input $\boldsymbol{x}$ to a neural network has no uncertainty, we may write $\hat{\boldsymbol{y}} = \mathrm{NN}(\boldsymbol{x})$. If $\boldsymbol{x}$ *does* have uncertainty, it can be modelled by some probability distribution, i.e.

$$\boldsymbol{x} \sim p(\boldsymbol{x}) \tag{2.118}$$

Then, an arbitrary number of samples $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ can be drawn from $p(\boldsymbol{x})$, and the corresponding neural network predictions $\{\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N\}$ can be computed. $\{\hat{\boldsymbol{y}}_1, ..., \hat{\boldsymbol{y}}_N\}$ can be used to obtain statistical diagnostics, such as error bands of the predictions. Note, that systematic errors (e.g. due to faulty implementations of radiative transfer for the computation of air mass factors) cannot be quantified using this method.

2.5.8 A review of ML models for the prediction of NO_2

In closing the fundamentals of this thesis, an overview of existing machine learning models in the context of predicting NO_2 air pollution from satellite data is given. The commonality of the existing literature is the limitation to empirical training sets, meaning that the proposed neural networks were trained to reproduce NO_2 surface concentrations from in situ measurements. Essentially, these studies differ by the choice of machine learning model and input variables.

- Gardner and Dorling (1999) train a feed-forward neural network for the prediction of hourly NO_2 and NO_x. The model was trained at a fixed location, and only uses meteorological data as input but no satellite observations. Correspondingly, it is expected to adapt to the specific properties of the chosen location, and cannot be reasonably applied elsewhere, which poses a strong limitation to its general usability. A correlation coefficient of $R = 0.69$ was achieved.

- Kang et al. (2021) predict surface NO_2 concentrations in Asia using different machine learning models, such as *random forests*, but no neural networks. Like all studies mentioned in the following, the models are trained on targets from in situ measurements, and use TROPOMI satellite observations as the main input. Correlations of up to $R = 0.84$ were achieved. In a feature relevance study, the TROPOMI NO_2 VCD was identified as most important, with a relevance of approx. 30 %.

- Chan et al. (2021) present a study very similar to Kang et al. (2021), but with a feed-forward neural network and in situ data from Germany instead. A correlation coefficient of $R = 0.80$ was achieved. The feature relevance of the TROPOMI NO_2 VCD was estimated around 27 %.

- Ghahremanloo et al. (2021) deploy a similar method using a convolutional neural network and in situ data from Texas, US. The use of a convolutional model allows for the aggregation of input data from nearby satellite observations in an efficient manner. A correlation coefficient of $R = 0.91$ was achieved. The feature relevance of the TROPOMI NO_2 VCD was estimated around 15 %. This is a comparably low value, most likely due to the use of a somewhat redundant "road density" variable of similar feature relevance.

- Zhang et al. (2022) present a model based on the ResNet architecture (a convolutional neural network with "skip connections", see He et al., 2015), trained on in situ data from China. A correlation coefficient of $R = 0.94$ was achieved. The feature relevance of the TROPOMI NO_2 VCD was estimated around 18 %.

- Cao (2023) use a similar setup as Ghahremanloo et al. (2021) with in situ data from the US and a much smaller convolutional neural network than Zhang et al. (2022). A correlation coefficient of $R = 0.94$ was achieved. The NO_2 VCD is claimed to be the most relevant feature, but the percentual feature relevance is not explicitly given. Furthermore, the model was made nearly independent from satellite observations by leveraging a plethora of redundant infrastructural variables (e.g. very detailed information on adjacent road distances, etc.).

Altogether, these studies give a clear picture on the current state of the art: Surface NO_2 concentrations can be reliably predicted from satellite observations using neural networks, and the prediction quality varies with the choice of model and input variables. A significant drawback of the presented models is their inability to account for the Mo-CL bias (see sect. 2.3.6). Furthermore, they can only predict NO_2 concentrations at the surface, as opposed to full concentration profiles. The NitroNet model presented in this thesis can overcome these shortcomings by training on synthetic model data, which includes the necessary information.

Chapter 3

Regional chemistry and transport modelling with WRF-Chem

The first step in the development of NitroNet is to generate a large training data set of NO_2 profiles and colocated input variables. This section describes the process of running the RCT model WRF-Chem for this purpose. However, obtaining realistic NO_2 profiles from RCT simulations is no triviality. As discussed in sect. 2.4, even state of the art simulation setups suffer from significant deviations from observational reference data, in particular the surface in situ measurements. A substantial amount of the research presented in this thesis has revolved around searching for the possible roots of these deviations, by experimenting with different changes to the model setup and intercomparing the results. The following chapter guides the reader through this analysis with the intent of documenting the underlying thought process, and making the findings accessible to the WRF-Chem modelling community. Lastly, an analysis on which model configuration was found to be most suitable for the generation of NitroNet training data is provided. The contents of this chapter correspond in large parts to Kuhn et al. (2024a).

3.1 General simulation setup

To begin with, the model setup used throughout this chapter is summarized. The WRF-Chem model (version 4.2.2, see Grell et al., 2005) is operated on a twofold nested domain in central Europe (see Fig. 3.1) for the month of May 2019. The spatial resolutions of the outer and inner domain (called D1, and D2 from hereon) are 15 km $\times$ 15 km on a grid of 320×245 cells, and 3 km $\times$ 3 km on a grid of 500×430 cells with 43 terrain-following vertical layers. The corresponding average layer extents in the lowest 5 km are given in Table 3.1. The exact layer bottoms and tops depend on location and time. Both domains require meteorological and chemical initial and boundary conditions. The chemical boundary conditions of D2 are taken

from the global CAM-Chem model (see Emmons et al., 2020). The inner domain D1 obtains model-consistent boundary conditions from the outer domain D2. Both domains receive their meteorological initial conditions from the ERA5 reanalysis dataset (see Hersbach et al., 2020). Spectral nudging to ERA5 meteorological data is deployed on D2. The simulation has a time step of 60 seconds, and model output is saved in hourly intervals.

Simulation schemes

layer	extent [m]	layer	extent [m]
1	0 – 8	13	1370 – 1546
2	8 – 33	14	1546 – 1697
3	33 – 66	15	1697 – 1841
4	66 – 125	16	1841 – 1937
5	125 – 209	17	1937 – 2035
6	209 – 310	18	2035 – 2183
7	310 – 429	19	2183 – 2374
8	429 – 575	20	2374 – 2661
9	575 – 741	21	2661 – 3142
10	741 – 935	22	3142 – 3907
11	935 – 1178	23	3907 – 4762
12	1178 – 1370	24	4762 – 5643

Table 3.1: Layer extents of the lowest 24 layers in the WRF-Chem simulation.

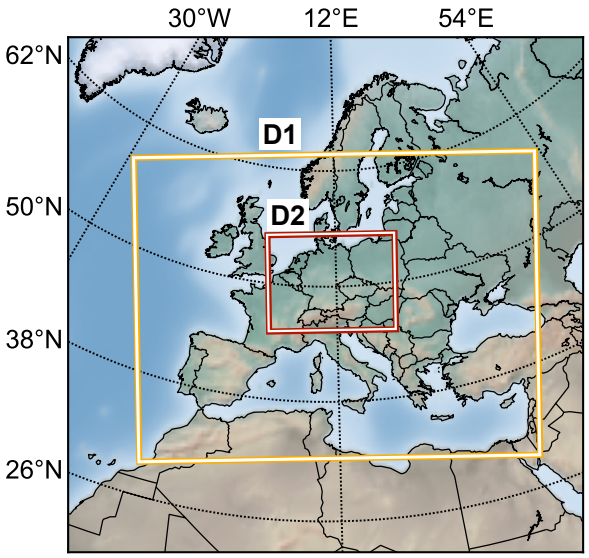

Figure 3.1: Model domains of the WRF-Chem simulation. All results presented in the following are taken from the inner domain D2.

The simulation uses the Thompson microphysics scheme (see Thompson et al., 2008), the RRTMG (Rapid Radiative Transfer Model for General Circulation Models long- and shortwave radiation scheme, see Iacono et al., 2008), the Monin-Obukhov similarity surface layer scheme (see Monin and Obukhov, 1954), the NOAH Land-Surface Model (see Niu et al., 2011), the Yonsei University boundary layer scheme (YSU, see Hong, 2010), and the Grell-Dévényi ensemble scheme for cumulus modelling (see Grell and Dévényi, 2002). Dry deposition is modelled following Wesely (1989). For photochemistry, the MOZART chemical mechanism (see Emmons et al., 2010) is coupled to the GOCART aerosol mechanism (see Chin et al., 2000) along with the TUV full photolysis scheme (see Madronich, 1987; Tie et al., 2003).

Emission inventories

The EDGARv5 global emission inventory (see Crippa et al., 2020, horizontal resolution of $0.1° \times 0.1°$, referring to the year 2015) is used for anthropogenic emissions on D1 and D2,

except for Germany, where the UBA-E regional emission inventory (see Hausmann et al., 2020, horizontal resolution of $0.01° \times 0.01°$) is used instead. Because UBA-E does not include VOCs, those are taken from EDGARv5 over the German domain. Biomass burning emissions are taken from the Fire Inventory from NCAR (FINN, see Wiedinmyer et al., 2011) with a spatio-temporal resolution of 1 km × 1 km × 24 h. FINN includes suitable temporal profiles and plume-rise parametrizations according to Freitas et al. (2007). Biogenic emissions are computed online using an implementation of the MEGAN model (see Guenther et al., 2006).

Temporal, vertical, and speciation profiles

The model configuration uses the temporal emission profiles from Kumar et al. (2021), which augment the established profiles, e.g. from Crippa et al. (2020), by measurements of car traffic in Germany. This yields arguably more realistic traffic emissions, particularly on the weekends. Figures B.1, B.2, and B.3 give an overview of the hourly, daily, and monthly emission profiles. The vertical emission profiles recommended by Bieser et al. (2011) are deployed, see Fig. B.4. Speciation profiles for non-methane organics are taken from Huang et al. (2017). Moreover, all NO_x emissions are speciated as 87.5 % NO and 12.5 % NO_2. This reflects the literature reports on NO_x speciation from combustion processes, ranging from approx. $5 - 40$ % (see Jimenez et al., 2000; Costantini et al., 2016; Wild et al., 2017; Richmond-Bryant et al., 2017).

3.2 Evaluation of a standard simulation run (S-YSU)

Here, a brief overview of the results from a "standard" simulation run ("S-YSU", in reference to the YSU boundary layer scheme) is presented. This run uses the model configuration described above, except for placing all emissions at the surface, as opposed to distributing them vertically. This aspect of the simulation is explored separately in sect. 3.6. It can be considered similar to the configurations used in recent literature, such as Poraicu et al. (2023) or Kuik et al. (2016). The aim is to give the reader a coarse outline of the model's performance with respect to the two main reference datasets available: The AirBase in situ measurements of the surface NO_2 concentration, and the TROPOMI satellite observations of the tropospheric NO_2 VCD. From this preliminary evaluation the need for a model optimization is motivated. In that pursuit, the validation is limited to the German model subdomain, on which said model optimization can be ideally conducted due to the exceptionally high horizontal resolution of the available emission data. A detailed model evaluation on the full domain (D2), comparison to MAX-DOAS data, and quantitative summaries in tabular form are presented in sect. 3.7 (in two versions; based on monthly-mean data and without any averaging). Here and in the following, the statistical diagnostics mentioned in the main text refer to the monthly means (e.g. RMSE values computed on monthly mean values of observations and model predictions),

unless specified otherwise.

3.2.1 Comparison to AirBase in situ measurements

At first, the simulation results are compared to AirBase in situ measurements at the surface. For this purpose the modelled NO_2 concentrations of the lowest layer (approx. $0 - 8$ m) are interpolated bilinearly to the locations of the in situ measurements. Because the in situ measurements represent hourly averages, matching the resolution of the model output, no temporal interpolation is required. Figure 3.2 shows the time series of the modelled and observed surface concentrations of NO_2, NO, NO_x, and O_3 at background measurement locations (for reference, see sect. 2.3.6). The left-side time series are restricted to the first 10 days of the simulation for easier readability. For NO_2 and NO_x, the Mo-CL correction factors F were computed according to eq. (2.78) using the simulated VMRs of PAN and HNO_3. However, instead of dividing the in situ measurement results by F, as suggested by eq. (2.79) and (2.80), the simulation results are multiplied by F instead. This way, all subsequent model runs can be compared to the same reference time series. The correction is only applied, where the molybdenum-based chemiluminescence method is used. Elsewhere, no correction is used, i.e. $F = 1$. Both time series, with and without application of F are shown.

Without correction of the Mo-CL bias the model's diurnal NO_2 cycle shows a low bias of -14.7 % at noontime, a daytime bias of $+14.9$ % and a nighttime bias of $+65.6$ %. This refers to the *mean relative bias*, defined as

$$\text{bias} = \sum_{i=1}^{N} (y_i - x_i)/x_i \tag{3.1}$$

where x_i denote the reference datapoints from AirBase, and y_i the corresponding modelled datapoints from the WRF-Chem simulations. The Mo-CL correction alleviates the noontime bias to $+5.5$ %, but increases the daytime and nighttime biases to $+29.2$ % and $+73.5$ %, respectively. The NO cycle is reproduced with similar deviations: In the transition from nighttime to sunrise (00:00 – 06:00), an overall bias of -26.3 % is observed. During daytime, NO is reproduced with a bias of -15.2 %. The NO_x cycle is dominated by NO_2, except for the morning period around sunrise (05:00), when the NO concentration peaks. With the Mo-CL correction applied, this results in a NO_x bias of -2.3 % at noontime, $+21.9$ % during the day, and $+60.2$ % during the night. Ozone is represented well throughout the first two days of the simulation, but is then overestimated by approximately 13 µg m^{-3} for the remaining simulation period. Most likely, the simulation suffers from a general tendency to overestimate ozone, which is suppressed by the initial conditions during the simulation's spin-up phase. Figure B.5 is a version of Fig. 3.2 that shows traffic measurements instead

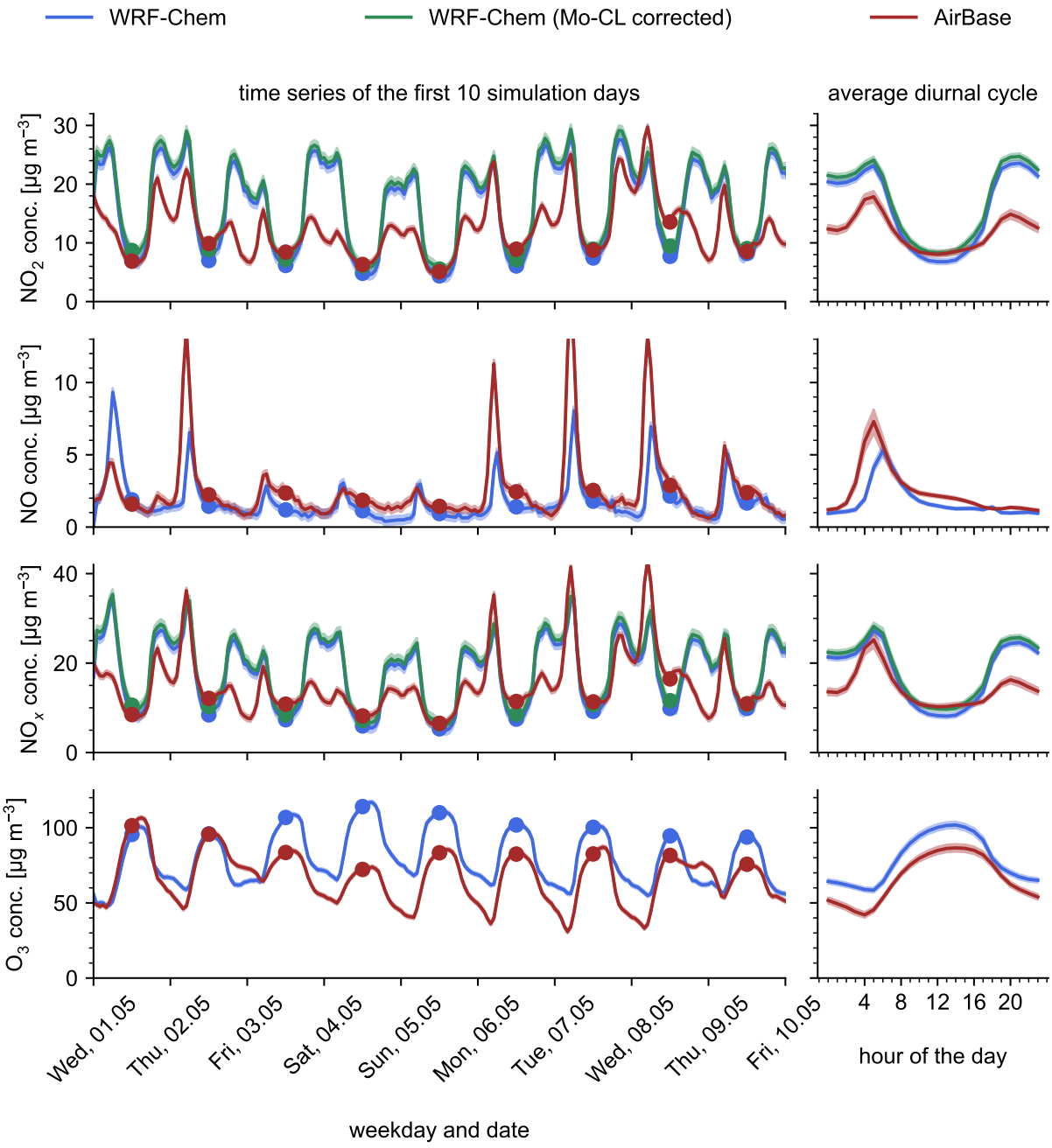

Figure 3.2: Evaluation of the simulation run S-YSU against background in situ measurements in Germany. The left-side time series shows the first 10 days of the simulation. The coloured dots represent the noon-time value of each day. The right-side plots show the average diurnal cycle of each trace gas for the entire simulation period.

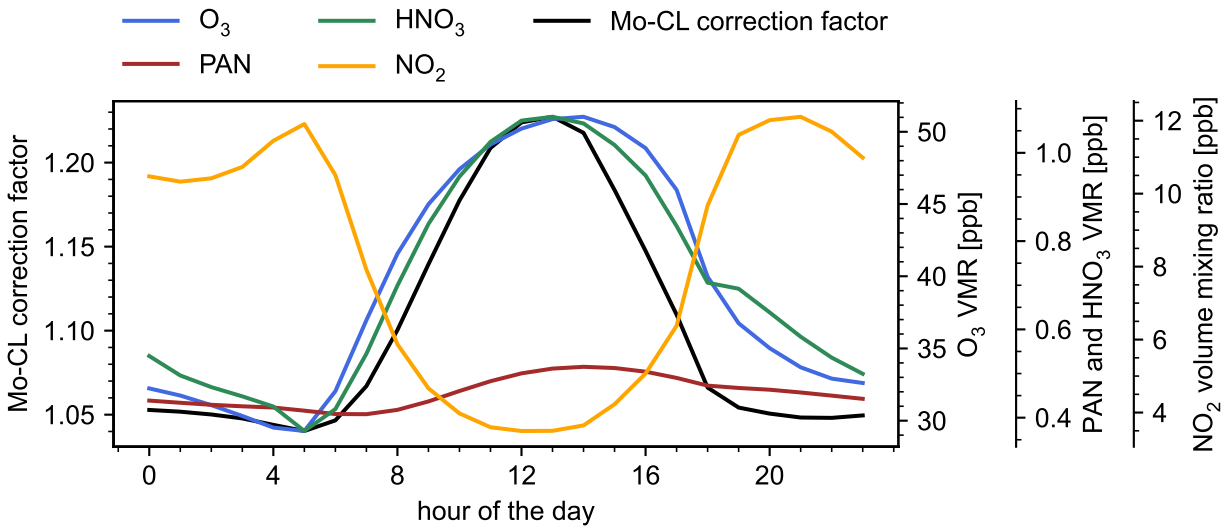

Figure 3.3: Average diurnal cycle of the Mo-CL correction factor in the simulation run S-YSU. The corresponding surface VMRs of HNO_3, PAN, NO_2, and O_3 are drawn in green, red, orange, and blue, respectively. Note, that the O_3 VMR has no effect on the Mo-CL correction factor and is only shown to demonstrate the strong correlation to PAN and HNO_3.

of background measurements. Here, drastic underestimations of NO_2, NO, and NO_x are observed, with deviations of much larger magnitude. This demonstrates, that any attempts to obtain adequate agreement to traffic measurements are futile, because the horizontal resolution of the simulation (and the emission data) are far too low. Throughout the rest of this chapter, all comparisons to in situ observations refer to background measurements.

The comparison to the in situ measurements shows, that the application of the Mo-CL correction significantly improves the model accuracy with respect to NO_2 and NO_x, particularly around noon. Figure 3.3 depicts the average diurnal cycle of the Mo-CL correction factor. As expected from the corresponding literature (see sect. 2.3.6), the diurnal cycle peaks around noon, when the VMRs of PAN and HNO_3 are high, and the VMR of NO_2 is low.

Further insight can be gained from a spatial validation of the model results, as displayed in Fig. 3.4. Figure 3.4a shows the comparison of monthly-mean noontime NO_2 surface concentrations. Besides a region of high model bias in the Ruhr region of western Germany, there appears to be no distinct spatial pattern in the over- and underestimations of the simulation. It was decided to show noontime values here (as opposed to other times of the day) in order to maintain intercomparability to other model evaluation papers in literature. Corresponding maps at other times of the day show the same qualitative spatial patterns, although modulated with the mean diurnal NO_2 cycle displayed in Fig. 3.2. An example of the comparison at 4 PM can be found in Fig. B.6.

3.2.2 Comparison to TROPOMI satellite observations

In order to validate the model results against tropospheric NO_2 VCDs from the TROPOMI satellite measurements, the simulated NO_2 profiles are interpolated horizontally and temporally to the locations of the satellite observations, and vertically to the pressure levels on which the averaging kernels are defined. Bilinear interpolation is used throughout. Data from the processor version v02.04.00 is used, and all observations with a quality score of $f_{\mathrm{QA}} \leq 0.75$ are dismissed. The simulated NO_2 VCDs are computed as described in sect. 2.4.3, eq. (2.84). These are then compared to the original VCDs from TROPOMI (see Fig. 3.4b, mean bias of +38.2 %) and those obtained by re-computing the air mass factors using the modelled NO_2 profiles as described in sect. 2.3.4 (see Fig. 3.4c, mean bias of +12.1 %). The results demonstrate the significant impact of re-computing the air mass factors with high-resolution NO_2 profiles. The largest model errors occur in western Germany, and reach up to 10^{16} molec. cm^{-2}. Intermediate model errors on the scale of $0.5 \cdot 10^{16}$ molec. cm^{-2} can be observed near large cities close to the Rhine river, e.g. in Frankfurt and Mannheim, and other locations, such as the Hamburg harbor and the nearby coal power plant Moorburg. The remaining parts of the domain show biases of $0.5 \cdot 10^{16}$ molec. cm^{-2} or lower. The errors of the modelled VCDs and surface concentrations correlate in some parts of the domain, particularly where the VCD errors are the largest. On the other hand, the model errors of the surface concentrations are mostly too noisy to identify a clear correlation to the overestimations of the VCD.

3.2.3 Intermediate conclusions

The preceding analysis allows to draw intermediate conclusions. Overall, the simulation results of S-YSU can be described as qualitatively similar to simulations reported in the previously mentioned literature (see Terrenoire et al., 2015; Visser et al., 2019; Kuik et al., 2016; Kuik et al., 2018; Poraicu et al., 2023). The model's diurnal NO_2 surface concentration shows a similar shape as reported in Kuik et al. (2016) and Poraicu et al. (2023) (see also Fig. 2.17), with slight underestimations at noon and strong overestimations at night. Correction of the Mo-CL bias strongly reduces the noontime bias, from −14.7 % to +5.5 %. With values of $F \approx 1.22$, the noontime correction factors are close to the results of Poraicu et al. (2023), who obtained values of $F \approx 1.25$. Re-computing TROPOMI's air mass factors using the modelled NO_2 profiles results in significantly better agreement between the simulated and observed NO_2 VCDs. While the literature discussed in sect. 2.3.4 reports corresponding enhancements of the TROPOMI NO_2 VCDs on the scale of 15 %, a larger average enhancement of 26.1 % is obtained here. This could be attributed to the strongly polluted regions of the model domain, where the effect is expected to be largest, or to the particularly high resolution of the emission data. With this, the two main deficits of the simulation can be identified as:

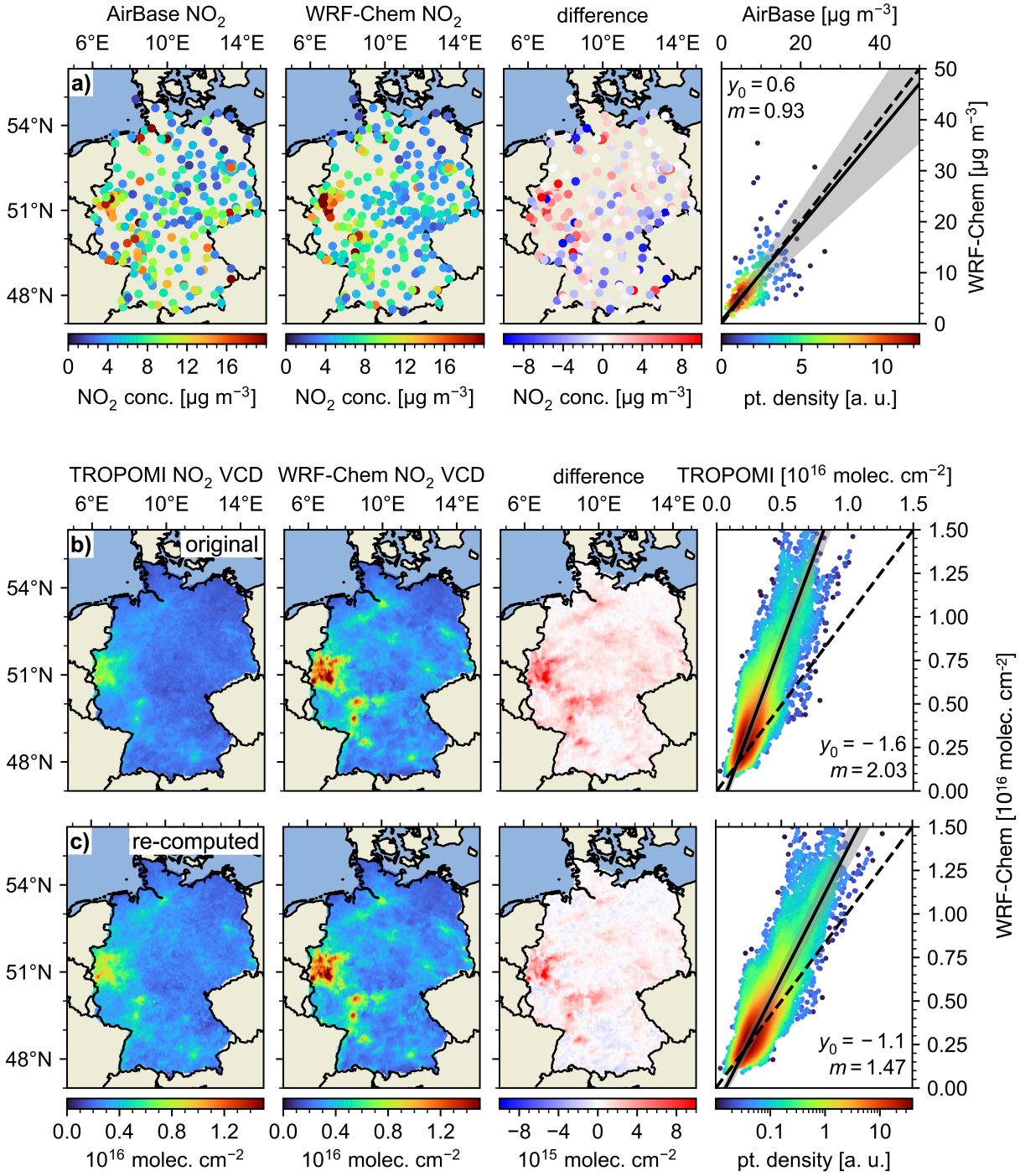

Figure 3.4: Validation of the simulation run S-YSU against monthly-mean in situ and satellite measurements from AirBase and TROPOMI. a) Comparison to noontime NO_2 in situ measurements from the AirBase network. The Mo-CL correction was applied. **b), c)** Comparison to NO_2 VCDs from TROPOMI without/with re-computation of the air mass factors using the modelled NO_2 profiles. Intercepts (y_0) are given in units of of µg m^{-3} for surface concentrations, and 10^{15} molec. cm^{-2} for NO_2 VCDs.

1. The mismatch in the diurnal cycle of NO_2 and NO_x surface concentrations, as shown in Fig. 3.2.

2. The overestimations of the NO_2 VCD on large parts of the model domain, as shown in Fig. 3.4.

Both are investigated in the following sections, starting with the mismatch in the diurnal cycles of NO_2 and NO_x. These can be affected by a multitude of model components, and an exhaustive analysis would be immensely complex. However, because the issues relate partially to diurnal cycles, it is reasonable to search for possible solutions in model routines that repeat on a daily basis with little day-to-day variations. In the following, two such routines are investigated: The diurnal modulation of emissions by temporal emission profiles, and the model's implementation of vertical mixing.

3.3 Simulations with tuned temporal emission profiles

In a first attempt at optimizing the simulation setup, a tuning of the temporal profiles of the most dominant emission sectors is conducted. This is motivated e.g. by the findings of Kumar et al. (2021), which demonstrate that NO_2 distributions in regional simulations are highly sensitive to the temporal emission profiles of specific sectors (in particular, the traffic sector).

Based on advice from the UBA (see Appendix B.1), the Mo-CL bias was not considered when the temporal profiles were optimized. However, the Mo-CL bias correction is applied to the final results shown in this subsection and taken into consideration in the interpretation of the results. Because the UBA emission inventory for the year 2019 was not published at that time, the optimization of the temporal emission profiles was conducted on simulation data of the year 2018. A simulation run for the year 2019 with the optimized temporal profiles is shown at the end of this section.

The idea behind tuning temporal emission profiles is to bring the model's diurnal cycles of surface NO_2 and NO_x to better agreement with the in situ measurements by redistributing emissions over the course of the day. As evident from Fig. 3.2, this requires to shift large portions of the NO_x emissions from nighttime to daytime. The tuning is conducted within four consecutive simulation runs, each spanning the first two weeks of May 2018. The first run (S-YSU-2018-1) uses the temporal emission profiles from Kumar et al. (2021). In the subsequent runs (S-YSU-2018-2, S-YSU-2018-3, S-YSU-2018-4), the diurnal profiles of the sectors "traffic", "power industry", "agricultural soils", "energy for buildings" and "manufacturing industry" are iteratively adjusted. Ideally, these adjustments would be determined e.g. by gradient descent, upon estimating the gradient of the model errors in surface NO_x with respect to the temporal profiles. However, this is not a feasible option, because the temporal

name	time frame	temporal profiles
S-YSU-2018-1	1 – 14 May 2018	as described in Kumar et al. (2021)
S-YSU-2018-2	1 – 14 May 2018	tuned (intermediate step 1)
S-YSU-2018-3	1 – 14 May 2018	tuned (intermediate step 2)
S-YSU-2018-4	1 – 14 May 2018	tuned (final)
S-YSU-TP	May 2019	tuned (final)

Table 3.2: WRF-Chem simulation configurations for the tuning of temporal emission profiles. For "intermediate step 1" and "intermediate step 2", refer to Fig. 3.5.

profiles possess too many degrees of freedom, and individual simulation runs are computationally too expensive. It was therefore decided to optimize the temporal profiles "empirically", meaning that the adjustments in between each run were guessed, while conserving the approximate shape of the temporal emission profiles as best as possible. The resulting temporal profiles are evaluated in a simulation run (S-YSU-TP) spanning the entire month of May 2019, see Table 3.2.

Figure 3.5 gives an overview of the optimization process. As shown in Fig. 3.5h, each iteration reduces the model's overall error in the diurnal NO_x cycle. Figure 3.5i shows the corresponding relative bias to the in situ measurements, whose daytime values are reduced from up to approximately -45 % in S-YSU-2018-1 to ± 25 % in S-YSU-2018-4. Similar improvements are observed at nighttime, where the bias is reduced from approximately $+25$ % in S-YSU-2018-1 to approximately ± 10 % in S-YSU-2018-4. Although the diurnal NO_x bias of the simulation can be significantly reduced by empirical tuning of the hourly temporal profiles, Fig. 3.5h and Fig. 3.5i are also illustrative of the pitfalls associated with this procedure. For example, S-YSU-2018-3 shows a significantly higher bias than S-YSU-2018-2 at 7 AM, which could be naively attributed to the traffic emissions during the morning rush hour. However, the comparison of the hourly emission profiles reveals, that the differences must instead relate to either the manufacturing industry (Fig. 3.5g), or the power industry (Fig. 3.5d), because these are the only sectors in which S-YSU-2018-3 (green dotted line) features higher emissions than S-YSU-2018-2 (red dashed line) around 7 AM. This exemplifies a situation, where the relative impact of the emission sectors was misjudged in the subjective tuning process. Moreover, the diurnal cycle obtained with tuned emission profiles is overall less smooth (see e.g. S-YSU-2018-4 at 4 PM). This, too, is a consequence of subjectively guessing supposedly better emission profiles, and indicates that some of the obtained emission profiles may be unrealistic.

Figure 3.6 shows the validation of the diurnal NO_2 and NO_x cycles of the simulation runs S-YSU and S-YSU-TP against the in situ measurements. For both simulation runs the diurnal cycle with and without the Mo-CL bias correction are shown (solid/dashes lines). The corresponding noontime and average nighttime biases are given in Table 3.3. Without

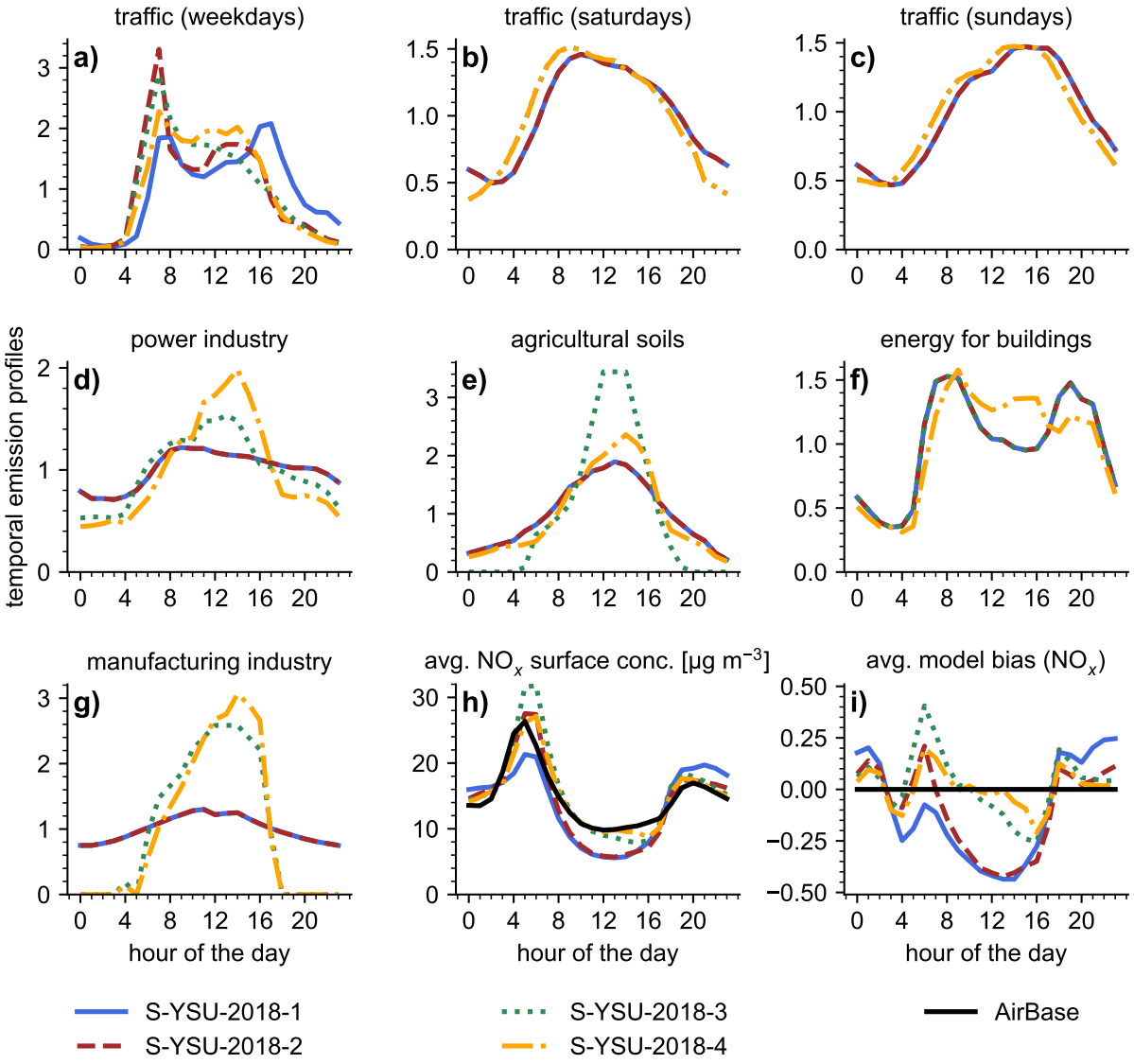

Figure 3.5: Overview of the optimization process for the hourly emission profiles based on German NO_x in situ measurements of May 2018. The run S-YSU-2018-1 uses the initial hourly emission profiles from Kumar et al. (2021). S-YSU-2018-2 and S-YSU-2018-3 represent the two intermediate steps of the optimization. S-YSU-2018-4 shows the final result of the optimization. The data shown here refer to background measurements/simulations in Germany from the first two weeks of May 2018.

name	NO_2, noontime	NO_x, noontime	NO_2, nighttime	NO_x, nighttime
S-YSU	−14.7 % (+5.5 %)	−18.3 % (−2.3 %)	+65.6 % (+73.5 %)	+53.3 % (+60.2 %)
S-YSU-TP	+6.0 % (+26.4 %)	+3.2 % (+19.3 %)	+28.2 % (+36.0 %)	+15.7 % (+22.5 %)

Table 3.3: Overview of the average noontime and nighttime NO_2 and NO_x biases of the simulation runs S-YSU and S-YSU-TP in 2019. In each cell of the table, the first entry denotes the bias without Mo-CL bias correction. The entries in brackets denote the bias with Mo-CL bias correction.

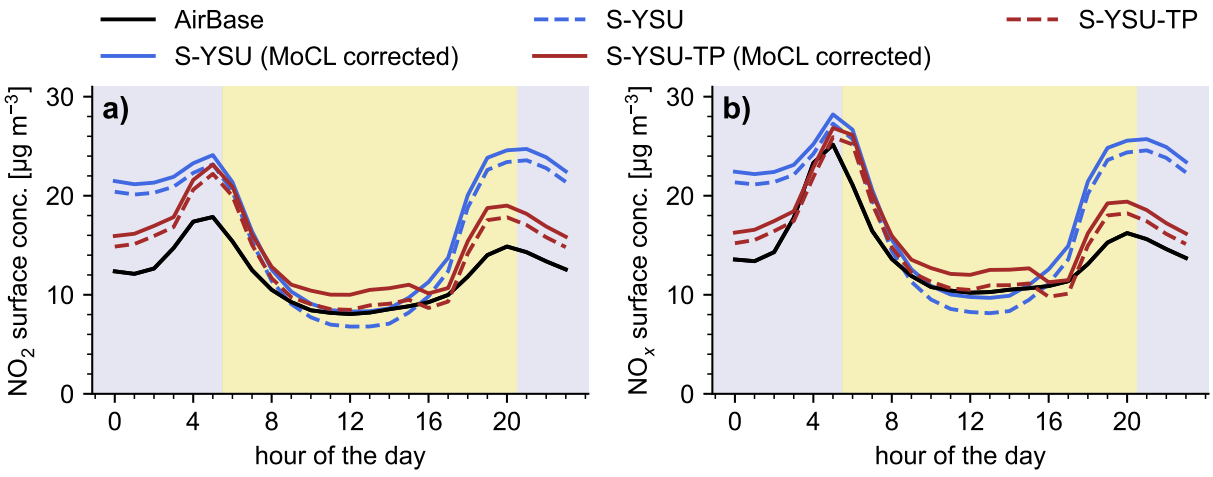

Figure 3.6: Comparison of diurnal NO_2 and NO_x cycles with original and tuned hourly emission profiles in 2019. The black solid line shows the corresponding AirBase measurements. The coloured lines represent simulation results with and without Mo-CL bias correction, respectively. The data shown refer to background measurements/simulations in Germany. The figure's background shading indicates day- and nighttime.

consideration of the Mo-CL bias, the simulation run S-YSU-TP with tuned temporal emission profiles results in significantly better agreement to the in situ measurements than S-YSU with the original temporal profiles. In particular, the noontime NO_2 underestimations of -14.7 % in S-YSU can be entirely compensated. On the other hand, application of the tuned temporal profiles has an adverse effect, if the Mo-CL bias is accounted for. In that case the original simulation run S-YSU already shows good noontime agreement and leaves essentially no room to reasonably redistribute emissions in the first place. Correspondingly, S-YSU-TP overestimates the NO_2 and NO_x noontime concentrations by $+26.4$ %, and $+19.3$ %. However, S-YSU-TP still performs considerably better at nighttime.

Overall, the results demonstrate that tuning the diurnal emission profiles is only helpful under the premise, that no Mo-CL bias correction is required. Although, as evident from Appendix B.1, there exists some controversy on the topic, the scientific literature referenced in sect. 2.3.6 clearly substantiates the relevance of the Mo-CL bias. The possibilities of improving the model results by tuning of the diurnal emission profiles was therefore not investigated further. Nonetheless, the results presented up to this point are helpful in giving a coarse estimate of the simulation's sensitivity to the choice of temporal profiles. In other simulation configurations, e.g. on domains where the available emission data can be expected to be outdated, tuning of the diurnal emission profiles, or even scaling of the emission inventory could be reasonable.

3.4 Revision of WRF-Chem's boundary layer schemes and vertical mixing

The analysis of diurnal emission profiles as a possible means to reduce the errors in the model's diurnal NO_2 and NO_x cycles has arguably lead to a dead end. If the literature consensus on the significant impact of the Mo-CL bias is trusted, an alternative method for model optimization is required. Referring back to Fig. 3.4a and 3.4c, a significant portion of the model domain is characterized by underestimations of the noontime NO_2 surface concentration, with no corresponding underestimations of the colocated NO_2 VCDs. This implies, that in many locations the tropospheric NO_2 load is accurately estimated by the model, but incorrectly distributed along the vertical axis, hinting towards a possible issue with the model's vertical mixing procedure. Note, that even when vertical emission profiles are used, the majority of NO_x emissions still occurs at the surface. If vertical mixing is suppressed, convection of the near-surface trace gases is inhibited, leading to larger surface concentrations. Vice versa, enhanced mixing results in lower trace gas surface concentrations. Besides, vertical mixing shows strong diurnal variability, which may explain the aforementioned differences between simulated and observed surface NO_2. Similar notions were raised, e.g. by Du et al. (2020), Poraicu et al. (2023), and Im et al. (2015), who recognized that trace gas surface concentrations are highly sensitive to the model's PBL scheme, which computes the strength of turbulent diffusion in the model.

In WRF-Chem, vertical mixing is a two-step procedure: First, the PBL scheme computes mixing coefficients k_h (called "`ekmfull`" in the source code), which represent the turbulent diffusion coefficients in eq. (A.41). Then, a mixing routine is called, which dilutes the trace gas concentrations of each model layer based on k_h. However, WRF-Chem implements an intermediate manipulation of the mixing coefficients in the source file `/chem/dry_dep_driver.F`, in lines 689 – 708, shown in Code Excerpt 3.1. Lines 678 – 683 explain, that this code was intended to boost vertical mixing, especially over urban areas. The local carbon monoxide emission strength E_{CO} is used as the criterion by which urban and rural areas are distinguished. A model grid cell is considered "on water" if $E_{CO} = 0$, "rural" if $0 < E_{CO} \leq 200$ mol km^{-2} h^{-1}, and "urban" if 200 mol km^{-2} h^{-1} $< E_{CO}$. From hereon, the classification of rural and urban model cells is referred to as the *rural-urban distinction* (RU-distinction). The warning in lines 686 – 687 does not apply to our model configuration, but might affect readers aiming to reproduce the results with a different chemical mechanism. The critical manipulation of the mixing coefficients follows in lines 691 – 696. In rural regions, the mixing coefficients of the lowest 10 model layers are set to $\max(k_h, 1)$ ("clipped to 1"). In urban regions, the mixing coefficients of the lower model half (meaning the lowest 50 % of all layers) are set to $\max(k_h, 2)$. In lines 696 – 706 further branches of the RU-distinction follow, which apply a clipping to $k_h = 2$ if

```
678  !!$! UNCOMMENT THIS AND FINE TUNE LEVELS TO YOUR DOMAIN IF YOU WANT TO
679  !!$! FORCE MIXING ESPECIALLY OVER URBAN AREAS TO A CERTAIN DEPTH:
680  !!$!
681  !!$! --- Mix the emissions up several layers in urban areas if no urban surface
         physics
682  !!$!      if e_co > 0., the grid cell should not be over water
683  !!$!      if e_co > 200., the grid cell should be over a large urban region
684  !!$!
685
686  ! this code is wrong - doesn't work if e_co is == param_first_scalar
687  ! (like it happened to be the case for MOZCART)
688  !      if (p_e_co > param_first_scalar) then
689  if (p_e_co >= param_first_scalar) then
690    if (sf_urban_physics .eq. 0) then
691      if (emis_ant(i,kts,j,p_e_co) .gt. 0) then
692        ekmfull(kts:kts+10) = max(ekmfull(kts:kts+10),1.)
693      endif
694      if (emis_ant(i,kts,j,p_e_co) .gt. 200) then
695        ekmfull(kts:kte/2) = max(ekmfull(kts:kte/2),2.)
696      endif
697      if (p_e_pm25i > param_first_scalar) then
698        if (emis_ant(i,kts,j,p_e_pm25i) + emis_ant(i,kts,j,p_e_pm25j) .gt. 8.19e
               -4*200) then
699          ekmfull(kts:kte/2) = max(ekmfull(kts:kte/2),2.)
700        endif
701      endif
702      if (p_e_pm_25 > param_first_scalar) then
703        if (emis_ant(i,kts,j,p_e_pm_25) .gt. 8.19e-4*200) then
704          ekmfull(kts:kte/2) = max(ekmfull(kts:kte/2),2.)
705        endif
706      endif
707    endif
708  endif
```

Code Excerpt 3.1: Manipulation of mixing coefficients in the WRF-Chem source code. Taken from `/chem/dry_dep_driver.F`. The code was slightly re-formatted for easier readability. The code comments are part of the original code.

certain thresholds of particulate matter emission are exceeded. The unitless clipping values of 1 and 2 should be understood in units of $m^2\ s^{-1}$.

The manipulation of such a crucial physical quantity without informing the user about it more directly is questionable in itself. However, this specific implementation is even more problematic for multiple reasons:

- The clipping values of 1 $m^2\ s^{-1}$ and 2 $m^2\ s^{-1}$ appear to be chosen empirically, and might need to be revised depending on the model setup. This is indicated by the code comments in lines 678 – 683, prompting the user to fine tune this parametrization to their domain.

```
690  if (sf_urban_physics .eq. 0) then
691    if (emis_ant(i,kts,j,p_e_co) .gt. 0) then
692      if (emis_ant(i,kts,j,urban_frac) .gt. config_flags%urban_threshold) then
693        ekmfull(kts+1:kte/2) = max(ekmfull(kts:kte/2),config_flags%k_min_urban)

694      else
695        ekmfull(kts:kts+10) = max(ekmfull(kts:kts+10),config_flags%k_min_rural)
696      endif
697    endif
698  endif
```

Code Excerpt 3.2: Revised mixing routine for WRF-Chem. See Code Excerpt 3.1 for context.

- Although apparently intended to be tunable, neither the clipping values nor any other parameters of this code are accessible through WRF-Chem's configuration interface (the `namelist.input` file). In consequence, any changes to the vertical mixing routine require a full re-compilation of the model.
- The RU-distinction based on emission criteria is inaccurate and unneccesary. WRF-Chem's geographical input data already contain a land use variable (called `LANDUSEF` in the `geo_em.d0*.nc` files produced by the WRF preprocessing system), which accurately separates the model domain into urban and rural regions.
- Because the emission thresholds of the RU-distinction refer to the *surface* emissions, the routine breaks when used with emission profiles. This is demonstrated and discussed further below in sect. 3.6.1.

The corresponding source code was committed to WRF-Chem's github repository on 8 May 2006 (commit `3c786f3`), meaning it has affected the community's simulation results for the past 18 years. It must be assumed that the WRF-Chem modelling community is mostly unaware of this, seeing that it is nowhere documented.

Code Excerpt 3.2 shows a revised mixing routine: Instead of the CO emissions, it accesses a variable named `urban_frac`, which describes the urban fraction of each model grid cell. `urban_frac` is obtained from the `LANDUSEF` variable mentioned earlier. In order to access `urban_frac` from within the mixing routine, it must be copied to the anthropogenic emission files first. The RU-distinction can then be based on `urban_frac`, returning "urban", where `urban_frac` is larger than a user-chosen threshold parameter `urb_threshold`, and "rural" elsewhere. The clipping of the mixing coefficients is no longer hard-coded to values of 1 and 2, but uses free parameters `k_min_rural` and `k_min_urban` instead. All parameters are made accessible through WRF-Chem's `namelist.input`. This requires a few additional code changes to the file `/Registry/registry.chem`, as shown in Code Excerpt 3.3. Afterwards, a full re-compilation of WRF-Chem is required.

```
166  state real urban_frac i+jf emis_ant 1 Z i5r "u_frac"   "u_frac" ""
     ...
3763 rconfig real k_min_urban   namelist,chem 1 2.0 - "k_min_urban"   "" ""
3764 rconfig real k_min_rural   namelist,chem 1 1.0 - "k_min_rural"   "" ""
3765 rconfig real urban_threshold namelist,chem 1 0.5 - "urban_threshold" "" ""
     ...
4018 package mozcem emiss_opt==8 - emis_ant:e_co,e_no,e_no2,e_bigalk,e_bigene,e_c2h4
        ,e_c2h5oh,e_c2h6,e_c3h6,e_c3h8,e_ch2o,e_ch3cho,e_ch3coch3,e_ch3oh,e_mek,
        e_so2,e_toluene,e_nh3,e_isop,e_c10h16,e_pm_10,e_pm_25,e_bc,e_oc,e_sulf,
        u_frac
```

Code Excerpt 3.3: Additions to the WRF-Chem registry for the revised mixing routine. The code refers to the WRF-Chem source file `/Registry/registry.chem`.

```
185  &chem
     ...
211    k_min_rural   = 1.0,
212    k_min_urban   = 2.0,
213    urban_threshold = 0.35,
```

Code Excerpt 3.4: Exemplary namelist entries for the revised mixing routine. The code refers to the WRF-Chem configuration file `namelist.input`.

The main advantages of this revised routine are:

- The RU-distinction is based on actual land-use data, which is more accurate and robust when used with temporal and vertical emission profiles.
- The clipping thresholds can be easily changed via the `namelist.input` file. An example for this is shown in Code Excerpt 3.4.

This allows for the systematic investigation of the influence of vertical mixing on simulated NO_2 and NO_x concentrations, as conducted in the following section. For easier readability, `k_min_urban`, `k_min_rural`, `urban_frac`, and `urban_threshold` are referred to as $k_{\text{min,urban}}$, $k_{\text{min,rural}}$, f_{urban}, and T_{urban} from hereon.

3.5 Analysis of the vertical mixing coefficients

The influence of vertical mixing on simulated NO_2 and NO_x is investigated on the basis of three further model runs (S-MYJ, S-BL, S-YSU-5-5), see Table 3.4. First, the trace gas cycles' dependence on the PBL scheme is assessed using WRF-Chem's original vertical mixing routine (S-YSU, S-MYJ, S-BL). Then, a run with the revised mixing routine using $k_{\text{min,urban}} = k_{\text{min,rural}} = 5$ is presented, based on which the benefits of the revised mixing routine are exemplified (S-YSU-5-5).

name	PBL scheme	mixing	$k_{\mathrm{min,urban}}$	$k_{\mathrm{min,rural}}$	vert. profiles	temp. profiles
S-YSU	YSU	original	2	1	no	yes[(1)]
S-MYJ	MYJ[(2)]	original	2	1	no	yes
S-BL	BL[(3)]	original	2	1	no	yes
S-YSU-5-5	YSU	revised	5	5	no	yes
S-YSU-5-5-B	YSU	revised	5	5	yes[(4)]	yes
S-YSU-0-5-B	YSU	revised	0	5	yes	yes
S-YSU-1-5-B	YSU	revised	1	5	yes	yes
S-YSU-2-5-B	YSU	revised	2	5	yes	yes

Table 3.4: WRF-Chem simulation configurations for the analysis of vertical mixing coefficients. The clipping thresholds $k_{\mathrm{min,\,urban}}$ and $k_{\mathrm{min,\,rural}}$ are given in units of $\mathrm{m^2\ s^{-1}}$.

(1) implemented according to Kumar et al. (2021).

(2) Mellor-Yamada-Janjić scheme, see Janjić (1994) and Mesinger (2020).

(3) Bougeault-Lacarrère scheme, see Bougeault and Lacarrère (1989).

(4) implemented according to Bieser et al. (2011).

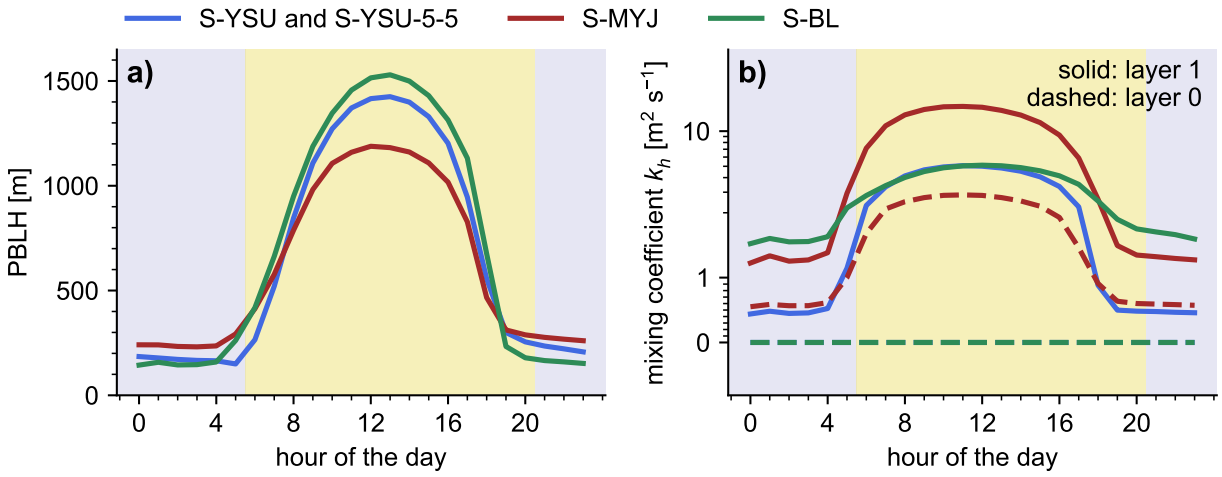

Figure 3.7: Analysis of boundary layer heights and mixing coefficients with different boundary layer schemes. a) Average diurnal cycle of the PBLH. **b)** Average diurnal cycle of the mixing coefficients in the lowest two layers of the model. The dashed lines represent the lowest model layer. The solid lines represent the layer above. The dashed lines for S-YSU and S-BL are close to (but not exactly) zero and overlap. The data shown here were sampled at the locations of the German background in situ measurements.

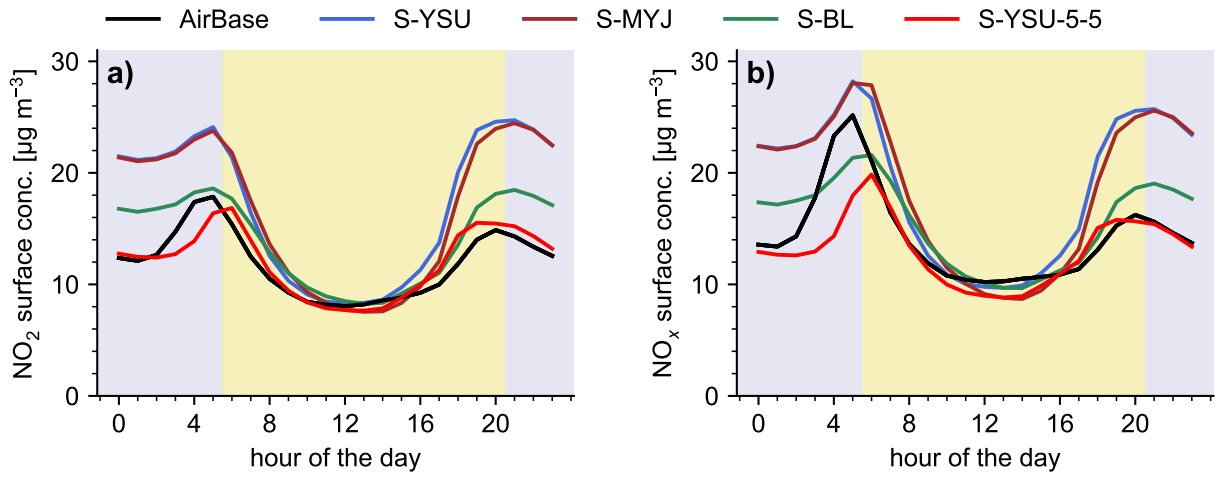

Figure 3.8: Comparison of diurnal cycles of surface NO_2 and NO_x with different boundary layer schemes and the revised mixing routine. **a)** Average diurnal cycle of NO_2. **b)** Average diurnal cycle of NO_x. The Mo-CL bias correction was applied. The black solid line shows the corresponding AirBase measurements. All data shown refer to background measurements/simulations in Germany.

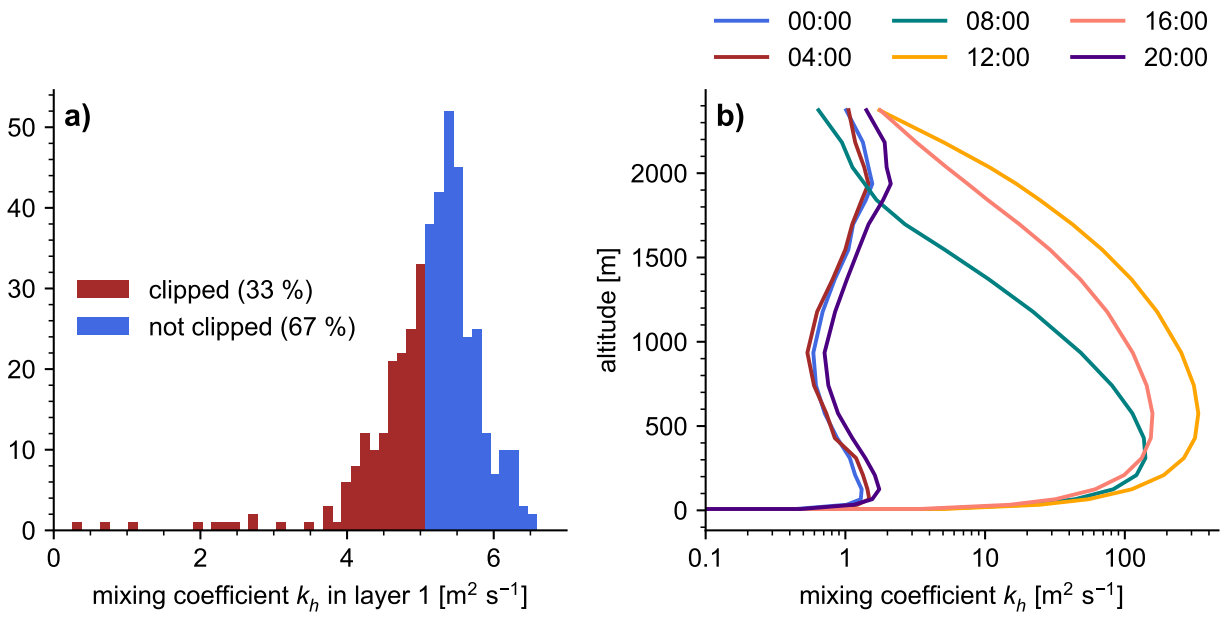

Figure 3.9: Further analysis of mixing coefficients in the simulation run S-YSU-5-5. **a)** Histogram of noontime mixing coefficients in layer 1 (the second layer above the ground). **b)** Vertical profiles of (unclipped) mixing coefficients at different hours of the day. All data shown here was sampled at the locations of the German background in situ measurements.

Figure 3.7a shows the diurnal cycle of the PBLH of the model runs S-YSU(-5-5), S-MYJ, and S-BL, at the background in situ measurement locations. The PBLH values of S-YSU and S-YSU-5-5 are identical throughout. The highest daytime boundary layers occur in S-BL, while S-MYJ yields the lowest daytime values. At nighttime, this relationship is inverted. Figure 3.8 shows the corresponding diurnal cycles of surface NO_2 and NO_x. One might expect that S-BL, where the trace gases are contained within a more shallow nighttime boundary layer and thus confined into a smaller total volume, would produce the largest nighttime surface concentrations. However, the opposite is the case: S-BL yields lower nighttime concentrations than S-YSU and S-MYJ. This is because the mixing of trace gases is not governed directly by the model's estimate of the PBLH, but by the mixing coefficients instead. Their values in the lowest two layers of the simulation are displayed in Fig. 3.7b. The critical observation to be made here is that all PBL schemes produce average nighttime mixing coefficients *smaller* than 1 m^2 s^{-1} in the lowest layer. Hence, the aforementioned clipping of the mixing coefficients is in effect, and the average nighttime mixing coefficient in the lowest layer is equalized among all schemes (with an effective value of 1 m^2 s^{-1}). One layer above, however, the mixing coefficients of S-BL are strictly larger than the rural clipping threshold of 1 m^2 s^{-1}, while S-MYJ and S-YSU still produce mixing coefficients of approximately 1 m^2 s^{-1} or below. Correspondingly, the trace gasses emitted into the surface layer are more strongly convected in S-BL, which explains the difference in the nighttime NO_2 and NO_x concentrations. This explanation is equally applicable to the daytime observations, where for example, S-MYJ shows the lowest surface NO_2 concentrations and the highest mixing coefficients out of the three runs.

The preceding investigation demonstrates, how the modelled trace gas concentrations are influenced by the strength of vertical mixing. The model's mixing coefficients are generally in a reciprocal relationship to the trace gas concentrations near the point of emission: stronger mixing results in lower concentrations, and vice versa. In the used model setup, the majority of NO_x emissions occurs at the surface, even if vertical emission profiles are used. Therefore, the statement can be further simplified: The vertical mixing coefficients (and by extension, their clipping values) are in an inverse relation to the *surface* NO_2 and NO_x concentrations. This helps to explain the results obtained from the simulation run S-YSU-5-5, as shown in Fig. 3.8. S-YSU-5-5 is identical to S-YSU, except that it uses the revised mixing routine with larger clipping values of $k_{\mathrm{min,rural}} = k_{\mathrm{min,urban}} = 5$ m^2 s^{-1}. This parametrization was chosen based on the literature recommendation of Du et al. (2020), and will be further optimized later on. Corresponding to the higher clipping thresholds, nighttime mixing is enhanced and the resulting NO_2 and NO_x concentrations are reduced. At daytime, the concentration differences are much smaller, because the mixing coefficients computed by all three PBL schemes are significantly larger, and thus less affected by the clipping procedure (see Fig. 3.7b). The caveat to this analysis is that it is based on *averaged* values of the mixing coefficient. Due

to spatial and temporal fluctuations, individual model cells may be subject to the clipping procedure, even if their average value exceeds the clipping threshold. Figure 3.9a shows a histogram of the noontime S-YSU mixing coefficients of layer 1 (meaning: the second layer above the ground, as in Fig. 3.7b). Here, although the average mixing coefficient is just above the clipping threshold of $k_h = 5$ m^2 s^{-1}, approximately 33 % of the model grid cells fall below it nonetheless. Correspondingly, the clipping routine may slightly enhance mixing, which an analysis based on average mixing coefficients alone could not indicate. However, due to the similarity of the diurnal NO_2 and NO_x cycles displayed in Fig. 3.8, this effect is likely negligible in most scenarios.

Another relevant question is, over which vertical range the influence of the clipping extends. According to Code Excerpt 3.1, the clipping can affect the lowest 10 layers in rural regions, or the lowest 50 % of the layers in urban regions. Figure 3.9b shows vertical profiles of the mixing coefficients from S-YSU at different times of the day. At nighttime, when the average mixing coefficients are between 1 m^2 s^{-1} and 2 m^2 s^{-1}, the clipping may extend far up into the troposphere, and even beyond the nocturnal boundary layer. During the day, however, the clipping essentially affects only the lowermost layers of the model (approx. 0 – 33 m), because the mixing coefficients increase rapidly with altitude. For example, the third-lowest model layer has an average noontime mixing coefficient of $k_h \approx 24$ m^2 s^{-1}, which far exceeds the clipping threshold.

3.6 Simulations with vertical emission profiles and optimized vertical mixing

No vertical emission profiles were deployed up to this point, meaning that all emissions were injected into the lowest layer of the model. The reason for holding back vertical emission profiles was to investigate the influence of the mixing coefficients in a simpler simulation setup. In order to achieve the most realistic simulation, they are now applied as described in the beginning of the chapter. The majority of NO_x is emitted from traffic, which remains in the lowest model layer. Still, the implementation of vertically distributed emissions leads to a decrease in surface emissions by approximately 33 % (see Fig. 3.10), and correspondingly lower surface concentrations. Therefore, the mixing thresholds of $k_{\mathrm{min,rural}} = k_{\mathrm{min,rural}} = 5$ m^2 s^{-1}, which previously produced reasonable results, must be recalibrated. This optimization is conducted on the basis of four simulation runs (S-YSU-5-5-B, S-YSU-0-5-B, S-YSU-1-5-B, and S-YSU-2-5-B, see Table 3.4). All runs use $T_{\mathrm{urban}} = 0.35$, meaning that the mixing routine's RU-distinction treats a model grid cell as urban, if more than 35 % of its area are covered by urban environments.

3.6.1 Influence of vertical emission profiles on the RU-distinction

The introduction of vertical emission profiles allows for a more detailed comparison of the original and revised RU-distinction of the vertical mixing routine. Figure 3.11a shows the model's revised RU-distinction by means of the urban fraction f_{urban}. Figure 3.11b-d shows the original RU-distinction based on the surface CO emissions. The comparison of Fig. 3.11a and 3.11b reveals, that both methods are in good qualitative agreement at noontime. However, this is only true for the German domain, where the emission data have particularly high resolution. Outside of Germany, the original RU-distinction fails at registering urban environments reliably. For example, not a single model grid cell of the Netherlands is identified as urban in Fig. 3.11b. This is most likely due to the dilution of low-emission rural environments with high-emission urban environments as a consequence of the much lower resolved emission data. Figure 3.11c shows, that the distinction capability of the original RU-distinction is further reduced when vertical emission profiles are used, because they reduce the surface CO emissions. Then, even metropolitan cities, like Brussels or Linz are no longer identified as urban. Lastly, as shown in Fig. 3.11d, the original RU-distinction is rendered completely useless during nighttime (shown here: at midnight), when the surface CO emissions are drastically lowered by the temporal emission profiles. As a result, the entire domain is classified as rural.

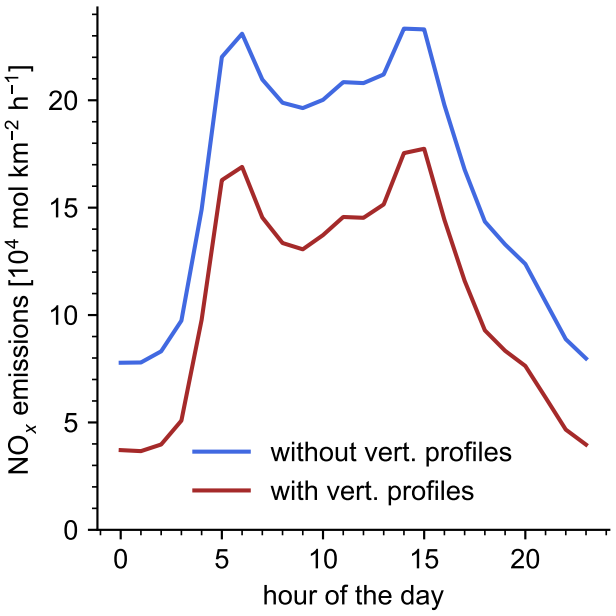

Figure 3.10: Diurnal cycle of average surface NO_x emissions with and without vertical emission profiles. Here: Using the vertical emission profiles from Bieser et al. (2011).

An interesting point of investigation is how well the RU-distinction of the model agrees with the classification of the AirBase instruments. For that purpose, instruments are labelled "all-urban" if they are classified as "urban" or "suburban" in the AirBase dataset and "all-rural" otherwise. Figure 3.11e-h shows the locations of the German background in situ instruments. All-urban and all-rural stations, whose colocated model grid cell is correctly identified as urban or rural are displayed as black and gray dots. If the colocated model grid cell's classification is false, the stations are displayed as red dots and green triangles, respectively. The revised RU-distinction shown in Fig. 3.11e agrees with the AirBase classification in

67 % of the all-urban and 93 % of the all-rural cases. The corresponding ratios for Fig. 3.11e-h are: 51 % and 96 %, 37 % and 99 %, 0 % and 100 %. The slightly better accuracy of the original RU-distinction with respect to all-rural stations is merely a consequence of it failing to register urban environments properly. As an extreme example take Fig. 3.11h, where the entire domain is classified as rural, and in consequence there can be no all-rural AirBase locations misclassified as urban. Lastly, the choice of $T_{\mathrm{urban}} = 0.35$ is briefly discussed. It might seem more reasonable to use a threshold of $T_{\mathrm{urban}} = 0.50$ instead, i.e. to classify a model grid cell as urban, if more than 50 % of its area are covered by urban land-use elements. However, this arguably worsens the classification accuracy to 58 % for all-urban and 96 % for all-rural instruments. Besides, the RU-distinction is generally not very sensitive to the choice of T_{urban}, because most model grid cells are either dominantly rural, or dominantly urban, but rarely an urban-rural mixture. More intuitively, most of the domain's area shown in Fig. 3.11a is either clearly blue (rural) or clearly red (urban), with very little white area in between. Overall, these results demonstrate the clear benefits of the revised RU-distinction based on land-use data, which is unambiguous and independent of the resolution of the emission data, as well as vertical and temporal emission profiles.

3.6.2 Evaluation of diurnal NO_2 and NO_x cycles

Vertical displacement of emissions results in lower NO_2 and NO_x concentrations at the surface. Hence, the model's mixing parametrization requires recalibration. This is exemplified in Fig. 3.12, showing the diurnal cycles of NO_2 and NO_x for S-YSU-5-5 and the four simulation runs with vertical emission profiles mentioned above. The simulation run S-YSU-5-5-B is identical to S-YSU-5-5, except that it uses vertically distributed emissions. This results in a reduction of the NO_2 concentrations by approx. 11 % at noontime, and 17 % over the full diurnal cycle. Correspondingly, the model's vertical mixing must be dampened via its clipping values in order to compensate the observed underestimations. The previously established classification into all-rural and all-urban instrument locations can be used to optimize $k_{\mathrm{min,rural}}$ and $k_{\mathrm{min,urban}}$ separately: $k_{\mathrm{min,urban}}$ by the agreement to the all-urban in situ measurements, and $k_{\mathrm{min,rural}}$ by the agreement to the all-rural measurements. Figures 3.12c and 3.12d reveal that the underestimations of NO_2 and NO_x occur mostly in the urban environments. When $k_{\mathrm{min,urban}}$ is reduced from 5 $\mathrm{m^2\ s^{-1}}$ (dark red line, S-YSU-5-5-B) to 1 $\mathrm{m^2\ s^{-1}}$ (bright red line, S-YSU-1-5-B), the modelled surface concentrations increase as expected, but the adjustment is too strong and results in an overestimation of NO_2 and NO_x. A value of $k_{\mathrm{min,urban}} = 2\ \mathrm{m^2\ s^{-1}}$ (orange line, S-YSU-2-5-B) results in intermediate NO_2 and NO_x concentrations and thus good noontime agreement to the all-urban in situ measurements.

It can be tested how the model responds to a clipping value of $k_{\mathrm{min,urban}} = 0$, meaning that the unmodified mixing coefficients of the YSU PBL scheme are used. As demonstrated by

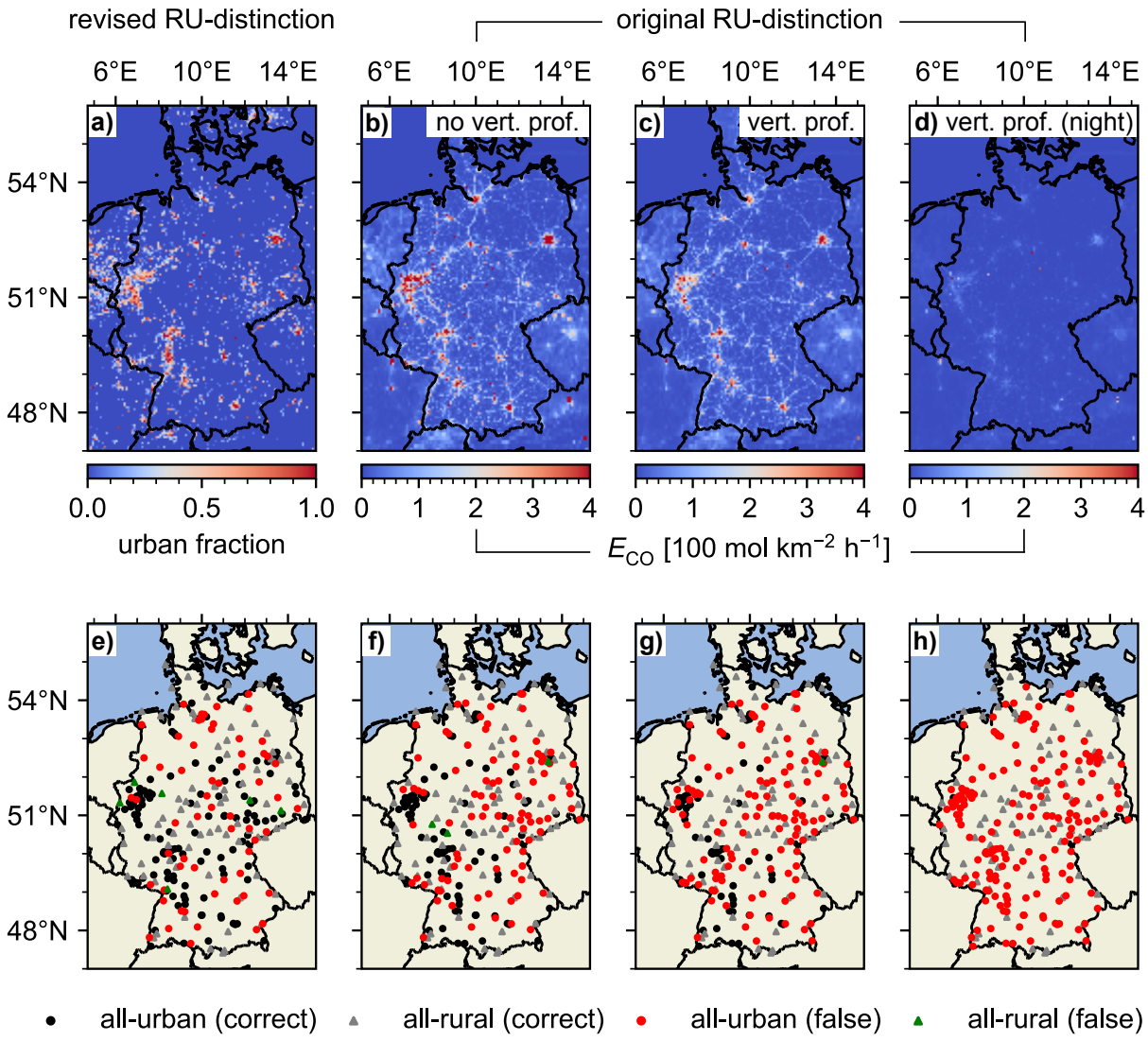

Figure 3.11: Comparison of the RU-distinction in the original and revised mixing scheme. **a-d)** Distinction criteria of the RU-distinction. From left to right: **a)** f_{urban} (revised mixing routine), **b)** noontime surface E_{CO} without vertical profiles, **c)** noontime surface E_{CO} with vertical profiles, and **d)** midnight surface E_{CO} with vertical profiles. E_{CO} denotes the surface CO emissions. **e-h)** Map of the urban and rural in situ measurement locations. Instruments whose AirBase classification into all-rural and all-urban stations agrees with the classification of the mixing routine are labelled "correct". Otherwise, they are labelled "false". For a definition of "all-rural" and "all-urban", refer to the main text.

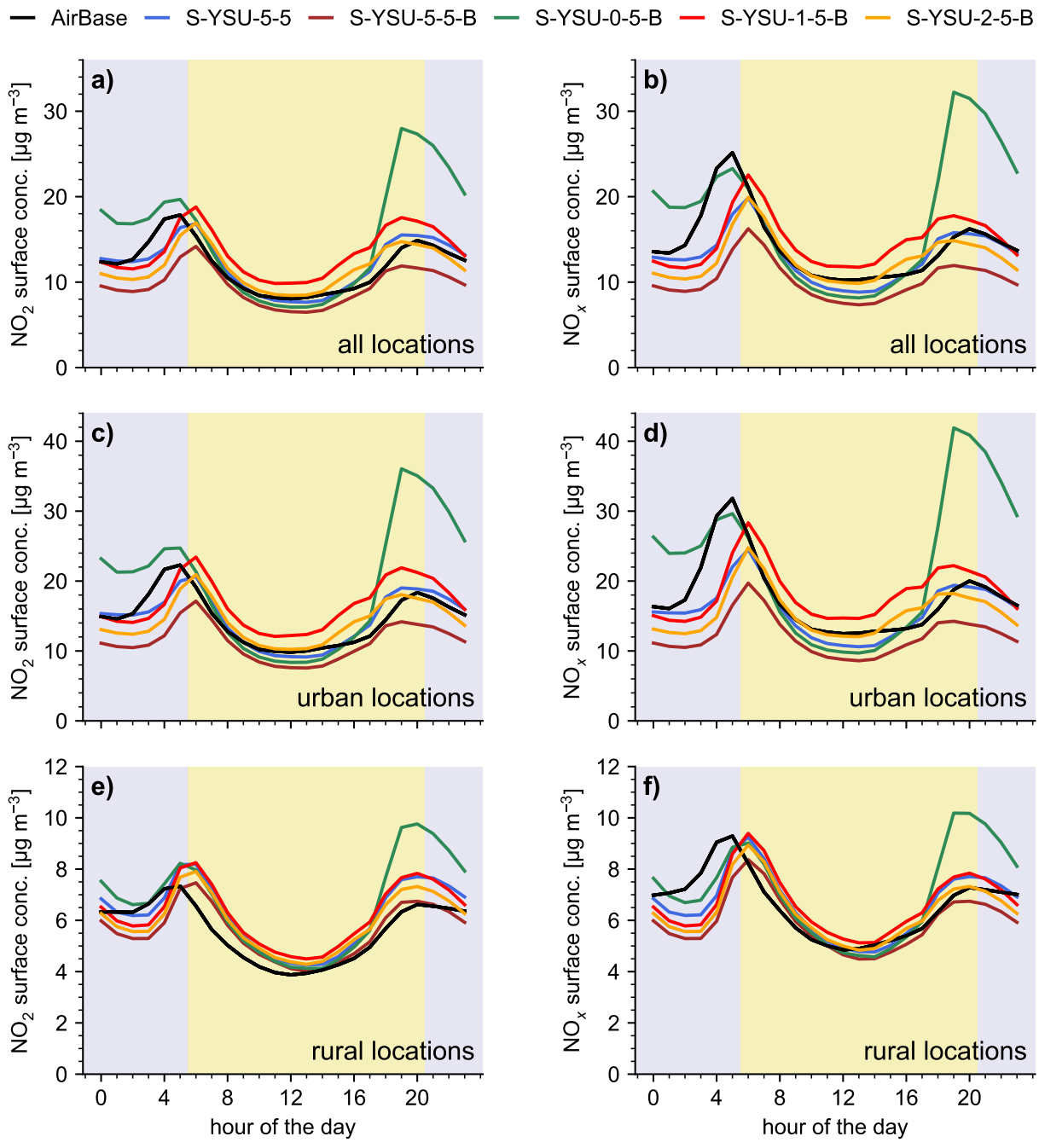

Figure 3.12: Comparison of diurnal NO_2 and NO_x cycles with vertical emission profiles and the revised mixing routine. a-b) In all locations. **c-d)** In all-urban locations. **e-f)** In all-rural locations. The Mo-CL bias correction was applied. S-YSU-5-5 is the only shown simulation run without vertical emission profiles. All data refer to background measurements/simulations in Germany.

the simulation run S-YSU-0-5-B (green line) the resulting diurnal cycles show an unexpected shape: At nighttime, strong overestimations corresponding to suppressed vertical mixing are observed as expected, but the noontime surface concentrations are almost as low as with $k_{\mathrm{min,urban}} = 5\ \mathrm{m^2\ s^{-1}}$. It is unclear, how these low noontime concentrations come about, but speculatively it could be an issue related to numerical (in)stability, seeing that Fig. 3.7b revealed un-clipped mixing coefficients very close to zero in the YSU PBL scheme.

Figures 3.12e and 3.12f show an evaluation of the different simulation runs in rural environments. The NO_2 cycle of S-YSU-2-5-B shows slight noontime overestimations of approximately 13 %. Note, that the simulation runs are quite similar to each other in rural regions, because they only differ in $k_{\mathrm{min,urban}}$, not in $k_{\mathrm{min,rural}}$. Better agreement in noontime NO_2 could obviously be achieved by increasing $k_{\mathrm{min,rural}}$. However, the simulation run S-YSU-2-5-B already shows good agreement in the daytime NO_x cycle. For the determination of a reasonable mixing configuration, this is ultimately a more reliable reference than the NO_2 cycle alone, because it validates $k_{\mathrm{min,urban}}$ and $k_{\mathrm{min,rural}}$ independently of other modelling aspects, such as NO_x chemistry and the resulting NO_2/NO_x ratios discussed in the following.

3.6.3 Analysis of NO_2/NO_x ratios

The daytime overestimations of rural surface NO_2 can not be resolved by further tuning of the vertical mixing, without breaking the good agreement to the NO_x in situ measurements. However, it can be demonstrated that the NO_2 overestimations are likely related to the model's (organic) NO_x chemistry. As shown in Fig. 3.13c-d the model tends to overestimate the NO_2/NO_x ratio by approximately 14 % in rural regions, and the O_3 surface concentration by approximately 20 µg $\mathrm{m^{-3}}$.

It can be simulated how the modelled NO_2, NO and NO_2/NO_x ratios would respond to certain chemical constraints. The results shown here are *not* obtained from actual new simulation runs, but from theoretical considerations applied to the existing simulation results shown in Fig. 3.12. For example, the modelled O_3 concentration can be fixed retroactively, so that it agrees with the in situ measurements. The corresponding change to the model's NO_2 and NO concentrations is estimated as follows: Under the assumption of steady state, the Leighton relationship from eq. (2.5) yields:

$$\frac{[\mathrm{O_3^*}]}{[\mathrm{O_3}]} = \frac{[\mathrm{NO_2^*}]}{[\mathrm{NO_2}]} \frac{[\mathrm{NO}]}{[\mathrm{NO^*}]} \tag{3.2}$$

where the star symbol * denotes the trace gas concentrations with adjusted ozone levels. Note, that eq. (3.2) refers to particle concentrations, not mass concentrations. Furthermore, the particle concentration of NO_x is conserved:

$$[\mathrm{NO}^*] + [\mathrm{NO}_2^*] = [\mathrm{NO}] + [\mathrm{NO}_2] \tag{3.3}$$

Solving for $[\mathrm{NO}^*]$ and $[\mathrm{NO}_2^*]$ yields

$$[\mathrm{NO}^*] = \frac{[\mathrm{NO}] + [\mathrm{NO}_2]}{1+a}, \quad [\mathrm{NO}_2^*] = a \cdot \frac{[\mathrm{NO}] + [\mathrm{NO}_2]}{1+a}, \quad \text{with} \quad a = \frac{[\mathrm{O}_3^*]}{[\mathrm{O}_3]} \frac{[\mathrm{NO}_2]}{[\mathrm{NO}]} \tag{3.4}$$

If $[\mathrm{O}_3^*]$ is set to the values of the in situ measurements, this results in a daytime surface NO_2 decrease of approximately 5 %, reducing the rural NO_2 overestimations to approximately 8 %. The NO concentration and the NO_2/NO_x ratio are slightly increased and decreased, correspondingly, see Fig. 3.13e-h. This demonstrates that the errors in the model's simulated NO_2, NO and NO_2/NO_x can not be explained by the overestimation of O_3 under steady state assumptions alone. This is not unexpected, because organic NO_x chemistry involves other reactions, which convert NO to NO_2 without loss of O_3, e.g. the oxidation of NO by HO_2 or RO_2 (see sect. 2.1.2).

The preceding analysis can be extended by using the NO_2/NO_x ratio measurements as an additional constraint. For that purpose, in addition to enforcing agreement to the observed O_3, modelled NO_2 is converted to NO without further loss of O_3, until agreement between the simulated NO_2/NO_x ratio and the corresponding in situ observations is reached. As evident by Fig. 3.13i-l, this is the key to explaining most of the remaining model discrepancies in daytime NO_2 and NO. Particularly the modelled NO concentrations in Fig. 3.13j show much better agreement to the in situ measurements under the imposed constraint. Likewise, the NO_2 concentrations are further reduced. This indicates, that the model's NO_x concentrations during daytime are realistic, but some unknown model component (e.g. the chemical mechanism, or biased VOC emissions) leads to a faulty partitioning into NO_2 and NO. At nighttime, the model errors can not be resolved entirely, which hints towards further issues with the model's nighttime NO_x chemistry, e.g. a tendency of the model to deposit too much NO_2 into the reservoir species NO_3 and N_2O_5, which escape the imposed constraints. Unfortunately, the nighttime reservoir species are not measured by the in situ instruments, and can therefore not be validated.

In conclusion, the mixing configuration of the simulation run S-YSU-2-5-B with $k_{\mathrm{min,rural}} = 2\ \mathrm{m^2\ s^{-1}}$ and $k_{\mathrm{min,urban}} = 5\ \mathrm{m^2\ s^{-1}}$ appears to be realistic, even if small deviations in the diurnal cycles of rural NO_2 and NO_x remain. At least during daytime, these can be explained by shortcomings in the model's NO_x chemistry, which is not investigated further within this thesis.

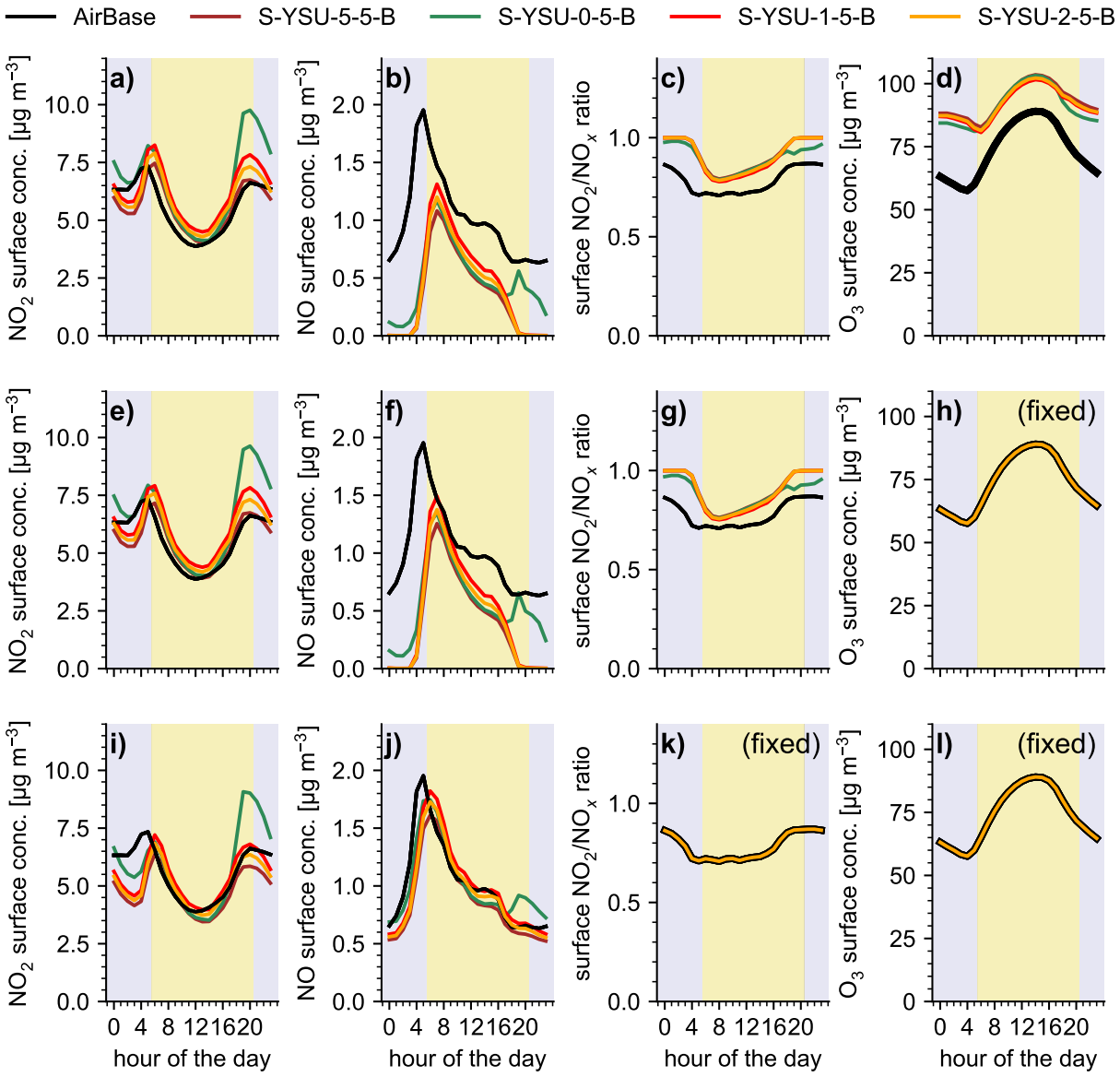

Figure 3.13: Diurnal cycles of rural NO_2, NO, NO_2/NO_x, and O_3 in a hypothetical model run with enforced chemical constraints. a-d) Original simulation results, corresponding to Fig. 3.12. **e-f)** Expected results with forced agreement to the in situ O_3 observations, and corresponding reduction of the NO_2 concentration and NO_2/NO_x ratio based on the steady state assumption. **i-l)** Expected results with additionally forced agreement to the in situ NO_2/NO_x observations. All data refer to all-rural background measurements/simulations in Germany.

3.7 Final evaluation and quantitative summary of all simulation runs

The preceding sections have highlighted the strong impact of the diurnal and vertical emission profiles, the parametrization of vertical mixing in the lowest model layers, and the RU-distinction. These model components were investigated and partly optimized based on the comparison of simulated diurnal NO_x and NO_2 cycles to background in situ measurements in Germany. Now, a final intercomparison of the different simulation runs listed in Table 3.2 and 3.4 against background AirBase in situ NO_2 measurements, tropospheric NO_2 VCDs from TROPOMI, and NO_2 profiles from MAX-DOAS measurements on the full simulation domain is presented. Due to the complexity of this task, model runs which represent intermediate optimization steps (e.g. S-YSU-1-5-B) are not shown. All figures show monthly means. Quantitative summaries based on monthly-mean/unaveraged data are given in tabular form.

3.7.1 Comparison to AirBase in situ measurements

Figure 3.14 gives an overview of the modelled noontime surface NO_2 concentrations and the corresponding AirBase in situ measurements. As evident from the subplots representing the differences (shown in the 3rd column) and comparison e.g. to Fig. 3.4, the WRF-Chem simulations show more underestimations on the European domain than they did in Germany. A possible reason for this could be the difference in the horizontal resolution of the emission data ($0.01° \times 0.01°$ in Germany, $0.1° \times 0.1°$ elsewhere), due to which emission hotspots outside of Germany could be diluted, and therefore misrepresented. On the other hand, the seemingly well-balanced simulation results obtained on the German domain must be at least partially attributed to the strong overestimations in western Germany, which are sufficient to compensate the underestimations in other parts of Germany, but not on the entire simulation domain. Nonetheless, the linear regressions through the scatter point clouds (see the 4th column) show promising results, with slopes close to 1 and offsets close to 0 for some of the simulation runs (S-YSU, S-BL, and S-YSU-2-5-B). However, this metric should be used with care, because the strong dispersion of the scatter points close to the origin makes the linear regression highly unstable. For example, Fig. 3.14a and 3.14b show very similar results for S-YSU and S-MYJ, but vastly different slopes ($m = 0.69$ and $m = 1.00$) nonetheless. Other than that, the differences between the individual simulation runs can be explained. As expected from the previous comparisons, S-BL and S-YSU give very similar noontime results. S-YSU-TP produces generally larger noontime surface concentrations due to the enhanced noontime emissions. S-YSU-5-5 is similar to S-YSU, but produces lower noontime surface concentrations due to enhanced vertical mixing. S-YSU-5-5-B is similar to S-YSU-5-5, but

shows even lower surface concentrations, because the use of vertical emission profiles reduces the emission strength at the surface. This effect is mostly compensated in S-YSU-2-5-B by reduction of the vertical mixing strength in urban regions.

A quantitative summary for the comparison to the AirBase in situ measurements is found in Tables 3.5 and 3.6. The statistical diagnostics given here are the mean relative bias as defined in eq. (3.1), as well as the RMSE, and the Pearson correlation coefficient, defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{3.5}$$

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3.6}$$

where x_i denote the reference datapoints, and y_i the corresponding modelled datapoints from the WRF-Chem simulations. The three diagnostics are computed for four timeframes: noontime (12 PM), daytime (averaged from 5 AM to 7 PM), nighttime (averaged from 8 PM to 5 AM), and for the entire 24 hours of the day ("diurnal"). The corresponding diurnal surface NO_2 and NO_x cycles are displayed in Fig. 3.15. The following text descriptions refer to Table 3.5 (monthly averages).

The differences between S-YSU, S-MYJ, and S-BL described in the previous analysis of WRF-Chem's mixing routine can be clearly verified here. Their results are similar at noontime, but vastly different at nighttime. This owes to the stronger vertical mixing in the Bougeault-Lacarrère scheme (see Fig. 3.7), resulting in lower biases and RMSEs. The lowest overall biases are obtained from the simulation run S-YSU-TP with tuned emission profiles (noontime bias of +2.9 %). This should be deemed a coincidence, as it appears to be caused mainly by the strong overestimations observed in Fig. 3.6 canceling out with the underestimations introduced when extending the evaluation from Germany to the full simulation domain. Furthermore, depending on the time window, S-YSU-TP often shows larger RMSE values than the other runs. This further indicates, that the lower simulation bias is not achieved by a higher model accuracy. Instead, compared to the other simulation runs, S-YSU-TP has reduced low biases, but increased high biases. This explanation also aligns well with the data shown in Fig. 3.14.

The simulation runs with the revised vertical mixing routine (S-YSU-5-5, S-YSU-5-5-B, and YSU-2-5-B) still show low biases in all four time windows, but their RMSE values are mostly lower than those of the other simulation runs. When averaged over the entire diurnal

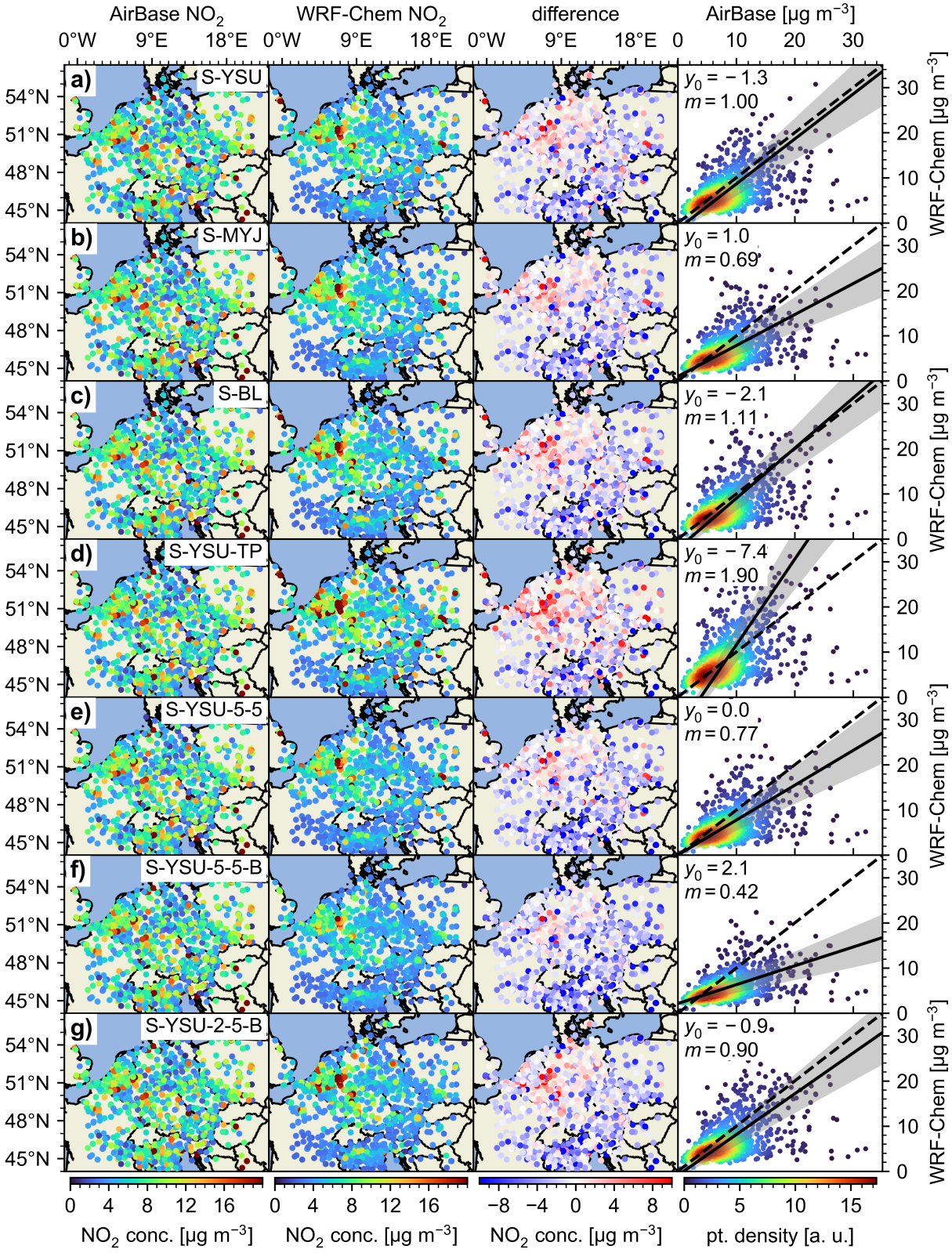

Figure 3.14: Comparison of monthly-mean noontime surface NO_2 concentrations from different simulation runs to AirBase in situ measurements. Shown here: background measurements on the full simulation domain. Intercepts (y_0) are given in units of µg m^{-3}.

		YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
noon	bias [%]	−15.9	−19.9	−13.4	+2.9	−22.4	−33.2	−20.3
	RMSE	5.3	5.0	5.3	6.2	5.2	5.2	5.1
	R	0.44	0.45	0.47	0.45	0.45	0.49	0.49
day	bias [%]	+6.4	+3.6	−7.5	−0.3	−16.4	−30.3	−17.9
	RMSE	7.6	6.9	6.6	7.4	6.5	6.2	6.3
	R	0.54	0.55	0.55	0.52	0.53	0.57	0.57
night	bias [%]	+52.2	+50.0	+18.4	+13.3	−10.0	−32.4	−23.7
	RMSE	15.4	14.9	10.4	10.7	7.8	7.0	6.9
	R	0.65	0.66	0.65	0.61	0.62	0.65	0.66
diurnal	bias [%]	+25.3	+22.8	+3.2	+5.3	−13.8	−31.2	−20.3
	RMSE	10.0	9.4	7.6	8.3	6.8	6.2	6.3
	R	0.60	0.61	0.61	0.58	0.58	0.62	0.63

Table 3.5: Validation of different simulation runs against NO_2 in situ measurements from AirBase based on monthly-mean data. All results refer to background measurements/simulations. RMSE values are given in units of µg m^{-3}. Daytime averages are computed from 5 AM to 7 PM. Nighttime averages are computed from 8 PM to 5 AM.

		YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
noon	bias [%]	−15.6	−19.7	−13.1	+3.2	−22.2	−33.0	−20.0
	RMSE	7.6	7.3	7.3	8.7	7.1	6.9	6.9
	R	0.38	0.40	0.41	0.39	0.41	0.43	0.44
day	bias [%]	+6.7	+3.9	−7.2	+0.0	−16.2	−30.2	−17.7
	RMSE	12.2	11.8	10.4	11.2	9.8	9.2	9.5
	R	0.47	0.49	0.49	0.46	0.47	0.50	0.49
night	bias [%]	+52.7	+50.5	+18.8	+13.8	−9.7	−32.3	−23.5
	RMSE	19.1	18.8	15.6	14.7	12.2	11.3	11.4
	R	0.46	0.48	0.50	0.45	0.43	0.45	0.46
diurnal	bias [%]	+25.8	+23.1	+3.5	+5.7	−13.5	−31.0	−20.1
	RMSE	15.2	14.8	12.6	12.6	10.8	10.1	10.2
	R	0.47	0.49	0.50	0.46	0.46	0.48	0.48

Table 3.6: Like Table 3.5, but based on hourly data. RMSE values are given in units of µg m^{-3}.

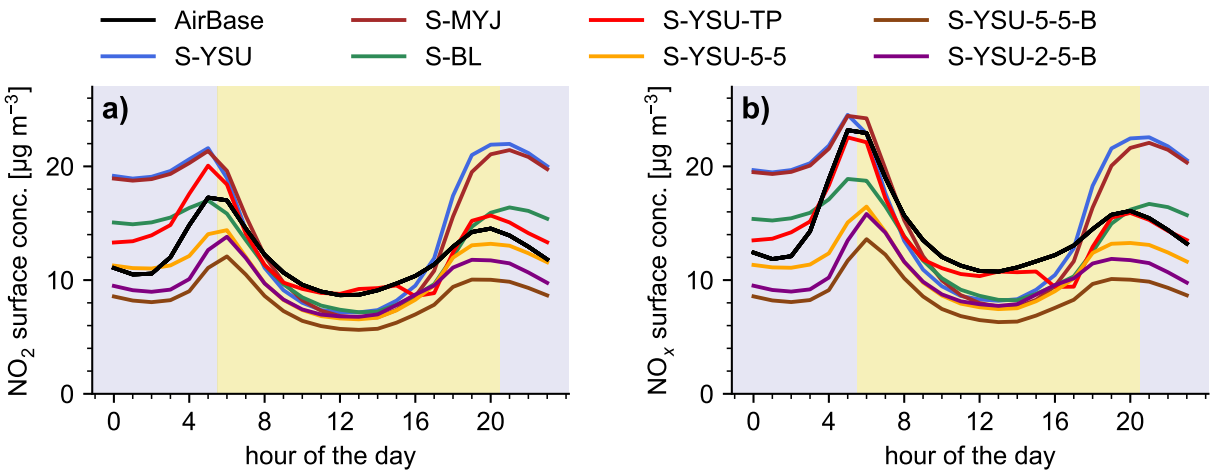

Figure 3.15: Comparison of diurnal surface NO_2 and NO_x cycles of different simulation runs to AirBase measurements. All data shown refer to background measurements/simulations in Europe.

cycle, this results in an RMSE decrease from approx. 8.8 µg m^{-3} to 6.4 µg m^{-3}. Note, that this is also the case for the run S-YSU-5-5 (without vertical emission profiles), meaning that the improvements must relate to the revised mixing routine instead. The results obtained from hourly data (see Table 3.6) are qualitatively very similar, except that the RMSE values are enhanced and the correlation coefficients reduced.

3.7.2 Comparison to TROPOMI satellite measurements

Figure 3.16 shows the comparison of modelled tropospheric NO_2 VCDs to TROPOMI satellite measurements in a similar manner as Fig. 3.14. Tables 3.7 and 3.8 list the corresponding statistical diagnostics. The air mass factors of the TROPOMI reference values were re-computed using the high-resolution NO_2 profiles from the WRF-Chem simulations. Like before, the qa-filter criterion is applied, i.e. data with $f_{QA} \leq 0.75$ are dismissed. The maps shown in Fig. 3.16 reveal that in some locations (e.g. Belgium and the Netherlands) there is an obvious spatial correlation between the over- and underestimations of the NO_2 VCDs and the surface concentrations, as described previously for the German domain. On the other hand, other regions show overestimations of the NO_2 VCDs (e.g. in Czechia and Paris) with no corresponding overestimations of the surface concentrations. When evaluated on the full simulation domain, the overestimations in western Germany are revealed to be exceptionally large, with similar levels only occurring in small parts of Belgium, and the city of Kaliningrad. However, these overestimations are (partly) statistically compensated by underestimations in other regions (e.g. northern Italy, France, Austria and Poland). As evident from Fig. 3.16e-g, the critical overestimations are significantly reduced in the simulation runs using the revised vertical

	YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
bias [%]	+2.5	−1.3	+2.6	+9.4	+1.7	+2.0	+1.6
RMSE	8.5	8.4	9.3	11.2	8.1	8.0	7.9
R	0.86	0.87	0.88	0.86	0.86	0.86	0.86

Table 3.7: Validation of different simulation runs against trop. NO_2 VCDs from TROPOMI based on monthly-mean data. RMSE values are given in units of 10^{14} molec. cm^{-2}. The air mass factors of the TROPOMI reference data were re-computed using the NO_2 profiles from the WRF-Chem simulations.

	YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
bias [%]	+2.5	−0.8	+2.6	+8.3	+1.7	+1.9	+1.6
RMSE	19.9	20.8	20.9	22.3	19.1	18.9	18.8
R	0.63	0.63	0.65	0.64	0.63	0.63	0.63

Table 3.8: Like Table 3.7, but based on data from individual orbits.

mixing routine (S-YSU-5-5, S-YSU-5-5-B, S-YSU-2-5-B). This indicates, that the enhanced mixing leads to more realistic profile shapes. Meanwhile, the simulation run S-YSU-TP with tuned emission profiles amplifies the described overestimations instead, resulting in poor overall agreement. In summary, all simulation runs produce monthly-mean NO_2 VCDs with mean biases of < 3 %, RMSE values of approx. $9.0 \cdot 10^{14}$ molec. cm^{-2}, and correlation coefficients of approx. $R = 0.86$. The sole exception is S-YSU-TP with a bias of +9.4 % and an RMSE of $11.2 \cdot 10^{14}$ molec. cm^{-2}. When evaluated against data from individual orbits, the RMSE increases to approx. $20.0 \cdot 10^{14}$ molec. cm^{-2} and the correlation coefficient drops to approx. $R = 0.63$. Figure B.8 shows a version of Fig. 3.16 evaluated for a single satellite orbit, and Fig. B.9 shows a version of Fig. 3.16 with the original air mass factors. The main observation to be made here is that re-computing the air mass factors based on WRF-Chem's high-resolution NO_2 profiles results in significantly larger TROPOMI NO_2 VCDs, with average enhancements between 18 % and 23 % depending on whether vertical emission profiles were used or not.

3.7.3 Comparison to NO_2 profiles from MAX-DOAS measurements

The last dataset to validate the WRF-Chem simulations against are the FRM$_4$DOAS NO_2 profiles obtained from MAX-DOAS measurements. For May 2019, FRM$_4$DOAS data is available from the stations Mainz, Bremen, Heidelberg, Uccle, and De Bilt (see Table 2.3). Retrieval version 3 (fv003) is used. The FRM$_4$DOAS NO_2 profiles extend from 0 – 4 km, with a layer height of 200 m. The WRF-Chem NO_2 profiles are interpolated to the FRM$_4$DOAS data

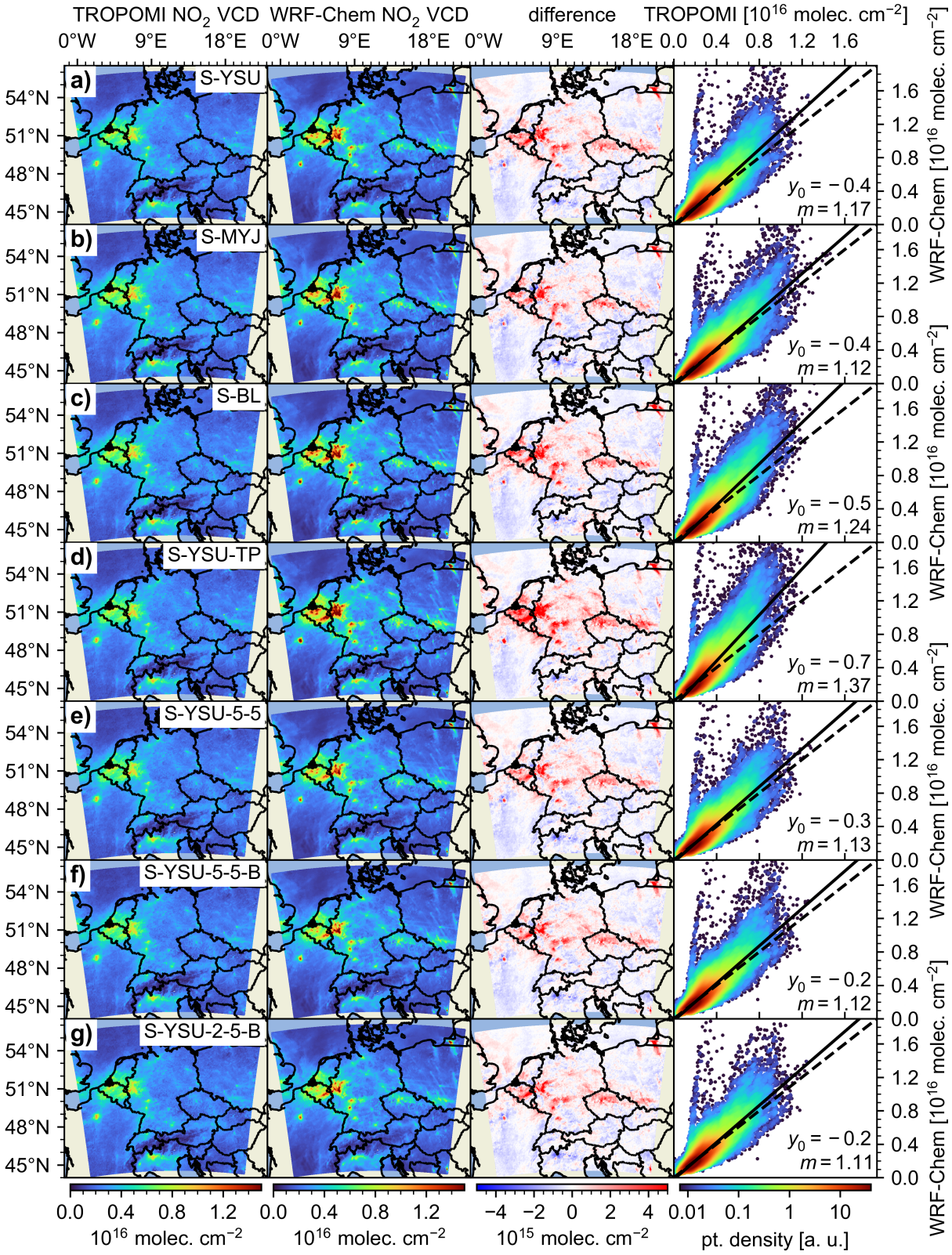

Figure 3.16: Comparison of monthly-mean tropospheric NO_2 VCDs from different simulation runs to TROPOMI satellite measurements. Intercepts (y_0) are given in units of 10^{15} molec. cm^{-2}. The air mass factors of the TROPOMI reference data were re-computed using the NO_2 profiles from the WRF-Chem simulations.

coordinates in horizontal and vertical direction. Both retrieval algorithms, MAPA and MMF, are used. The averaging kernels obtained from the MMF retrieval are applied according to eq. (2.69). The MAPA retrieval is used with an O_4 scaling factor of 0.8, and all entries flagged as "erroneous" are filtered out (for details see Beirle et al., 2019). In locations with multiple operational instruments (e.g. with different azimuthal viewing directions) the average of all instruments is computed.

Figure 3.17 shows a direct comparison of MMF, MAPA, and all WRF-Chem simulation runs separated by instrument location. The left column presents the monthly-mean NO_2 profiles obtained from a time window of 11 AM – 2 PM. The right column displays scatter plots, where each marker represents the average NO_2 concentration within a single retrieval layer (here: drawn against the NO_2 concentrations from MMF). Table 3.9 lists the corresponding statistical diagnostics for the surface layer (approx. 0 – 200 m) and the lowest 2 km. At higher altitudes, MAX-DOAS measurements typically show very low sensitivity (see e.g. the averaging kernel matrix of the measurements at Heidelberg in Fig. 2.13). Note, that the data of Table 3.9 is not specific to individual instrument locations; a quantitative summary of the simulation biases separated by location is presented in Table 3.11.

The agreement between WRF-Chem and the MAX-DOAS retrievals varies strongly with the instrument location. For example, excellent agreement between WRF-Chem and MMF is obtained in De Bilt, whereas MAPA shows severe disagreement. This, however, appears to be related to De Bilt specifically and is not observed elsewhere. Nonetheless there are considerable differences between MAPA and MMF at all locations, indicating the magnitude of the retrieval uncertainty. Furthermore, MAPA tends to produce larger average concentrations and more complex profile shapes than MMF (with the exception of Bremen). For example, the characteristic concentration falloff at the top of the PBL is still faintly resembled in the MAPA profiles at Heidelberg and Mainz, with approximate PBL heights of 1 km. The corresponding MMF profiles, on the other hand, have a mostly smooth exponential shape. An exception is Bremen, where both MMF and MAPA yield complex profiles with more "jagged" shapes. According to the descriptions found in Bösch (2019), the MAX-DOAS instruments in Bremen were operated in a viewing direction of two nearby power plants, which could emit NO_2 plumes aloft, leading to more complex profile shapes. A similar conclusion was drawn in Kuhn et al. (2024a), using only one instrument per location (as opposed to the average over all colocated instruments) and an older retrieval version (fv002). Furthermore, horizontal NO_2 gradients can cause apparently elevated profiles in MAX-DOAS retrievals. The definite reason for the resulting profiles in Bremen remains unknown. However, an important observation to be made here is that none of the WRF-Chem simulations (with or without vertical emission profiles) show matching elevated layers. An explanation for this could be the spatial resolution of the WRF-Chem simulations of approximately 3 km × 3 km × 100 m at 500 m altitude, while

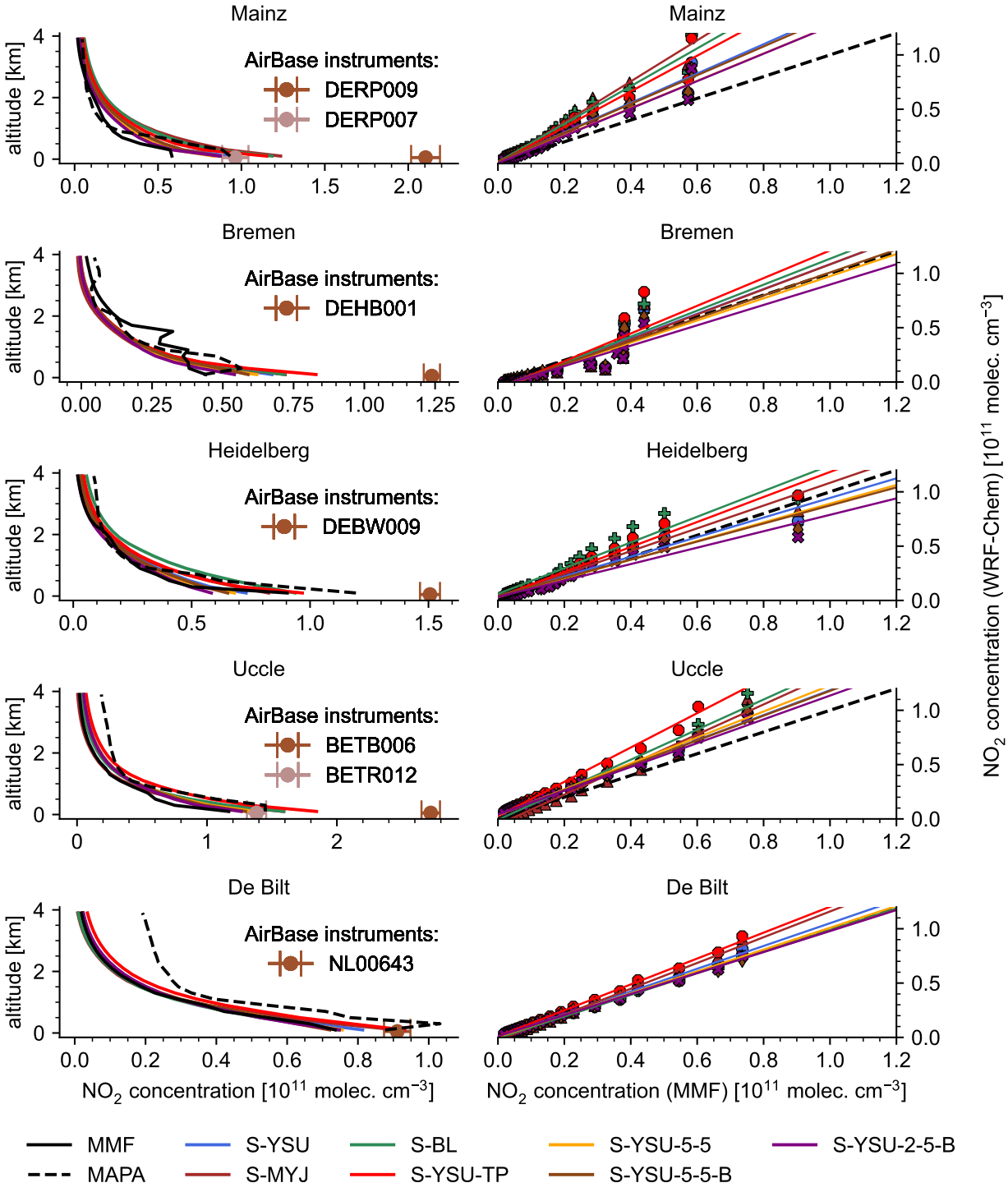

Figure 3.17: Comparison of monthly-mean NO_2 profiles from different simulation runs to NO_2 profiles from MAX-DOAS retrievals. The left column shows the average NO_2 profiles from 11 AM – 2 PM. Colocated AirBase background measurements within a 5 km radius are drawn at an altitude of 0 m. The right column shows scatter plots (WRF-Chem vs. MMF), where each point represents the monthly-mean concentration of one retrieval layer.

		YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
lowest layer vs. MMF	bias [%]	+23.0	+48.8	+47.8	+58.2	+16.1	+13.7	+10.9
	RMSE	2.8	4.7	4.4	5.0	2.5	2.5	2.6
	R	0.63	0.49	0.57	0.68	0.67	0.61	0.55
lowest 2 km vs. MMF	bias [%]	+19.7	+35.7	+41.0	+43.3	+13.5	+16.9	+7.9
	RMSE	1.3	2.2	2.2	2.3	1.2	1.3	1.1
	R	0.90	0.85	0.86	0.90	0.90	0.89	0.88
lowest layer vs. MAPA	bias [%]	−9.2	+9.8	+9.1	+16.8	−14.3	−16.1	−18.1
	RMSE	4.4	4.7	4.4	4.7	4.4	4.6	4.7
	R	0.48	0.39	0.48	0.49	0.51	0.45	0.45
lowest 2 km vs. MAPA	bias [%]	−11.2	+0.7	+4.6	+6.3	−15.8	−13.2	−19.9
	RMSE	2.2	2.2	2.1	2.0	2.2	2.2	2.4
	R	0.86	0.82	0.83	0.86	0.87	0.86	0.86

Table 3.9: Validation of different simulation runs against NO_2 profiles from FRM_4DOAS based on monthly-mean data. RMSE values are given in units of 10^{10} molec. cm^{-3}. All values were obtained in from the time window of 11 AM - 2 PM.

		YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
lowest layer vs. MMF	bias [%]	+21.4	+45.0	+43.5	+56.7	+15.6	+12.4	+9.4
	RMSE	9.1	10.3	10.0	10.5	8.7	8.8	8.9
	R	0.39	0.38	0.36	0.44	0.42	0.40	0.34
lowest 2 km vs. MMF	bias [%]	+17.4	+29.3	+35.1	+41.3	+12.4	+14.7	+6.5
	RMSE	5.1	5.6	5.6	5.7	4.9	5.0	4.8
	R	0.44	0.45	0.42	0.46	0.45	0.43	0.43
lowest layer vs. MAPA	bias [%]	−7.4	+10.7	+9.6	+19.5	−11.6	−14.4	−16.3
	RMSE	32.9	33.5	33.3	32.6	32.9	32.9	33.2
	R	0.20	0.13	0.15	0.24	0.21	0.20	0.16
lowest 2 km vs. MAPA	bias [%]	−13.6	−4.7	−0.5	+4.2	−17.1	−15.6	−21.3
	RMSE	13.6	13.9	13.9	13.6	13.6	13.6	13.7
	R	0.25	0.22	0.22	0.27	0.26	0.24	0.23

Table 3.10: Like Table 3.9, but based on individual profiles.

	YSU	MYJ	BL	YSU-TP	YSU-5-5	YSU-5-5-B	YSU-2-5-B
Mainz (vs. MMF)	+59.1 %	+111.8 %	+103.0 %	+97.5 %	+43.5 %	+48.3 %	+49.8 %
Bremen (vs. MMF)	+53.0 %	+63.8 %	+63.3 %	+88.5 %	+40.9 %	+33.9 %	+22.9 %
Heidelberg (vs. MMF)	−19.2 %	−8.7 %	+3.6 %	+7.0 %	−24.7 %	−27.9 %	−35.5 %
Uccle (vs. MMF)	+13.9 %	+36.1 %	+36.3 %	+57.8 %	+16.9 %	+11.0 %	+8.2 %
De Bilt (vs. MMF)	+10.6 %	+23.7 %	+1.1 %	+26.6 %	+2.7 %	−2.1 %	+1.6 %
Mainz (vs. MAPA)	−0.5 %	+32.4 %	+26.9 %	+23.5 %	−10.3 %	−7.3 %	−6.4 %
Bremen (vs. MAPA)	+56.1 %	+67.0 %	+66.6 %	+92.2 %	+43.8 %	+36.6 %	+25.4 %
Heidelberg (vs. MAPA)	−38.8 %	−30.9 %	−21.5 %	−19.0 %	−43.0 %	−45.4 %	−51.1 %
Uccle (vs. MAPA)	−8.2 %	+9.7 %	+9.8 %	+27.1 %	−5.8 %	−10.6 %	−12.9 %
De Bilt (vs. MAPA)	−7.3 %	+3.8 %	−15.2 %	+6.2 %	−13.8 %	−17.9 %	−14.8 %

Table 3.11: Biases of different simulation runs against FRM$_4$DOAS NO_2 concentrations in the lowest retrieval layer (approx. 0 - 200 m) based on monthly-mean data. All values were obtained from the time window of 11 AM - 2 PM.

stack plumes typically have a cross sectional area of e.g. 100 m × 100 m near the point of emission (compare Fig. 2.16). This leads to a strong dilution of the plume concentration in the model. However, there are scenarios, mainly close to very strong emitters, where WRF-Chem indeed predicts elevated NO_2 layers, e.g. in the vicinity of the Bełchatów power plant (see Fig. B.7).

Another relevant aspect in this validation is the agreement to colocated AirBase background measurements within a 5 km radius, drawn as additional markers in the left-side subplots of Fig. 3.17. In many cases, MMF, MAPA, and WRF-Chem significantly underestimate the AirBase measurements. This was already described by Bösch (2019) for the measurements in Bremen, however, without consideration of the Mo-CL bias. The Mo-CL bias is accounted for here, but the underestimations persist, demonstrating that they require further reasoning. It is justified to assume, that the vertical NO_2 gradients near the surface cannot be well resolved with a grid spacing of 200 m. This claim can be verified as follows: Figure 3.18 shows the left-side subplots of Fig. 3.17, but without application of the averaging kernels and no vertical interpolation of the WRF-Chem NO_2 profiles, i.e. the "raw" profiles. Due to the higher vertical resolution of the raw simulated profiles near the surface, this results in significant enhancements of the surface NO_2 concentrations, which even overestimate the AirBase measurements in some cases. Although this brief analysis can not guide towards further optimization of the WRF-Chem simulation setup, it is nonetheless helpful in demonstrating the influence of vertical resolution near the surface, for RCT simulations and MAX-DOAS retrievals alike. For example, the simulation run S-YSU-2-5-B produced NO_2 concentrations of approx. $2.1 \cdot 10^{11}$ molec. cm^{-2} at the surface (0 – 8 m), but $1.9 \cdot 10^{11}$ molec. cm^{-2} averaged over the lowest five WRF-Chem layers (0 – 209 m). This corresponds to a relative loss of 10 % just from reduction of the vertical resolution.

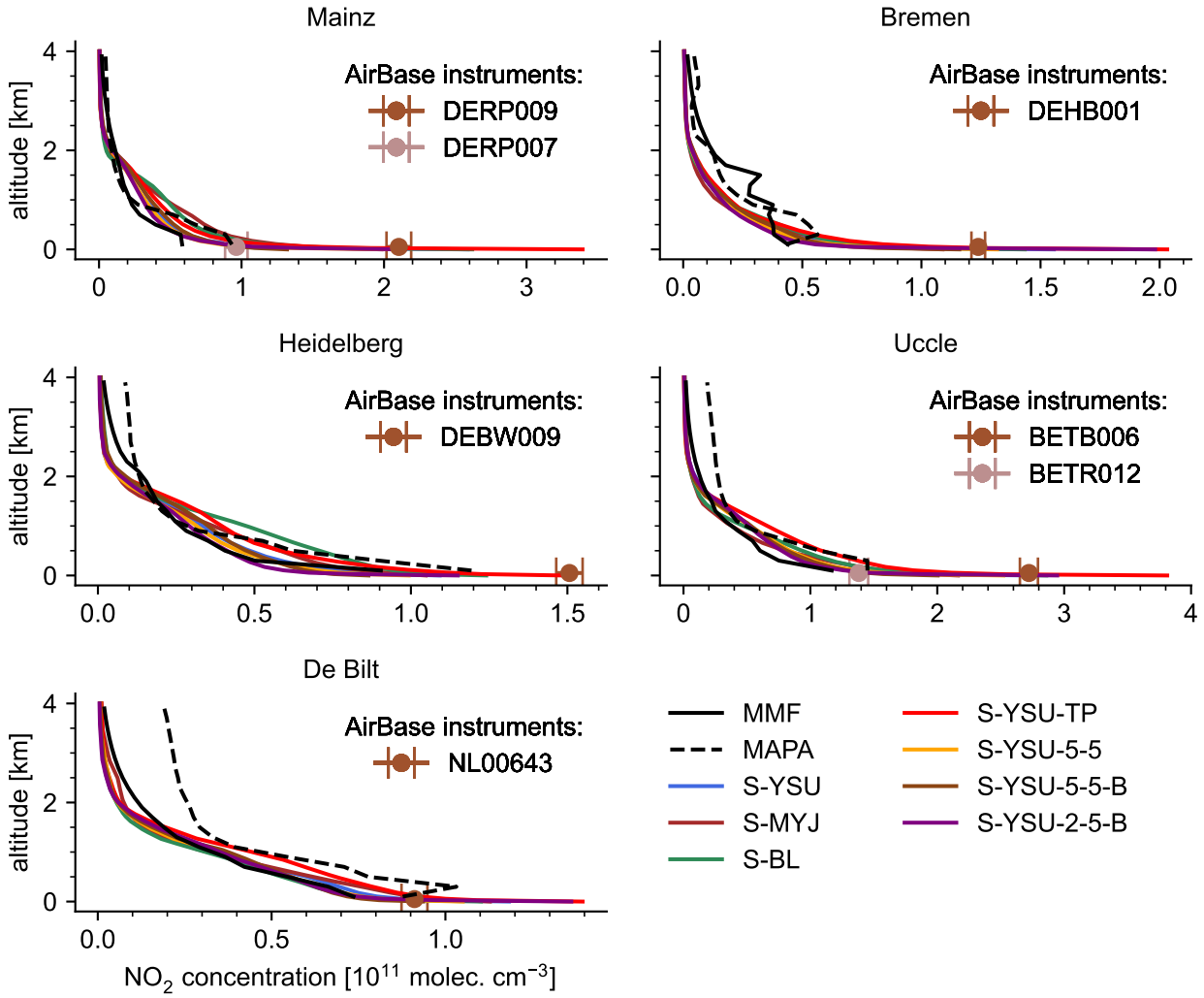

Figure 3.18: Comparison of monthly-mean NO_2 profiles from different simulation runs to NO_2 profiles from MAX-DOAS retrievals without averaging kernels or vertical interpolation. Shown here are average NO_2 profiles from 11 AM – 2 PM.

Up to this point, the analysis was restricted to profiles within the 11 PM – 2 AM time window. It can be extended by a brief overview of the full daytime cycles (5 AM to 6 PM) of surface layer NO_2 concentrations, as shown in Fig. 3.19. On one hand, there is qualitative agreement between WRF-Chem, MMF, and MAPA with respect to the shape of the diurnal cycles, characterized by enhanced NO_2 concentrations in the morning and evening. On the other hand, many of the WRF-Chem simulations overestimate the NO_2 concentration during the early morning. Similar results were found in the comparison to the spatially averaged AirBase in situ measurements, shown in Fig. 3.15. When analyzed on a per-location basis, these overestimations are revealed to vary strongly across the model domain. Notably, the simulation runs with revised vertical mixing tend to outperform the others, with particularly good agreement to MMF and MAPA in Heidelberg, and to MAPA in De Bilt. A corresponding

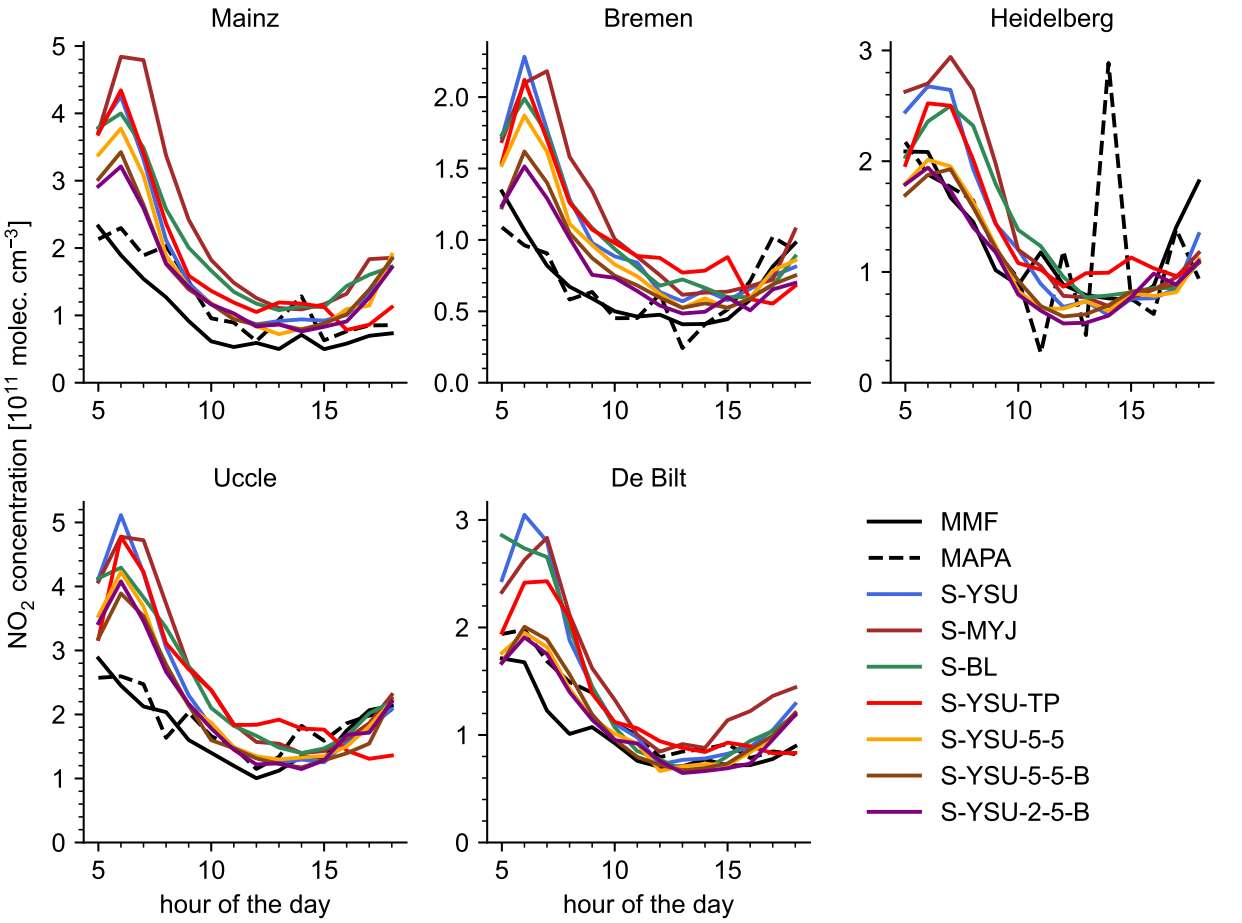

Figure 3.19: Comparison of monthly-mean NO_2 concentrations from WRF-Chem and MAX-DOAS in the lowest retrieval layer (approx. 0 – 200 m).

quantitative summary is found in Table 3.11.

3.8 Summary, discussion, and conclusions

This section has explored the capabilities of the RCT model WRF-Chem in numerically simulating tropospheric NO_2 in central Europe during summertime. The model was operated with different configurations, and the results were compared to three observational datasets: in situ NO_2 and NO_x measurements from the AirBase network, tropospheric NO_2 VCDs from the TROPOMI satellite instrument, and NO_2 concentration profiles from MAX-DOAS measurements. In the comparison of the model's surface NO_2 concentrations to background AirBase in situ measurements, the main findings from previous model evaluation papers (e.g. Poraicu et al., 2023; Kuik et al., 2016) were successfully reproduced, particularly the strong overestimations at night and moderate underestimations around noon (if using vertical emission profiles). Furthermore, similar overestimations of the NO_2 VCD in comparison to TROPOMI

data were identified in western Germany, the Netherlands, and Belgium. This motivated a detailed analysis of different model components and their influence on the observed model errors, including the model's vertical mixing routine, its temporal and vertical emission profiles, its NO_x chemistry, and the correction of Mo-CL biases for in situ measurements based on modelled photooxidant VMRs. The main findings of this study can be summarized as follows:

Correction of the Mo-CL bias

The Mo-CL bias correction based on modelled PAN and HNO_3 VMRs for molybdenum-based chemiluminescence in situ measurements of NO_2 was implemented following Lamsal et al. (2008). The correction factors showed a diurnal cycle peaking at noon, which relates to the photochemical production of PAN and HNO_3 (as well as the alkyl nitrates, which were not considered here). The noontime correction factors in Germany reached an average value of 1.22. This is in good agreement to the results reported by Poraicu et al. (2023) in a comparable WRF-Chem simulation setup, and previous estimates of the Mo-CL bias based on instrument studies (see the references listed in sect. 2.3.6).

Temporal and vertical emission profiles

The simulation results revealed a strong sensitivity to the choice of temporal and vertical emission profiles. A mass-conservative redistribution of the emissions along the diurnal cycle ("tuned temporal emission profiles") was found to be potentially useful for compensating the model's over- and underestimations of NO_2 at different times of the day. However, the feasibility of this approach highly depends on the model setup (e.g. whether the Mo-CL bias is corrected for, and whether vertical emission profiles are used) and is associated with considerable disadvantages summarized further below. Vertical emission profiles were found to decrease the modelled NO_2 surface concentration by approx. 17 % and produce NO_2 VCDs in significantly better agreement to the TROPOMI measurements in regions of average to high pollution.

Vertical mixing routine

A revised vertical mixing routine was implemented, which circumvents many problems of WRF-Chem's original mixing routine. With the revised routine, the clipping thresholds ($k_{\mathrm{min,urban}}$ and $k_{\mathrm{min,rural}}$) of the model's turbulent diffusion coefficients can be fine-tuned for urban and rural domains, as intended by the developers. Furthermore, the distinction between rural and urban regions (RU-distinction) was changed to use land-use data instead of arbitrary emission thresholds, which was identified to be more accurate, as well as robust towards the use of vertical and temporal emission profiles. It was found that the nighttime overestimations

and daytime underestimations of NO_2 and NO_x could be reduced significantly by increasing or reducing the clipping thresholds of the turbulent diffusion coefficients. An optimization of $k_{\mathrm{min,urban}}$ and $k_{\mathrm{min,rural}}$ was conducted based on the comparison of modelled diurnal surface NO_2 and NO_x cycles to background AirBase in situ measurements. This analysis used data from the German subdomain with particularly high emission resolution ($0.01° \times 0.01°$ in Germany vs. $0.1° \times 0.1°$ elsewhere). For simulation configurations with vertical emission profiles, a parametrization of $k_{\mathrm{min,urban}} = 2\ \mathrm{m^2\ s^{-1}}$ and $k_{\mathrm{min,rural}} = 5\ \mathrm{m^2\ s^{-1}}$ was found to be ideal. The original mixing routine of WRF-Chem marks a serious shortcoming of the model. The WRF-Chem developer team should consider a revised implementation that is less opaque to the user and makes the mixing thresholds accessible through the namelist interface.

NO_x chemistry

After the optimization of vertical mixing, the model's NO_x chemistry was investigated on the basis of its surface NO_2/NO_x ratios in rural regions, where some remaining daytime discrepancies to the in situ NO_2 and NO measurements were identified. Here, the modelled NO_2 was found to be too large, while NO was too low. It was shown, that under the assumption of steady state, only a small part of these differences could be attributed to the model's consistent overestimation of surface O_3. The remaining discrepancies were found to be likely related to inaccuracies in the model's organic NO_x chemistry (e.g. an overestimation of VOCs).

Deliberations on the ideal model configuration

An important question is, which of the presented WRF-Chem simulation runs should be selected to provide the training data for the NitroNet neural network. The agreement to MAX-DOAS measurements is an interesting aspect in this consideration, but arguably not very substantial due to the measurements' sparsity. It is therefore more reasonable to base the assessment of model configurations mainly on the comparison to the AirBase in situ measurements and TROPOMI satellite observations. The statistical diagnostics summarized in the following were computed on monthly-mean data, unless specified otherwise.

The comparison to TROPOMI NO_2 VCDs gave clear results. As expected, the agreement between simulated and observed NO_2 VCDs was found to improve significantly upon re-computing the air mass factors using the model's high resolution NO_2 profiles. The corresponding VCD enhancements of approx. 20 % were slightly larger than what is reported in the literature (approx. 15 %, see sect. 2.3.4). This could be due to the particularly high spatial resolution of our WRF-Chem simulation compared to other model studies. All simulation configurations produced generally acceptable results, with monthly-mean biases below 10 %, correlation coefficients of approximately $R = 0.86$ and RMSE values of approximately

$1 \cdot 10^{15}$ molec. cm^{-2}. However, the runs with the revised mixing routine (S-YSU-5-5, S-YSU-5-5-B, and S-YSU-2-5-B) slightly outperformed the others. The improvements were revealed to be most significant in the strongly polluted regions of the domain, e.g. in western Germany, where the simulations showed a tendency to overestimate the NO_2 VCD. This aligns well with the aforementioned literature, which indicates that the TROPOMI NO_2 VCD is most strongly affected by exchange of the a priori profiles in strongly polluted regions. On the whole, the comparison to TROPOMI NO_2 VCDs favours the simulation run S-YSU-2-5-B. This was the simulation run with the optimized vertical mixing routine ($k_{\mathrm{min,urban}} = 2$ m^2 s^{-1}, $k_{\mathrm{min,rural}} = 5$ m^2 s^{-1}), the original temporal emission profiles proposed by Kumar et al. (2021), and the vertical emission profiles proposed by Bieser et al. (2011).

The comparison to AirBase in situ measurements is more complicated. When evaluated against the German AirBase data, the simulation runs with optimized vertical mixing clearly outperformed the standard model configurations. The main advantage of the revised mixing routine is the significant reduction of nighttime overestimations and daytime underestimations of NO_2 and NO_x. This was demonstrated to be the case regardless of whether vertical emission profiles were used (see Fig. 3.12) or not (see Fig. 3.8). In the former case, the few remaining differences in the simulated and measured diurnal cycles of NO_2 and NO could be attributed to faulty NO_2/NO_x ratios, as shown in Fig. 3.13. Up to this point, still restricted to the German subdomain, it appeared evident that S-YSU-5-5, and S-YSU-2-5-B were the best performing simulation runs. However, this became less clear when extending the validation to the full domain, where the simulations showed far stronger underestimations of the noontime surface NO_2 concentrations. Then, judging by the statistical diagnostics presented in Tables 3.5 and 3.6, the simulation run S-YSU-TP with tuned temporal emission profiles appeared to be the top candidate. However, there are multiple arguments against that conclusion, listed in the following:

- The fundamental premise of S-YSU-TP, that the model errors are caused by unrealistic diurnal emission patterns, is highly questionable. The supposedly improved temporal emission profiles were obtained by "empirical optimization". Even worse, the resulting emission profiles are in obvious qualitative disagreement to those derived from scientifically sound methods, for example from measurements of car traffic (see e.g. Kumar et al., 2021). For these reasons, the approach was heavily criticized during the discussion phase of an early publication on the topic (see Kuhn et al., 2024a for the final publication, Kuhn et al., 2023 for the preprint, and the reviewer's comment RC1 from 7 February 2023), and subsequently abandoned.

- S-YSU-TP contained no vertical emission displacement, which skews the interpretation in its favour compared to the runs S-YSU-5-5-B, and S-YSU-2-5-B. Comparing the sim-

ulation runs S-YSU-5-5 and S-YSU-5-5-B, the use of vertical emission profiles results in an average reduction of the modelled surface NO_2 by approx. 17 % (in reference to the full diurnal cycle) or 11 % (at noontime), as shown in Table 3.5. S-YSU-TP had an average surface NO_2 bias of +2.9 % at noontime, and +5.3 % throughout the day. A hypothetical simulation run S-YSU-TP-B with vertical emission profiles is therefore expected to result in equally lower model biases. However, as the approach was abandoned for the reason given above, no such simulation run was produced.

- Based on the comparison of Fig. 3.8 and 3.15, extending the validation from the German subdomain to the full model domain resulted in approx. 23 % lower bias in modelled NO_2 surface concentrations. It is reasonable to at least partly attribute this to the comparably poor resolution of the emission data outside of Germany. An exact quantification of this aspect is not necessary here, but could be obtained from corresponding simulation runs with artificially degraded emission data in the future. Similar notions were raised by Kuik et al. (2016). Combined with the strong overestimations of S-YSU-TP on the German subdomain, this makes it plausible that the tuned emission profiles are in fact erroneous, but happen to cancel out the deficits resulting from the limited emission resolution on the full model domain.

- S-YSU-TP was the worst performing run in the validation against TROPOMI satellite observations.

Based on these considerations, a conclusion in favour of S-YSU-TP can not be substantiated. It should be noted, that a combination of tuned temporal emission profiles and revised mixing would not escape the preceding arguments either. Therefore, the simulation run S-YSU-TP is discarded from hereon. The remaining simulation configurations can be separated into two classes:

- configurations without vertical emission profiles, (S-YSU, S-MYJ, S-BL, S-YSU-5-5)

- configurations with vertical emission profiles (S-YSU-5-5-B, S-YSU-2-5-B, and runs corresponding to intermediate optimization steps for the vertical mixing parametrization)

It is obviously unrealistic to inject all emissions into the model's surface layer. If it leads to better agreement nonetheless, it is most likely by compensation of some other unidentified model deficiency. Therefore it is reasonable to search for the optimal model configuration among those which utilized vertical emission profiles. As demonstrated in sect. 3.6.2, this is the run S-YSU-2-5-B. The arguments in favour of the configuration S-YSU-2-5-B can be summarized as follows:

WRF-Chem compared to	bias [%]	RMSE	R
TROPOMI	+1.6 **(+1.6)**	7.9 **(18.8)** $\cdot 10^{14}$ molec. cm^{-2}	0.86 **(0.63)**
AirBase (noontime)	−20.3 **(−20.0)**	5.1 **(6.9)** µg m^{-3}	0.49 **(0.44)**
FRM$_4$DOAS (MMF)	+7.9 **(+6.5)**	1.1 **(4.8)** $\cdot 10^{10}$ molec. cm^{-3}	0.88 **(0.43)**
FRM$_4$DOAS (MAPA)	−19.9 **(−21.3)**	2.4 **(13.7)** $\cdot 10^{10}$ molec. cm^{-3}	0.86 **(0.23)**

Table 3.12: Quantitative summary of the performance of model run S-YSU-2-5-B against TROPOMI, AirBase, and FRM$_4$DOAS. The black numbers represent the results obtained from monthly-mean data. The coloured bold numbers represent the results obtained without averaging (e.g. individual orbits from TROPOMI). The corresponding data was taken from Tables 3.5, 3.6, 3.7, 3.8, 3.9, and 3.10. The FRM$_4$DOAS data were evaluated in a time window from 11 PM – 2 AM for the lowest 2 km of the atmosphere.

- its implementation follows the WRF-Chem developers' advice on tuning the vertical mixing parametrization, see Code Excerpt 3.1 and the corresponding discussion in sect. 3.4 and 3.5.
- it benefits from the demonstrably better distinction of urban and rural regions (see Fig. 3.11 and the discussions in sect. 3.6.1).
- compared to the other simulation runs, it performed best in the validation against TROPOMI satellite observations, and produced among the lowest RMSE values in the validation against AirBase in situ data. It is also one of the best performing configurations in the comparison to MAX-DOAS measurements in most locations, particularly in comparison to runs without vertical emission profiles.

A final, brief summary of S-YSU-2-5-B with respect to all three reference datasets is given in Table 3.12.

Outlook

This chapter has explored various approaches to improving future WRF-Chem simulations of tropospheric NO_2 on European domains during summertime. Nonetheless a few open questions remain, which warrant further research and methodical refinements. The fact that WRF-Chem relies on the described clipping of the vertical mixing coefficients shows, that numerical modelling of atmospheric mixing is still poorly understood. Furthermore, fine-tuning of the clipping thresholds should ideally be conducted in a more rigorous manner, as opposed to the brute-force approach used here. On the other hand, standard methods like gradient descent are hard to combine with complex simulations that take days to compute. Either way, this problem should ultimately be solved by improved PBL schemes, which produce correct mixing coefficients and require no clipping.

The validation of the simulation could be further extended in the future. Comparisons to TROPOMI's tropospheric formaldehyde VCD may be used along with CO in situ measurements at the surface to assess, whether the atmosphere's oxidative capacity and VOC loads are accurately represented in the simulation. Once available, NO_2 VCDs from the geostationary Sentinel-4 satellite mission (Stark et al., 2013) could be used for model validation studies with enhanced temporal resolution. Another significant step could be to increase the horizontal resolution of the model, e.g. to 1 km $\times$ 1 km, and investigate the associated benefits.

In closing this chapter, a perspective on using the WRF-Chem model results for training the NitroNet neural network is given. The vast majority of RCT model evaluation literature is produced with the intent of testing a model's overall performance, and possibly deducing means to improve it. In such analyses, it is necessary to include and discuss all deviations between the model and the reference data, even if they are clearly confined to small, localized sub-regions of the model domain, and could in theory be filtered out easily. An example would be the severe NO_2 overestimations over western Germany described previously. It could be argued that such model errors are perhaps linked to inaccuracies in the emission inventory or other input data, and should not be attributed to the RCT model itself. Nonetheless, model evaluation papers usually do not resort to discarding sub-regions of the model domain in order to assess the model performance. However, this thesis aims not to validate and optimize the WRF-Chem model as an end in itself, but to use it to produce suitable training data for the NitroNet neural network. In that scope, it is reasonable to filter out model results, if they can be determined as faulty prior to training. This procedure is called *data wrangling* (or *data filtering*, *data curation*). By rejecting certain training data, e.g. from the polluted regions of western Germany, the neural network can be partly prevented from adopting systematic errors of the regional model. This means that the performance of the WRF-Chem simulation, quantified here by the bias, RMSE, and R-value in comparison to reference data, does *not* pre-define the performance of the resulting neural network. NitroNet's performance can not be estimated beforehand, and must be determined empirically later on.

Chapter 4

The NitroNet model

The preceding chapter has demonstrated, what level of accuracy can be expected in numerical simulations of tropospheric NO_2 with the WRF-Chem RCT model. Note, that these results are highly specific to the chosen season (summertime) and geographic region (central Europe). On one hand, the optimization of the model's vertical mixing parametrization and subsequent validation against observational reference data resulted in far lower model errors compared to out-of-the-box simulation setups. On the other hand, the presented optimization approach has considerable limitations. Due to WRF-Chem's long runtimes, the optimization of the summertime mixing clipping thresholds took months to complete. Furthermore, it was not yet investigated whether the obtained thresholds are approriate for other spatio-temporal domains. Lastly, even the optimized simulation was plagued by significant errors, e.g. the severe NO_2 overestimations in western Germany. The central premise of this thesis is that the aforementioned hurdles can be partly overcome with a machine learning model, e.g. a neural network, trained to predict NO_2 profiles based on the dataset of WRF-Chem simulation results, referred to as "WRF-2019" from hereon. The following chapter presents an implementation of such a neural network based profile retrieval named "NitroNet". NitroNet takes NO_2 VCDs along with ancillary variables as input and returns matching NO_2 profiles. The input data are mostly observational (e.g. the TROPOMI NO_2 VCD), but include some variables from atmospheric reanalyses (e.g. the PBLH from ERA5). Compared to classical RCT models like WRF-Chem, NitroNet is orders of magnitude faster, which saves time in operational use and enables regional and seasonal validation studies. Most importantly, however, NitroNet is trained on a selected subset of WRF-2019, which meets certain validation criteria based on model-external reference data (*data curation*) and is thereby partly prevented from adopting WRF-Chem's systematic errors.

The following part of this thesis revolves around the design, the training process, and the validation of NitroNet. This includes an intercomparison of NitroNet predictions, WRF-Chem simulation results, and the independent measurements used for validation of the WRF-Chem

simulations before (in situ NO_2 measurements from AirBase, tropospheric NO_2 VCDs from TROPOMI, and MAX-DOAS NO_2 profiles from FRM_4DOAS). The impact of NitroNet's individual input variables on its overall prediction quality is investigated in a feature relevance analysis. Lastly, a regional and seasonal validation study of NitroNet is presented. The content of this chapter corresponds in large parts to Kuhn et al. (2024b).

4.1 Model description

parameter	value
hidden layers	8
neurons per layer	326
activation function	PReLU
loss function	L_1
batch size	2048
optimizer	NAdam
initial learning rate	$3.42 \cdot 10^{-4}$
Δ_{VCD}	0.2
Δ_{PBLH}	0.1

Table 4.1: Hyperparameters of the NitroNet neural network. For a combined reference of the terms used here, refer to sects. 2.5, 4.2, 4.2.6, and the references therein.

The NitroNet model consists of an artificial feed-forward neural network at its core, together with additional non-machine learning code for fetching and pre-processing of the required input data and for input uncertainty propagation as described in sect. 2.5.7. A feed-forward network is used instead of a more complex neural network (e.g. a CNN) for being easier to implement and train, while possessing theoretically universal approximation capabilities nonetheless. NitroNet was designed to integrate seamlessly into high performance computing (HPC) environments. In particular, NitroNet can run on multiple graphics processing units (GPUs) in parallel, which, compared to central processing units (CPUs), accelerates its use by orders of magnitude.

NitroNet's neural network uses the feed-forward topology, as described in sect. 2.5.1. The neural network consists of 8 hidden layers, with 326 neurons each, corresponding to approximately 850000 trainable parameters. The model's hyperparameters listed in Table 4.1 were selected within the scope of a hyperparameter optimization, in which over 100 different variants of the neural network were tested (see the results in sect. 4.2.6). The network configuration described here and in the following was found to be the best-performing.

4.1.1 Input variables

The main input variable of NitroNet is the tropospheric NO_2 VCD from TROPOMI. As depicted in Fig. 4.1a-b, it further receives a variety of additional input variables (e.g. meteorological data, emissions, other satellite observations, etc.), referred to as the *ancillary variables*. These input variables were selected on the assumption that they provide helpful information for the

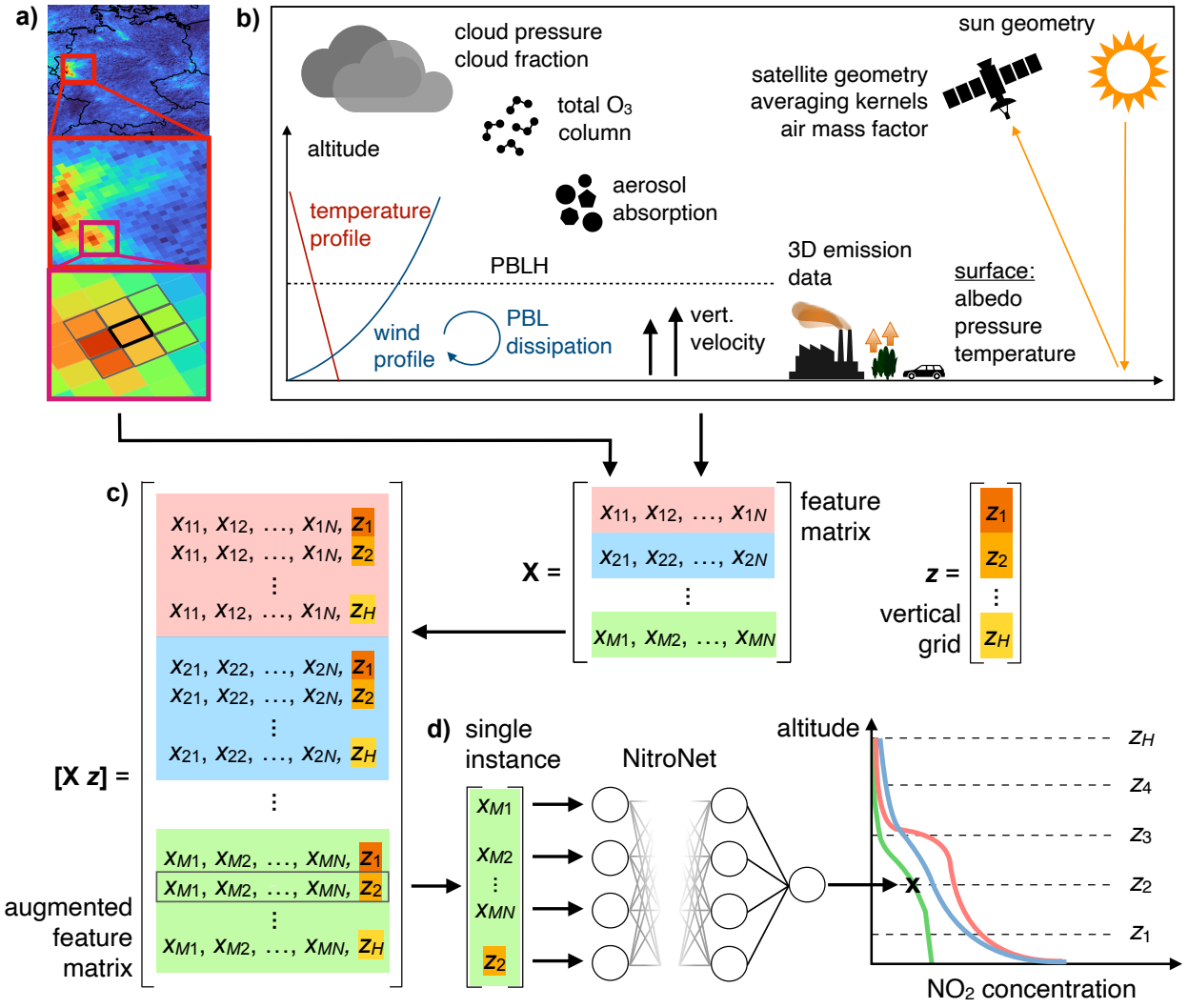

Figure 4.1: Overview of the NitroNet model. a) The TROPOMI NO_2 VCD, used as the main input. The zoomed-in maps show individual TROPOMI observations. For each TROPOMI ground pixel, NitroNet receives the NO_2 VCD as input and also computes an "NO_2 VCD influx" variable from its 8 immediate neighbours. **b)** Exemplary ancillary input variables. **c)** Construction of the augmented feature matrix [**X** ***z***] from the feature matrix **X** and the vertical output grid ***z*** (see sect. 4.1.2). **d)** Forward pass of a single instance of the augmented feature matrix through the neural network, returning a prediction of the NO_2 concentration at a specified altitude above ground.

prediction of tropospheric NO_2 profiles (see also the results of the "feature relevance analysis" in sect. 4.2.9). If required (e.g. for meteorological data from ERA5), the ancillaries are bilinearly interpolated to the horizontal grid of the TROPOMI observations. NitroNet works on a per-pixel basis, i.e. each network call receives input data from a single TROPOMI ground pixel (with exception of the NO_2 influx variable discussed further below) and predicts a corresponding NO_2 profile. Thus, its horizontal output grid matches that of the TROPOMI observations. Unlike convolutional neural networks, the feed-forward topology of NitroNet

is not designed to process input data in awareness of spatial context automatically (e.g. by including meteorological data and satellite observations belonging to adjacent satellite pixels). Nonetheless, a rudimental spatial contextualization is achieved by means of the "NO_2 VCD influx" input variable, discussed further below. When choosing suitable input variables for NitroNet, there are two relevant limitations, namely:

1. The input variables must be in some relationship (e.g. causal, functional, or by correlation) to the desired targets (the NO_2 concentration profiles). It does not matter, whether this relationship can be formulated in analytical terms or not.

2. The input variables must be accessible at runtime. This is obviously the case for continuously produced observational data, e.g. the TROPOMI satellite observations, but may also include variables from operational atmospheric reanalyses (e.g. the PBLH from ERA5). Note, that WRF-2019 contains many variables, which can certainly be expected to be highly informative for the prediction of NO_2 profiles (e.g. the simulated NO profiles), but cannot be used in NitroNet, because they are not available at runtime without running a WRF-Chem simulation. Some variables, e.g. the tropospheric NO_2 VCD, exist both as a modelled variable in WRF-2019 and from some other external data source (here: from TROPOMI measurements). In that case, the variable is taken from the external data source during training and at runtime. NitroNet's training targets (the NO_2 profiles) are exempt from this rule and taken from WRF-2019, because they are only required during training, not during runtime. For an overview of machine learning terms such as "targets", "features", etc. refer to sect. 2.5.

TROPOMI input data

Table 4.2 gives an overview of all input variables used by NitroNet. For the tropospheric NO_2 VCD and total O_3 VCD the TROPOMI product v.02.04.00 is used. Although the O_3 VCD is mostly stratospheric, it is assumed here that some helpful information can be extracted from its small tropospheric component. The tropospheric averaging kernels are computed as described in sect. 2.3.4, and defined on the vertical grid of the TM5-MP model. Here, the tropospheric averaging kernels of the 9 lowest TM5-MP layers were used (reaching up to $\sim$ 2300 m above ground). The surface classification of TROPOMI (see Eskes et al., 2022) is used for ternary distinction of surface types into "urban", "cropland", and "forest". The variable "NO_2 VCD influx" describes the influx of NO_2, that a TROPOMI pixel receives from its 8 immediate neighbours due to horizontal transport. The influx variable is computed as

$$F_{\text{in}} = \sum_i w_i V_{t,i}^{\text{trop}} \tag{4.1}$$

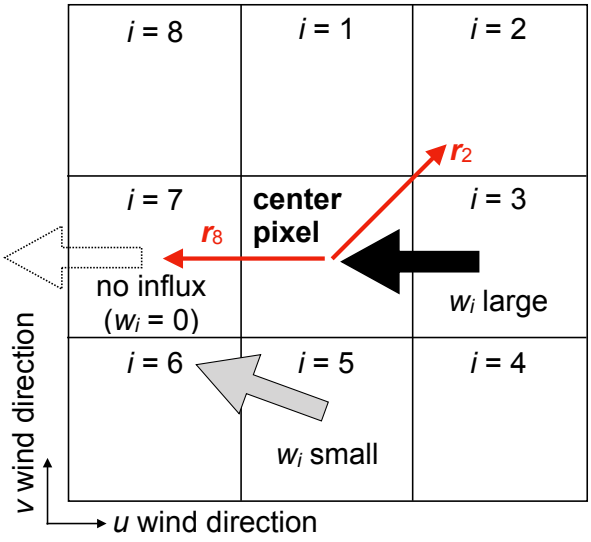

Figure 4.2: Computation of the "NO_2 VCD influx" variable. The thick arrows symbolize wind direction. The thin red arrows symbolize the unit vectors $\boldsymbol{r}_i$ (here: for $i = 2$ and $i = 8$) pointing from the center pixel to its immediate neighbours.

where:

i : index of neighbour pixel i, with $1 \leq i \leq 8$

w_i : inflow wind component of neighbour pixel i

$V_{t,i}$: tropospheric NO_2 VCD in of neighbour pixel i

The influx wind component depends on the wind direction in the neighbouring pixels. Let $\boldsymbol{r}_i$ denote the unit vector pointing from the center pixel to any neighbour pixel i, as depicted in Fig. 4.2. Then, w_i is computed as

$$w_i = \max\left(-\boldsymbol{r}_i \cdot (u, v)^{\mathrm{T}}, 0\right) \tag{4.2}$$

where:

$\boldsymbol{r}_i$: unit vector pointing from the center pixel to neighbour pixel i

u : eastward wind component

v : northward wind component

The max function is required to ensure that the resulting influx is always positive. The computation of w_i currently uses the surface wind speeds from ERA5, but future implementations might use the average wind speeds within the boundary layer instead. The resulting flux represents line density per time, and has units of molec. cm^{-1} s^{-1}. Note that there is no necessity

of computing an "outflux" variable, because NitroNet receives the center pixel's tropospheric NO_2 VCD and ERA5 wind speeds, which fully determine the outflux.

ERA5 input data

All ERA5 variables are documented on the website `https://codes.ecmwf.int/grib/param-db`, and accessible via their variable IDs given in Table 4.2. The variables "wind speed" and "vertical velocity" are resolved at 6 different pressure levels (1000 hPa, 950 hPa, 900 hPa, 850 hPa, 750 hPa, and 700 hPa). "Wind speed" refers to the absolute horizontal wind speeds. "Vertical velocity" refers to the speed of air motion in upward direction. "Boundary layer dissipation" quantifies the conversion of kinetic energy into heat due to small-scale eddies (turbulences) in the PBL.

Emission input data

NitroNet receives the total NO_x emissions (i.e. the sum over all emission sectors) from the EDGARv5 emission inventory. In addition, the contribution of four emission sector groups based on the "SNAP" categorization (see sect. 2.4.5) is used. Thereby, the neural network is informed about the horizontal (EDGARv5) and vertical (SNAP) distribution of emissions. The SNAP sectors used here are "1" (public power, cogeneration, and district heating plants), "3" (industrial combustion), "4" (production processes), and "surface emissions", i.e. road traffic, agricultural emissions, etc. EDGARv5 refers to the year 2015, and might be outdated in some use cases, e.g. in geographic regions with rapid economic development (an example of this is briefly discussed in sect. 4.3.3). However, as shown for the years 2019 and 2022 in sect. 4.3.1 and 4.3.2, the use of emission data from 2015 seems to be adequate for predictions in central Europe.

4.1.2 Predictions on arbitrary vertical grids

NitroNet can predict NO_2 profiles on vertical output grids of the user's choice (as opposed to a fixed vertical output grid determined prior to training). The WRF-2019 dataset, on which NitroNet is trained, is resolved on 43 fixed vertical pressure levels, expressed in the terrain-following vertical η-coordinate (see the definition in sect. 2.4.3). When expressed in "meters above ground", the layer centers of the simulation output are slightly different for each model grid cell. A histogram of the layer centers of WRF-2019 can be found in Fig. 4.3.

Although it may seem intuitive to use a neural network with multiple output neurons (e.g. 10 neurons, which compute the NO_2 concentration at 10 different altitudes), a more efficient and flexible design can be achieved using only a single neuron. As shown in Fig. 4.3, the WRF-2019 layer centers almost cover the entire troposphere. Therefore, "altitude

input variable name	data source	note
NO_2 VCD (tropospheric)	TROPOMI	v.2.04.00
O_3 VCD (total)	TROPOMI	v.2.04.00
tropospheric air mass factor	TROPOMI	
tropospheric averaging kernels	TROPOMI	9 lowest TM5-MP layers
cloud radiance fraction	TROPOMI	
cloud pressure	TROPOMI	
aerosol absorbing index	TROPOMI	
surface albedo	TROPOMI	
surface pressure	TROPOMI	
sun geometry (zenith and azimuth angle)	TROPOMI	
viewing geometry (zenith and azimuth angle)	TROPOMI	
surface classification	TROPOMI	ternary mask (urban/cropland/forest)
NO_2 VCD influx	TROP. + ERA5	
planetary boundary layer height (PBLH)	ERA5	ERA5 variable ID: 159
planetary boundary layer dissipation	ERA5	ERA5 variable ID: 145
surface temperature	ERA5	ERA5 variable ID: 167
vertical velocity (profile)	ERA5	ERA5 variable ID: 135
wind speed (profile)	ERA5	total horizontal wind speed, i.e. $\sqrt{u^2+v^2}$
NO_x emissions (total)	EDGARv5	
NO_x emissions (SNAP 1)	EDGARv5	
NO_x emissions (SNAP 3)	EDGARv5	
NO_x emissions (SNAP 4)	EDGARv5	
NO_x emissions (surface sources)	EDGARv5	
day	—	binary mask (weekday/weekend)
altitude	user choice	see sect. 4.1.2

Table 4.2: NitroNet's input variables. The surface classification masks include land-use data from the USGS Global Land Cover Characteristics Database, see Eskes et al. (2022).

above ground" can simply be used as an additional neural network input variable. A single training instance then consists of a feature vector $\boldsymbol{x} = (x_1, x_2, ..., x_N)$ and a single altitude z, which can be written as the augmented feature vector $[\boldsymbol{x}\ z] = (x_1, x_2, ..., x_N, z)$. The corresponding training target y is the NO_2 concentration at altitude z. After training, the neural network is operated in a similar manner. In order to make predictions with NitroNet, the user must provide the required input variables (stored in a feature matrix $\mathbf{X}$) and a vector $\boldsymbol{z} = (z_1, z_2, ...z_H)$ with H entries, defining the vertical grid in "meters above ground" at which NitroNet is queried. The choice of $\boldsymbol{z}$ depends on the intended use case. For example, studies on surface level air pollution might use $\boldsymbol{z} = (1, 1.5, 2)$, which requests NitroNet to return NO_2 concentrations at 1 m, 1.5 m, and 2 m, respectively. The caveats of such low-altitude queries are discussed in more detail in sect. 4.2.10. In order to obtain full tropospheric NO_2 profiles, the user might choose a uniform sampling from $z_1 = 0$ to $z_H = 13000$ m (or some other

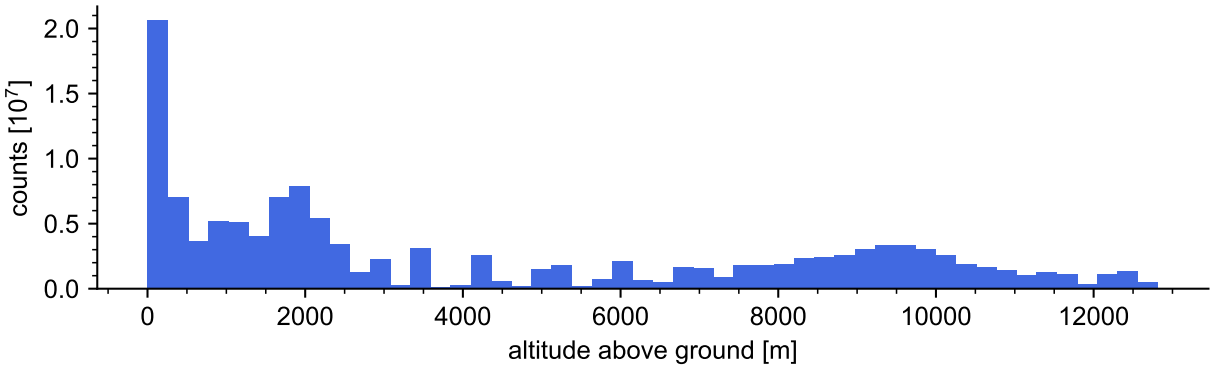

Figure 4.3: Histogram of the layer centers in WRF-2019 in meters above the ground. Drawn based on 300000 random samples from WRF-2019.

altitude, estimating the tropopause height), instead. NitroNet then processes the input data as depicted in Fig. 4.1c-d: An augmented feature matrix [$\mathbf{X}$, $\boldsymbol{z}$] is constructed, by repeating each row (representing one instance each) of the feature matrix $\mathbf{X}$ H times. Then, the altitude vector $\boldsymbol{z}$ is repeated as many times as required (= the number of rows in the original feature matrix) and appended horizontally to $\mathbf{X}$. The instances in the resulting augmented feature matrix can be used as input to the neural network, whereby a single forward pass returns the NO_2 concentration at a specified height z_i with $1 \leq i \leq H$.

By suitable choice of $\boldsymbol{z}$, the computational load of NitroNet can be considerably reduced. Moreover, the spacing of the vertical output grid can be chosen very small (e.g. 1 m), allowing to resolve fine-scale vertical NO_2 gradients (with caveats discussed in further detail in sect. 4.2.10). Additionally, the described design allows for a convenient memory-saving trick during the training of NitroNet. Depending on the size of the neural network and the training set, the video memory of consumer-grade GPUs can become a bottleneck. In that case, multiple neural networks can be trained sequentially, each in a different vertical range. Then, at runtime, their predictions are merged to yield full tropospheric profiles. This was done historically in the training of NitroNet, which internally consists of two sub-networks (one for the range from the surface to 8000 m, and the other for the range from 8000 m to the tropopause). Meanwhile the amount of accessible video RAM on NitroNet's training machine ("raven", see `https://docs.mpcdf.mpg.de/doc/index.html`) and the I/O routines of the PyTorch package have improved, making the described procedure obsolete. In future training runs of NitroNet, the hyperparameter optimization described in sect. 4.2.6 should be repeated in order to maintain the model's overall predictive capacity with only a single neural network (and correspondingly fewer trainable parameters).

4.2 Training of NitroNet's neural network

NitroNet's neural network is trained as explained in sect. 2.5.3. All required input data as listed in Table 4.2 are interpolated bilinearly to the spatio-temporal domain of WRF-2019. The resulting dataset is partitioned into a training set (80 %), a validation set (15 %), and a test set (5 %) by random sampling. The aspects described in the two previous sects. 4.1.1 and 4.1.2 represent principal design choices independent of the data partitioning. The aspects described in the following require distinction between training, validation, and test data.

4.2.1 Data transformations

NitroNet deploys data transformations, as described in sect. 2.5.4. The aim of this procedure is to bring the features (input variables) and the training targets to a mutual scale (approximately in a numerical range between 0 and 1) and normalize features with strongly skewed or long-tailed distributions. The corresponding bijective transformations are determined on the training set. Upon transformation, the variables become unitless. The type of data transformation applied to each variable depends on its distribution:

- Variables with long-tailed distributions are transformed logarithmically, e.g. the NO_2 concentrations shown in Fig. 4.4a. Although the transformed NO_2 concentration's distribution rather resembles a Gaussian mixture, it is overall much closer to a Gaussian than the original distribution.

- Some variables have more irregular distributions, e.g. the NO_x emissions from SNAP sector 3 shown in Fig. 4.4b. This variable spans a large range of values, including many cases where it equates to zero. Here, the *quantile transformation* is applied (see Pedregosa et al., 2011), which may result in considerable gaps (see the histogram of the transformed variable in Fig. 4.4b below values of 0.30). However, these pose no problem in the training process of the neural network.

- Most other variables are simply normalized, i.e. they are re-scaled approximately to the interval $[0, 1]$, while their histogram shape remains unchanged. This applies e.g. to the total wind speeds shown in Fig. 4.4c and the NO_2 VCDs shown in Fig. 4.4d. Slight exceedances of the target interval $[0, 1]$ (e.g. in the form of negative values) pose no problem.

- Categorical features (e.g. land surface classes) are not transformed.

For the technical details and the transformation equations, refer to Appendix C.1.

Figure 4.5 shows the transformation of the "altitude" variable from WRF-2019. Note that Fig. 4.5a and the histogram shown in Fig. 4.3 only differ by choice of the bin size. The altitude

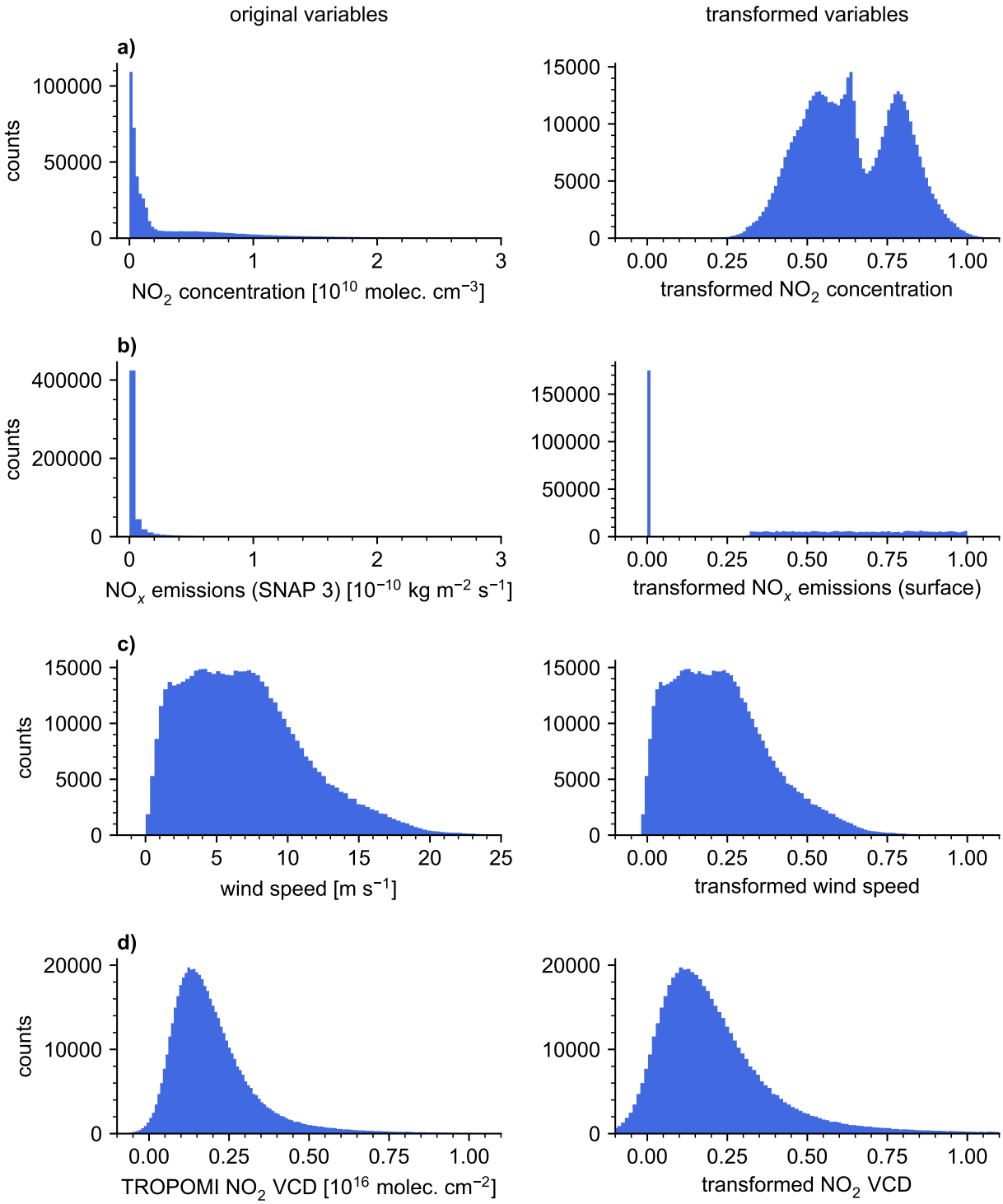

Figure 4.4: Data transformations of exemplary input features and the training targets of NitroNet. The left panel shows histograms of the original variables. The right panel shows histograms of the transformed variables. Due to the large size of the training set, the histograms were generated based on 300000 random samples of each variable. **a)** The NO_2 concentrations from WRF-2019 (i.e. the training targets). **b)** The NO_x emissions from SNAP sector 3. **c)** The total wind speeds from ERA5 (at the pressure levels given in sect. 4.1.1). **d)** The tropospheric NO_2 VCDs from TROPOMI.

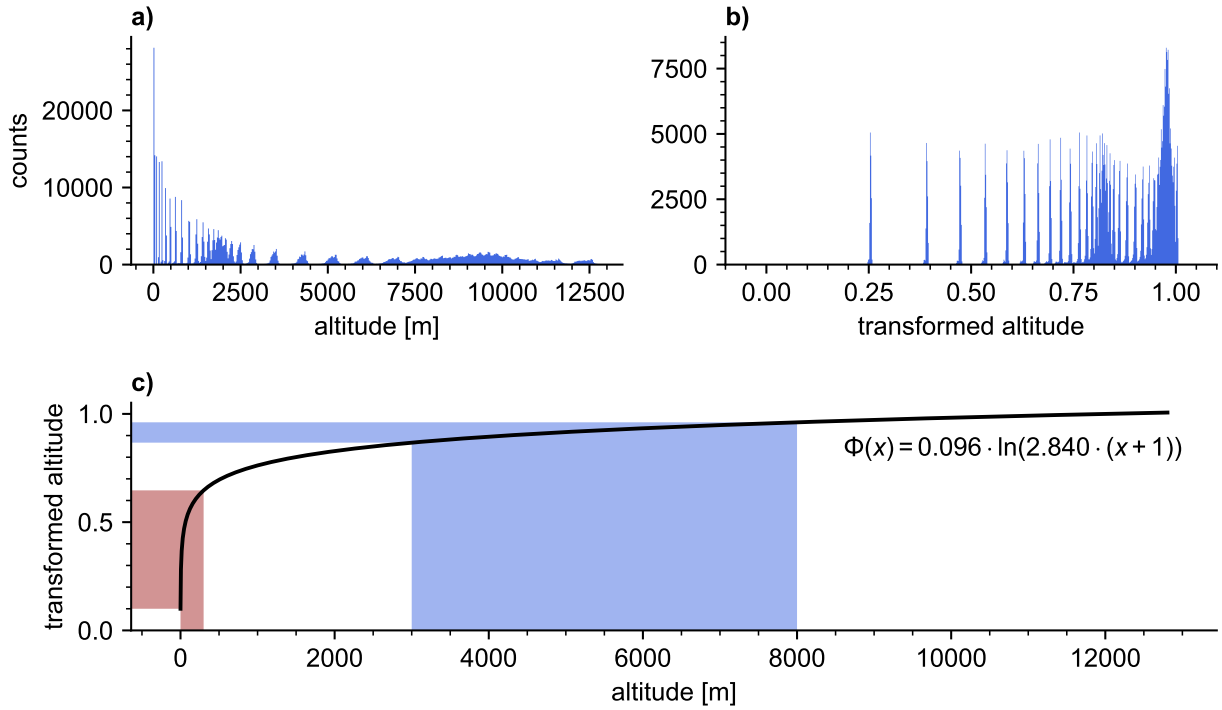

Figure 4.5: Data transformation of the "altitude" variable. a) The original altitude variable from WRF-2019. **b)** The transformed altitude variable. Due to the large size of the training set, the histograms were generated based on 300000 random samples. **c)** Plot of the corresponding data transformation function Φ, written out in the top right. For the red and blue patches, refer to the text description in sect. 4.2.1.

variable is treated with a logarithmic transformation as depicted in Fig. 4.5c. The histograms shown in Fig. 4.5a-b may give the impression, that this results in a mere redistribution of the variable's values across the interval $[0, 1]$ with no obvious benefit. However, consider the following: NO_2 profiles show steep gradients at the surface and at the top of the PBL. In the free troposphere, however, the NO_2 concentrations are usually considerably lower with less vertical dependence. The original altitude variable ranges approximately from $0 - 13000$ m. As represented by the blue and red patches in Fig. 4.5c, approximately 38 % of this range are occupied for the representation of altitudes between $3000 - 8000$ m (belonging to the free troposphere), while only 2 % are reserved for altitudes between $0 - 300$ m. With the transformed altitude variable, this relationship is inverted: The range from $3000 - 8000$ m is mapped to the interval $[0.86, 0.96]$, corresponding to 10 % of the target interval $[0, 1]$, while the range from $0 - 300$ m is mapped to the interval $[0.10, 0.65]$, corresponding to 55 %. In other words, the transformation stretches the relevant portions of the variable's domain, relative to its less relevant portions.

NitroNet applies the data transformations to the input features at runtime. The neural network predicts transformed NO_2 concentration, which are transformed back to their original

form via the inverse transformation. The entire process is automated and requires no attention on behalf of the user.

4.2.2 Data curation

As discussed in Chapter 3, the training data obtained from WRF-Chem still suffer from significant systematic errors in specific subregions of the model domain (see e.g. the strong overestimations of NO_2 in western Germany in Fig. 3.16). If the NO_2 profiles in these model regions can be identified as erroneous via comparison to reference data (e.g. TROPOMI observations), it is reasonable to dismiss them from the training set – otherwise, the neural network would learn to reproduce them. This procedure is referred to as *data curation* (or *data wrangling*, *data filtering*), and it is applied to the training, the validation, and the test set (unless mentioned otherwise, as for example in sect. 4.2.8). For that purpose, two thresholds $\Delta_{\mathrm{VCD}} > 0$ and $\Delta_{\mathrm{PBLH}} > 0$ are defined. Then, all instances for which either

$$\left| \frac{V_{\text{WRF-2019}} - V_{\text{TROPOMI}}}{V_{\text{TROPOMI}}} \right| > \Delta_{\mathrm{VCD}} \tag{4.3}$$

or

$$\left| \frac{\mathrm{PBLH}_{\text{WRF-2019}} - \mathrm{PBLH}_{\text{ERA5}}}{\mathrm{PBLH}_{\text{ERA5}}} \right| > \Delta_{\mathrm{PBLH}} \tag{4.4}$$

or

$$f_{\mathrm{QA}} \leq 0.75 \tag{4.5}$$

where:

V : the tropospheric NO_2 VCD from either WRF-2019 or TROPOMI

PBLH : the planetary boundary layer height from either WRF-2019 or ERA5

f_{QA} : qa-value of the colocated TROPOMI observation

are filtered out. The subscript t indicating tropospheric NO_2 VCDs is omitted for brevity. Altogether, an NO_2 profile in the WRF-2019 dataset set only qualifies for training/validation if the bias of its simulated tropospheric VCD (under consideration of TROPOMI's averaging kernels) is below Δ_{VCD}, and the bias of the simulated PBLH is below Δ_{PBLH}, with TROPOMI and ERA5 yielding the corresponding reference values. The rationale behind these two filters is the following:

1. The simulated and observed tropospheric NO_2 VCDs represent the total tropospheric NO_2 load. Any disagreement between the two indicates an over- or underestimation of

NO_2 in the WRF-Chem training data. If not removed, the neural network would be presented with training examples for which the (observed) NO_2 VCD does not even approximately represent the (simulated) vertical concentration integral, defying its original definition.

2. The PBLH is arguably the most important shape parameter for the prediction of NO_2 profiles, marking the transition between the planetary boundary layer and the free troposphere. ERA5 is an atmospheric reanalysis, i.e. a fusion of numerical simulations and a wide range of observational data, whereas the WRF-Chem simulation only assimilates rudimentary meteorological data via nudging. Therefore, the ERA5 PBLH values are deemed more trustworthy than those from the WRF-Chem simulation and all WRF-Chem data with strong disagreement to the ERA5 PBLH are dismissed.

A relative VCD filter is useful here, because it ensures a consistent linear relationship between the NO_2 VCDs and the target NO_2 profiles regardless of magnitude. On the other hand, relative filters lead to the removal of many training examples with low NO_2 VCDs (and shallow PBLHs), where even small absolute errors can correspond to large relative errors. Furthermore, the chosen filter criteria remove negative NO_2 VCDs (which can occur due to retrieval noise) entirely. NitroNet currently only deploys relative data filters. The benefits of absolute filters might be explored in future revisions.

Obviously, lower threshold values Δ_{VCD} and Δ_{PBLH} result in higher quality data, but fewer overall training examples. This poses a trade-off problem between data quality and quantity, which is addressed by including Δ_{VCD} and Δ_{PBLH} in the hyperparameter optimization discussed in sect. 4.2.6. Thereby, optimal values of $\Delta_{\mathrm{VCD}} = 0.2$ and $\Delta_{\mathrm{VCD}} = 0.1$ were determined. In theory, additional filter criteria could be implemented (e.g. based on the agreement between WRF-2019 surface NO_2 concentrations and AirBase measurements). However, this is infeasible here because Δ_{VCD} and Δ_{PBLH} already reduce the training data from ~ 1800000 NO_2 profiles (after colocating with the TROPOMI overpass and applying the quality filter $f_{\mathrm{QA}} > 0.75$) to just ~ 130000 (~ 7 % of the original training data). Figure 4.6 shows how the data curation procedure affects the training set across the spatial domain. As depicted in Fig. 4.6b, the aforementioned filter criteria are strict enough to remove entire geographic regions from the training set (e.g. large parts of Belgium or the Baltic Sea, where the NO_2 VCDs are also typically very small). Note, that this map shows the monthly mean of the TROPOMI NO_2 VCDs, ignoring any missing data along the temporal axis. In other words, the gaps occurring in Fig. 4.6b-c correspond to locations, in which not a single instance remained after filtering. Overall, depending on the location, the fraction of remaining instances is between 0 and 0.5.

An overview of the NO_2 VCDs and PBLH values in the training set before and after the

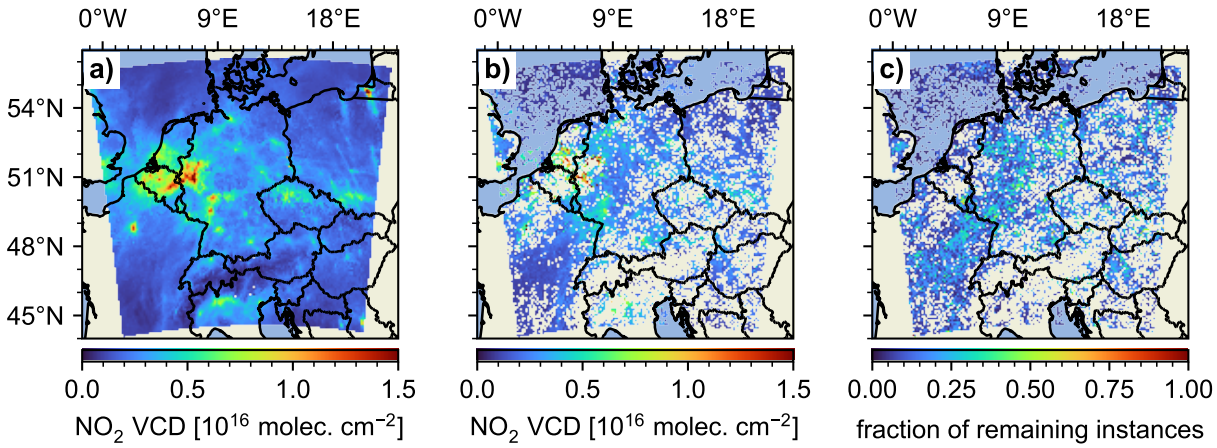

Figure 4.6: Overview of the remaining training instances before and after filtering. a), b) The monthly-mean tropospheric NO_2 VCDs from TROPOMI before and after filtering. **c)** The remaining fraction of instances after filtering.

filtering is given in Fig. 4.7. The figures were drawn based on 10000 samples from the training set, so that the right-side histograms remain intercomparable. As shown, the distributions of the variables in the filtered training set can differ significantly from those of the unfiltered reference variables. This means, that the training set might be skewed towards certain physical scenarios. For example, the filtered WRF-2019 training set appears to over-represent cases with higher PBLHs, lower NO_2 VCDs, and lower NO_2 surface concentrations, compared to the original TROPOMI and ERA5 reference distributions. This does not necessarily pose a problem; a neural network trained on a limited feature domain may still extract data relationships of general validity (e.g. that the NO_2 VCDs relate to the NO_2 profiles' amplitudes), and thereby generalize to make reasonable predictions on unseen input data. The true influence of these effects cannot be quantified until the neural network has been trained. The issue is therefore put aside for now, and addressed later in sect. 4.2.8.

4.2.3 Winsorization for out-of-distribution instances

Neural networks may struggle with out-of-distribution (OOD) instances, i.e. input data which lie far outside the joint distribution of the features in the training set. OOD instances can have a detrimental impact on a neural network's performance, even if the neural network's sensitivity to the OOD feature was found to be low in the in-distribution case. NitroNet aims to reduce the impact of OOD instances by implementation of a variant of the *winsorization* method (see e.g. Ruppert, 2014) at prediction time. Winsorization usually refers to the procedure of removing outliers from 1-dimensional datasets by computing a lower and upper quantile (e.g. the $5\ \% - 95\ \%$ quantile) and clipping all values above and below. For

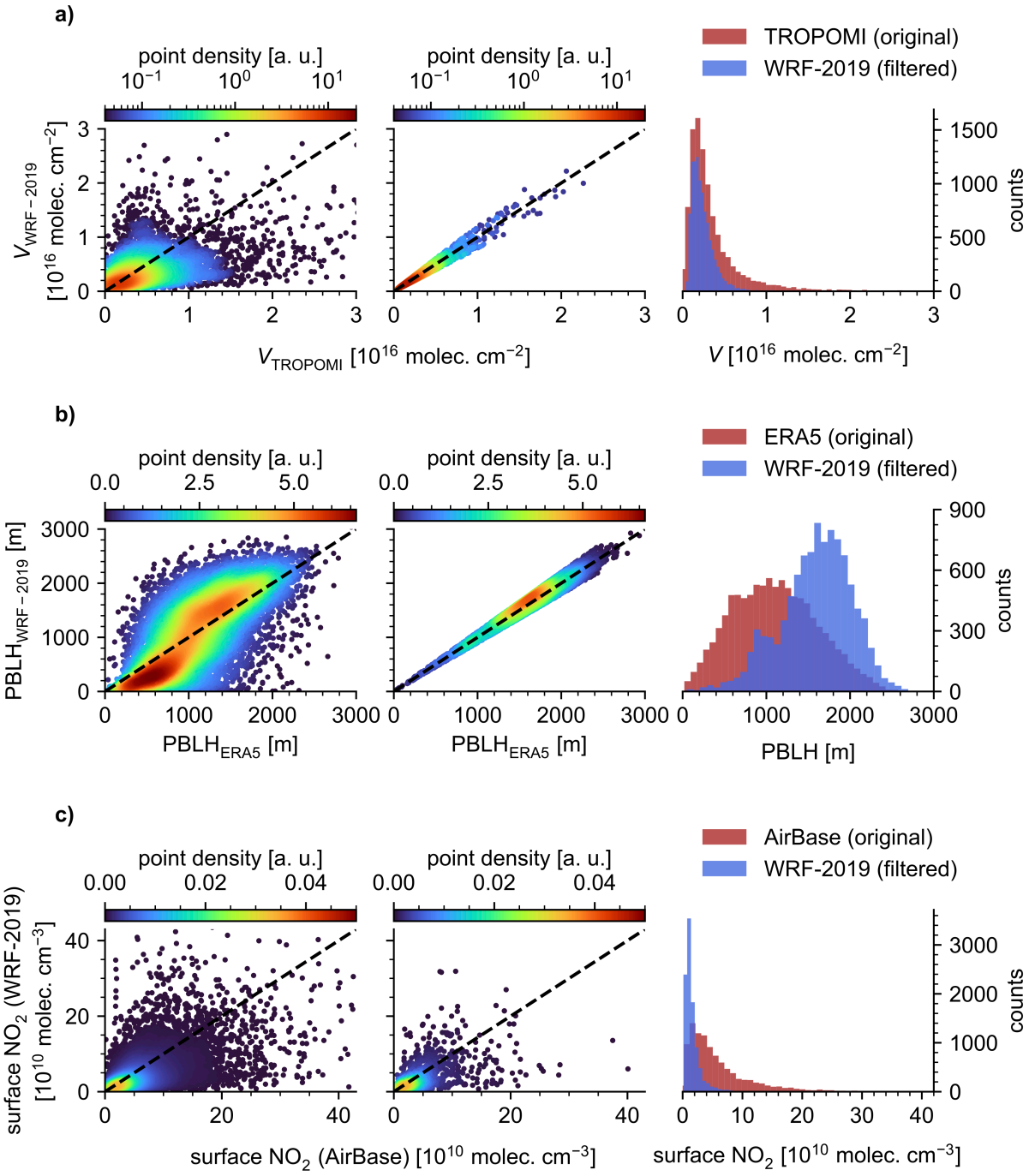

Figure 4.7: Comparison of WRF-2019 and reference data from TROPOMI and ERA5 before and after data filtering. Shown here are scatter plots of **a)** the tropospheric NO_2 VCD, **b)** the PBLH, and **c)** the NO_2 surface concentration. The left panel shows the unfiltered WRF-2019 dataset. The middle panel refers to the dataset after filtering. The right panel shows histograms of the original reference values (TROPOMI NO_2 VCDs, ERA5 PBLHs, and AirBase NO_2 concentrations) and the corresponding WRF-2019 values after filtering. The Mo-CL bias of the surface concentrations was accounted for, following sect. 2.3.6. All figures were drawn based on 10000 samples from the training set.

neural networks the approach can be modified by estimating the marginal probability density distributions $p(x_i)$ of the features x_i with kernel density estimation (KDE, see Parzen, 1962) on the training set. Instance entries are considered OOD if $p(x_i) < \alpha$, in which case they are replaced with a sample from $p(x_i)$. NitroNet uses a threshold value of $\alpha = 0.15$. The NO_2 VCD and categorical input variables are exempt from the winsorization procedure.

4.2.4 Prediction of Mo-CL bias correction factors

The widely used molybdenum-based chemiluminescence method (Mo-CL) for NO_2 in situ measurements was discussed in sect. 2.3.6. The Mo-CL bias is an instrument effect, and does not occur in the WRF-Chem model data used to train NitroNet. Although this is favourable, it raises the issue that NitroNet's predictions at the surface cannot be directly validated against the bias-contaminated surface in situ measurements. As explained in sect. 2.3.6, there exists a correction method for the Mo-CL bias in the form of multiplicative correction factors F. NitroNet was trained to predict F alongside the NO_2 concentrations, which is achieved by instantiating a copy of NitroNet's neural network but training it on F targets instead of NO_2 concentration targets. From hereon, the two neural networks of NitroNet are referred to as the "main network" and the "F-network". Because predictions of the Mo-CL bias are only required at the surface, the F-network is only trained on data from the lowest layer of the WRF-2019 dataset. The required F targets are computed according to eq. (2.79) using modelled surface VMRs of NO_2, PAN, and HNO_3. At runtime, the two neural networks are used one after another, yielding a combined prediction of the NO_2 concentration and the Mo-CL bias correction factor F. When comparing NitroNet predictions to Mo-CL in situ measurements, the biased measurements are divided by F.

4.2.5 Training process and loss curves

Now a detailed explanation of NitroNet's training routine is given. First, an artificial feed-forward neural network is instantiated. Its initial weights and biases are randomly sampled using a variant of the *Kaiming uniform distribution* (also referred to as *He initialization*, see He et al., 2015). Then, the neural network iteratively makes predictions on training batches, and its trainable parameters are optimized via backpropagation. After each full iteration over the training set (one epoch), the neural network's validation loss (i.e. the loss on the validation set) is computed as well. Whenever the validation loss reaches a new global minimum, the neural network is saved to disk. Note, that the batch size of 2048 listed in Table 4.1 refers to 2048 NO_2 concentration targets, not 2048 full NO_2 profiles. Correspondingly, one batch consists of 2048 surface-level F targets in the training of the F-network.

While traversing the loss landscape in search of a global loss minimum, training can stall

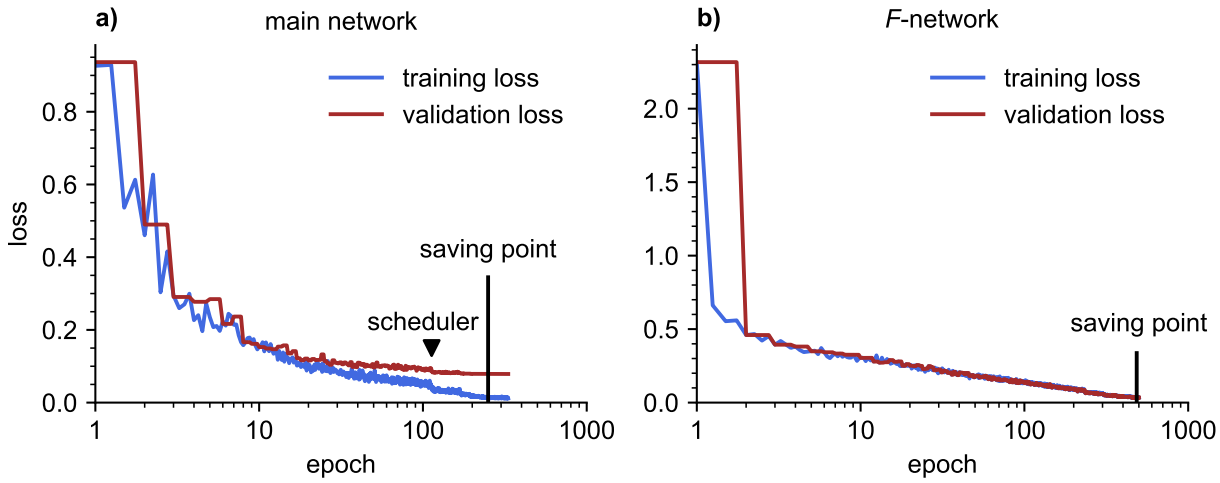

Figure 4.8: Loss curves of NitroNet's training. **a)** Loss curve of the main network, trained on NO_2 concentration targets. The "scheduler" annotation (epoch 113) marks the first activation of the learning rate scheduler. Training is terminated by early stopping after 331 epochs. The network is last saved at epoch 251. **b)** Loss curve of the *F*-network, trained on *F* targets. The network is last saved at the end of training, i.e. after 500 epochs.

if the learning rate is chosen too large, causing the optimizer to overshoot nearby minima. In order to address this issue, a *learning rate scheduler* is used, which halves the learning rate whenever the validation loss does not decrease over the span of 20 epochs. Training ends after 500 epochs, or when the learning rate drops below $7 \cdot 10^{-5}$ (*early stopping*). Early stopping helps to reduce the computational burden of the hyperparameter optimization (see sect. 4.2.6), which includes many training runs that stall after a few epochs, making a full iteration over 500 epochs futile.

Figure 4.8a shows the loss curves (i.e. training and validation loss as functions of the training step), of NitroNet's main network. The figure reveals steady training progress over the first $\sim$ 100 epochs. After approx. 20 epochs, the training and validation loss begin to diverge. This poses no fundamental issue, as long as the validation loss keeps decreasing. After 113 epochs, the training and validation loss suddenly drop by a small amount, while the validation loss becomes considerably less noisy. This is due to the learning rate scheduler being triggered after a period of stagnation. Early stopping occurs at epoch 331, at which point the scheduler has been triggered multiple times without any improvements to the validation loss. Its minimum occurs at epoch 251, where the final trained version of the network is saved to disk.

Figure 4.8b shows the loss curves of NitroNet's *F*-network. In contrast to the training of the main network, neither the learning rate scheduler nor early stopping are triggered, meaning that the training spans 500 epochs over which both training and validation loss

continuously decrease. As mentioned in sect. 4.2.4, only surface data is used when training NitroNet's F-network. This makes for a simpler training objective, because a far smaller altitude range must be captured. Note, that the loss values of Fig. 4.8a and 4.8b are not expected to be similar, because they represent errors of entirely different physical quantities. The downwards trend of the loss curves in Figure 4.8b, indicates, that even lower loss values could have been obtained with longer training. However, the F-network's relative prediction error on the validation set amounts to $\sim$ 6 % at the end of training, which is considered sufficient.

4.2.6 Hyperparameter optimization

The hyperparameters of NitroNet, as listed in Table 4.1, were obtained from an extensive hyperparameter study (see also sect. 2.5.5). In this study over 100 variants of the neural network were trained as described in the previous section. Then, the network variants were ranked by their *mean absolute percentage error* (MAPE) on the validation set. The MAPE is defined as

$$\text{MAPE} = \sum_{i=1}^{N} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{4.6}$$

where:

$\hat{y}_i$: the neural network predictions

y_i : the ground truth

The parameter combinations were determined by random sampling from a pre-defined parameter space ("randoms search", see Bergstra and Bengio, 2012), described in Table 4.3. The samples were obtained from uniform distributions over the specified intervals, although the interval boundaries were at some points adjusted. For example, the number of hidden layers was originally sampled from between 3 and 7, but values of lower than 6 were found to lead to a stagnant training loss early on. Therefore, the sampling range was changed to range from 6 to 10.

The results of the hyperparameter optimization are summarized in Fig. 4.9. The figure shows the entire ensemble of neural network configurations in a parallel coordinate view. Each vertical axis (e.g. "hidden layers", "neurons per layer", etc.) corresponds to one hyperparameter, and each neural network configuration is represented by a contiguous line crossing the vertical axes corresponding to its hyperparameter selection. The rightmost axis denotes the validation MAPE achieved by each neural network. This way, patterns in the relation between the hyperparameter selection and network performance can be identified. For ex-

hyperparameter name	sampling space	optimal value
hidden layers	$3 - 10$	8
neurons per layer	$200 - 400$	326
activation function	ReLU, PReLU, CELU[(1)], GELU[(2)], SELU[(3)]	PReLU
loss function	MSE, L_1, smooth L_1[(4)], RMSLE[(5)]	L_1
batch size	$2^7 - 2^{12}$	2048
optimizer	SGD[(6)], NAdam, AdamW	NAdam
initial learning rate	$5 \cdot 10^{-5} - 10^{-3}$	$3.42 \cdot 10^{-4}$
batch normalization	true, false	false
dropout probability	$0 - 0.15$	0
Δ_{VCD}	$0 - 0.7$	0.2
Δ_{PBLH}	$0 - 0.7$	0.1

Table 4.3: Parameter space used in NitroNet's hyperparameter optimization. For a combined reference of the terms used here, see sects. 2.5, 4.2, the references therein, and the table footnotes below.

(1) *continuously differentiable exponential linear unit*, see Barron (2017)

(2) *gaussian error linear unit*, see Hendrycks and Gimpel (2023)

(3) *scaled exponential linear unit*, see Klambauer et al. (2017)

(4) see Girshick (2015)

(5) *root mean squared logarithmic error*, see e.g. Jadon et al. (2022)

(6) all runs with SGD diverged, hence why they do not occur in Fig. 4.9

ample, the activation function "SELU" resulted in much higher average MAPE values than "GELU". Interestingly, the neural network does not benefit from popular regularization methods such as batch normalization or dropout. The supposedly best configuration, described in Table 4.1 and used for NitroNet's main network from hereon, is drawn as a thick line and achieves a MAPE of 11 %. Overall, the MAPE values can reach up to 30 %, meaning that NitroNet's relative prediction errors can vary by up to a factor 3, depending on the choice of hyperparameters. This emphasizes the importance of the hyperparameter optimization.

4.2.7 Evaluation on the test set

During training, the network's trainable parameters were optimized on the training set and validated on the (independent) validation set. The main purpose of this procedure is to identify overfitting, if it occurs, and to obtain a realistic assessment of how well the neural network would perform on previously unseen data. Notice, however, that the hyperparameter selection was influenced by the training data *and* the validation data. Therefore, the resulting neural network must still be tested on another completely independent dataset. This is the

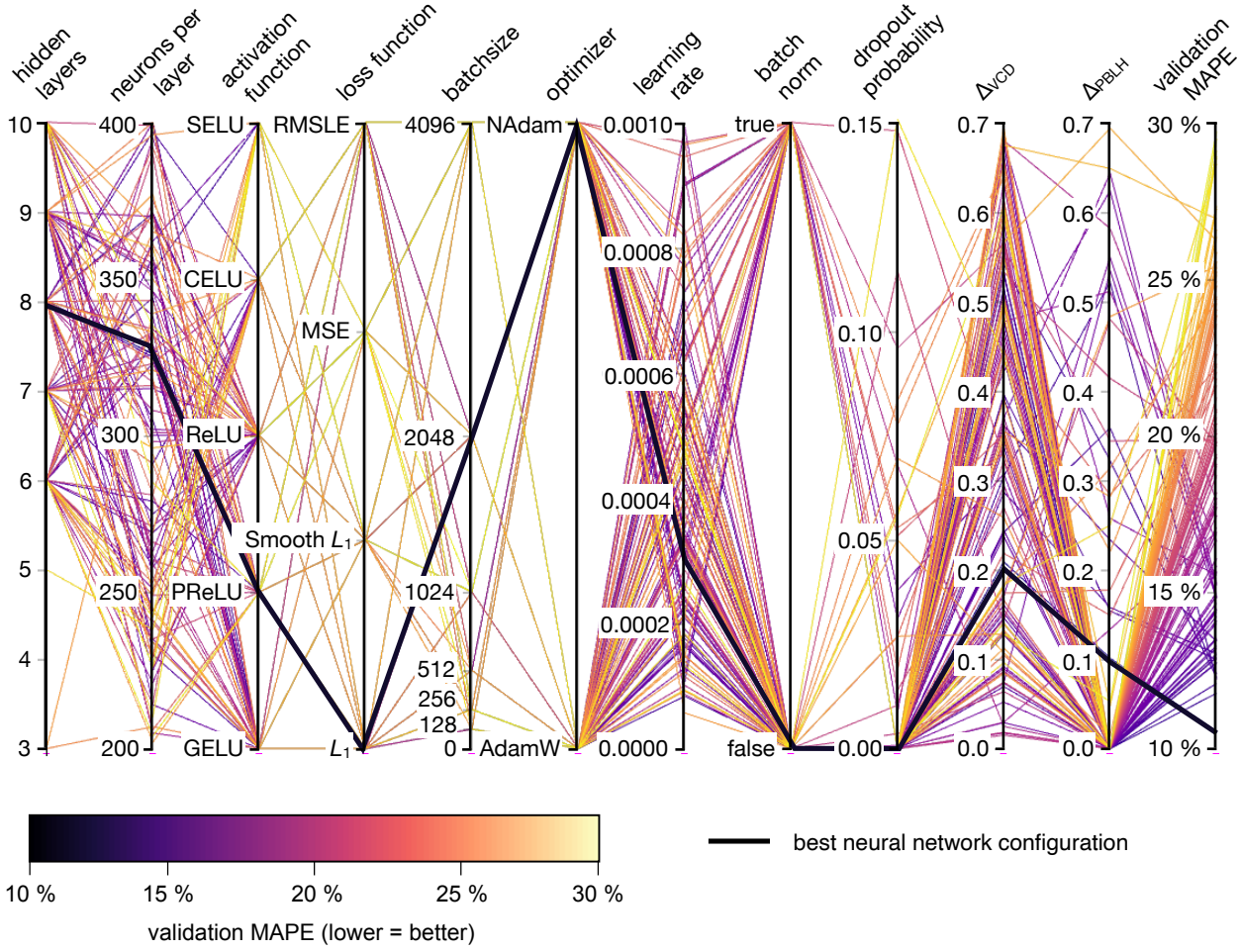

Figure 4.9: Results of NitroNet's hyperparameter optimization study. Each vertical axis corresponds to a single hyperparameter. Each neural network configuration is represented by a contiguous line, crossing the vertical axes corresponding to its hyperparameters. The rightmost vertical axis shows the validation MAPE. NitroNet's optimal network configuration is drawn as a thick line. In total, 122 random parameter combinations were tested.

purpose of the test set, mentioned in the beginning of sect. 4.2.

Table 4.4 gives an overview of different performance diagnostics NitroNet achieves on the test set, resolved at the individual vertical layers of WRF-2019. Although WRF-2019 has 43 vertical layers, some of them extend beyond the tropopause. Therefore, only the lowest 38 tropospheric layers are used here. For now, the data filters described in sect. 4.2.2 remain applied, following the notion, that the neural network should not be evaluated against data that were identified as possibly erroneous and would have been rejected from training. Table 4.4 can be broadly divided into three vertical regions:

- From the surface to ~ 3 km, representing mostly the planetary boundary layer. Here, NitroNet is practically unbiased, with relative mean biases of < 2 %. The RMSE is largest at the surface ($3.9 \cdot 10^9$ molec. cm^{-3}). The relative prediction error, as quantified

by the MAPE, is between 2 % and 13 %. The entries of the columns "slope" and "intercept" were determined by a least-squares linear fit for each vertical layer, and reveal close to ideal values (1 for the slope, and 0 for the intercept, denoted in units of 10^9 molec. cm^{-3}). The correlation coefficients are as high as $R \geq 0.96$.

- From $\sim 3 - 8$ km, representing a large portion of the free troposphere. In this vertical range, a significant falloff in NitroNet's performance is observed. Although the RMSE values are far smaller than at lower altitudes, so are the corresponding ground-truth NO_2 concentrations. This results in MAPE values of up to 206 %, and mean biases of up to +65 %. The slopes are between 0 and 1, and the correlation coefficients lie in the range of $0.15 - 0.83$.

- From ~ 8 km to the tropopause, representing the upper troposphere. Here, NitroNet's performance recovers, with MAPE values and absolute mean biases dropping below 66 % and 14 %, respectively. The slopes and correlation coefficients range from $0.6 - 0.8$ and $0.66 - 0.80$, respectively.

The interpretation of Table 4.4 is assisted by Fig. 4.10a. Shown here are four randomly selected NO_2 profiles from the test set and their corresponding NitroNet predictions. The shape and the amplitude of the profiles is reproduced well, particularly throughout the first few kilometers above ground. In the free troposphere small deviations occur, e.g. in the green profile between $2 - 6$ km altitude, as shown in the inset axis. However, these deviations can still contribute significantly to relative performance metrics, such as the MAPE and the mean bias. The third vertical range discussed above, from 8 km upwards, coincides with the sudden increase of the NO_2 concentration. This surplus NO_2 could originate from aircraft emissions or lightning, or from stratospheric downwelling (see sect. 2.2.6). Figure 4.10b shows a scatter plot of 10000 NO_2 targets, randomly sampled from the test set and drawn against the corresponding NitroNet predictions. Again, the largest relative prediction errors occur in low-pollution scenarios, for the reasons discussed above.

Altogether, the evaluation on the filtered test set demonstrates, that NitroNet is well capable of reproducing the NO_2 concentrations from WRF-2019 based on the supplied input data. However, the relative precision depends on altitude, with particularly large relative errors in the free troposphere from $\sim 3 - 8$ km above ground. Near the surface the average relative prediction errors are as low as 5 %.

WRF-2019 layer	altitude [m]	bias [%]	RMSE	MAPE [%]	slope	intercept	R
1	4	−1.2	3.9	4.5	1.0	0.0	0.98
2	20	−0.8	2.4	3.4	1.0	0.1	0.99
3	48	−0.3	1.7	2.9	1.0	0.0	0.99
4	93	−0.5	1.6	2.8	1.0	0.1	0.99
5	161	−0.4	1.6	2.7	1.0	0.1	0.99
6	251	−0.3	1.5	2.6	1.0	0.1	0.99
7	358	−0.4	1.4	2.6	1.0	0.1	0.99
8	486	−0.3	1.2	2.7	1.0	0.0	0.99
9	637	−0.2	1.1	2.8	1.0	0.0	0.99
10	812	−0.2	1.0	3.1	1.0	0.0	0.99
11	1023	−0.2	1.0	3.5	1.0	0.0	0.99
12	1234	−0.2	1.0	5.2	1.0	0.1	0.99
13	1413	−0.3	0.8	7.5	1.0	0.0	0.99
14	1572	−0.3	0.7	9.8	1.0	0.0	0.99
15	1715	−0.5	0.6	10.5	1.0	0.0	0.99
16	1832	−0.6	0.5	11.9	1.0	0.0	0.99
17	1927	−0.3	0.4	10.7	1.0	0.0	0.99
18	2047	−0.5	0.4	10.1	1.0	0.0	0.99
19	2212	−0.1	0.3	11.6	1.0	0.0	0.97
20	2446	−0.5	0.2	11.7	1.0	0.0	0.97
21	2823	+0.2	0.1	12.5	1.0	0.0	0.96
22	3435	+6.0	0.2	34.6	0.8	0.1	0.83
23	4232	+9.9	0.2	63.1	0.4	0.2	0.66
24	5091	+17.8	0.2	94.0	0.2	0.2	0.43
25	5965	+38.9	0.3	131.9	0.1	0.2	0.27
26	6826	+64.1	0.3	176.4	0.1	0.2	0.18
27	7532	+64.9	0.3	205.7	0.1	0.2	0.15
28	8049	+3.2	0.2	93.0	0.5	0.1	0.35
29	8470	−13.1	0.2	65.3	0.6	0.0	0.66
30	8822	−12.4	0.2	65.7	0.6	0.1	0.71
31	9125	−10.6	0.2	65.0	0.8	0.0	0.74
32	9384	−9.1	0.2	65.4	0.8	0.1	0.76
33	9619	−8.3	0.2	65.2	0.8	0.1	0.77
34	9874	−7.5	0.3	61.1	0.8	0.1	0.78
35	10211	−6.2	0.3	47.1	0.8	0.1	0.79
36	10699	−3.9	0.3	30.9	0.8	0.1	0.80
37	11371	−2.6	0.2	17.0	0.8	0.2	0.80
38	12277	−2.5	0.2	11.4	0.8	0.2	0.71

Table 4.4: NitroNet performance metrics on the filtered test set. Slope and intercept were obtained from a least-squares linear regression through the point cloud of each vertical layer. RMSE values and intercepts are given in units of 10^9 molec. cm^{-3}.

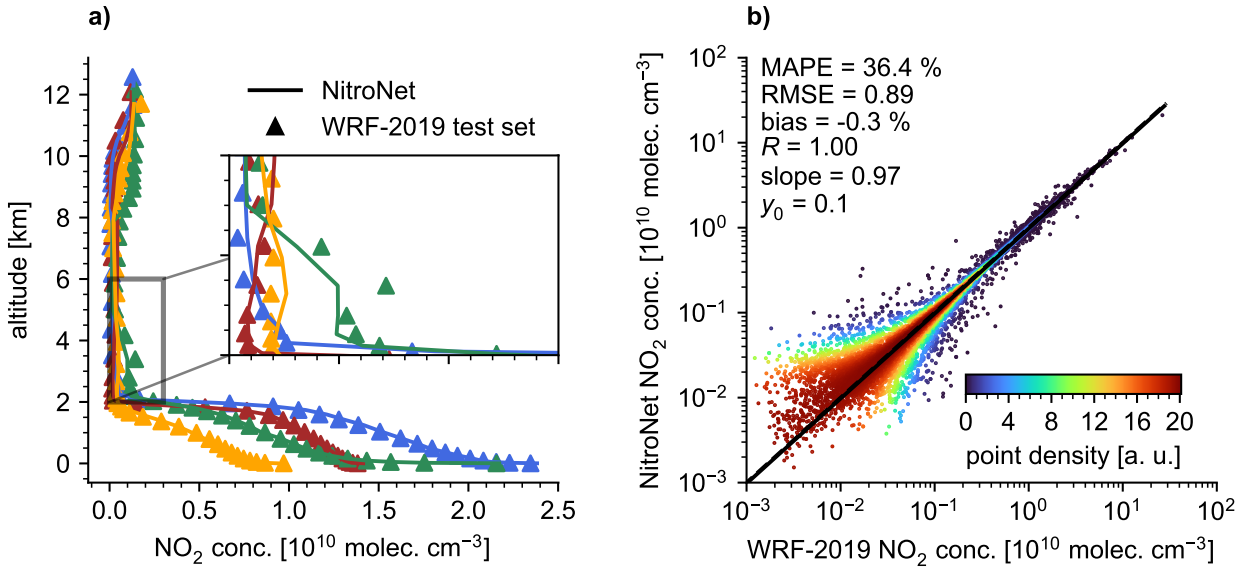

Figure 4.10: Exemplary NitroNet predictions on the filtered test set. a) Four randomly selected profiles from WRF-2019 (drawn as triangle scatter markers) with corresponding NitroNet predictions (drawn as solid lines). **b)** Scatter plot of the NO_2 targets in the filtered test set, regardless of altitude, drawn against their corresponding NitroNet predictions. RMSE and intercept (y_0) are given in units of 10^9 molec. cm^{-3}. The plot was drawn on the basis of 10000 random samples from the test set.

4.2.8 Negative impacts of data filtering, empirical bias correction, and evaluation on the unfiltered test set

So far, NitroNet was evaluated on the filtered test set, meaning that the data filters described in sect. 4.2.2 were applied. For the reasons discussed there it is fair to assume, that the filtering improves the *quality* of the data by removing possibly erroneous instances. However, selection of a subrange of data can also have a negative impact on the *variety* of the data. For example, the right-side panel of Fig. 4.7 indicates, that filtering based on the agreement to TROPOMI NO_2 VCDs and ERA5 PBLH values shifts the WRF-2019 dataset towards lower NO_2 VCDs, lower NO_2 surface concentrations, and higher PBLHs. In other words, the procedure may lead to the training data over/underrepresenting certain physical scenarios compared to the real world. In this section the associated consequences on the training of NitroNet are investigated.

In order to address the described issue, NitroNet is first evaluated on the unfiltered training set (note: *not* the test set). There, compared to the evaluation on any of the filtered datasets (training, validation, or test set), the neural network predictions shows slight biases (see Table 4.5). The prediction bias near the surface (-8 %) is comparably low, but reaches larger negative values of ~ -24 % near the top of the boundary layer, and even larger positive values of $+10$ % to $+50$ % in the lower-to-mid free troposphere. Overall, this empirical

WRF-2019 layer	altitude [m]	mean bias [%]	WRF-2019 layer	altitude [m]	mean bias [%]
1	4	−8.0	2	20	−8.9
3	48	−8.8	4	93	−8.5
5	161	−7.5	6	251	−6.6
7	358	−6.3	8	486	−6.3
9	637	−6.9	10	812	−8.2
11	1023	−11.4	12	1234	−12.9
13	1413	−17.3	14	1572	−21.1
15	1715	−24.2	16	1832	−24.3
17	1927	−23.0	18	2047	−21.1
19	2212	−17.0	20	2446	−13.2
21	2823	−9.8	22	3435	−1.6
23	4232	+10.8	24	5091	+22.4
25	5965	+39.5	26	6826	+55.9
27	7532	+48.4	28	8049	+1.6
29	8470	−7.5	30	8822	−6.4
31	9125	−5.2	32	9384	−4.7
33	9619	−5.7	34	9874	−7.5
35	10211	−9.1	36	10699	−8.2
37	11371	−6.5	38	12277	−5.5

Table 4.5: Look-up table for NitroNet's empirical bias correction. The mean biases given here were determined by evaluation on the full training set.

assessment indicates that the data filtering procedure leads to biased NitroNet predictions on unfiltered instances. Because NitroNet is intended to operate on unfiltered input data (with the exception of thresholding based on TROPOMI's qa-value) in practical use, this calls for the implementation of a bias correction scheme, here in the form of a look-up table. More precisely, for each layer given in Table 4.5, an altitude dependent bias correction factor $c(z)$ is defined, so that

$$\langle y(z) \rangle = c(z) \cdot \langle \hat{y}(z) \rangle \tag{4.7}$$

where:

$\langle \hat{y}(z) \rangle$: the average NitroNet prediction at altitude z

$\langle y(z) \rangle$: the average target value at altitude z

Following the definition in eq. (3.1), $c(z)$ can be computed from the mean relative bias of

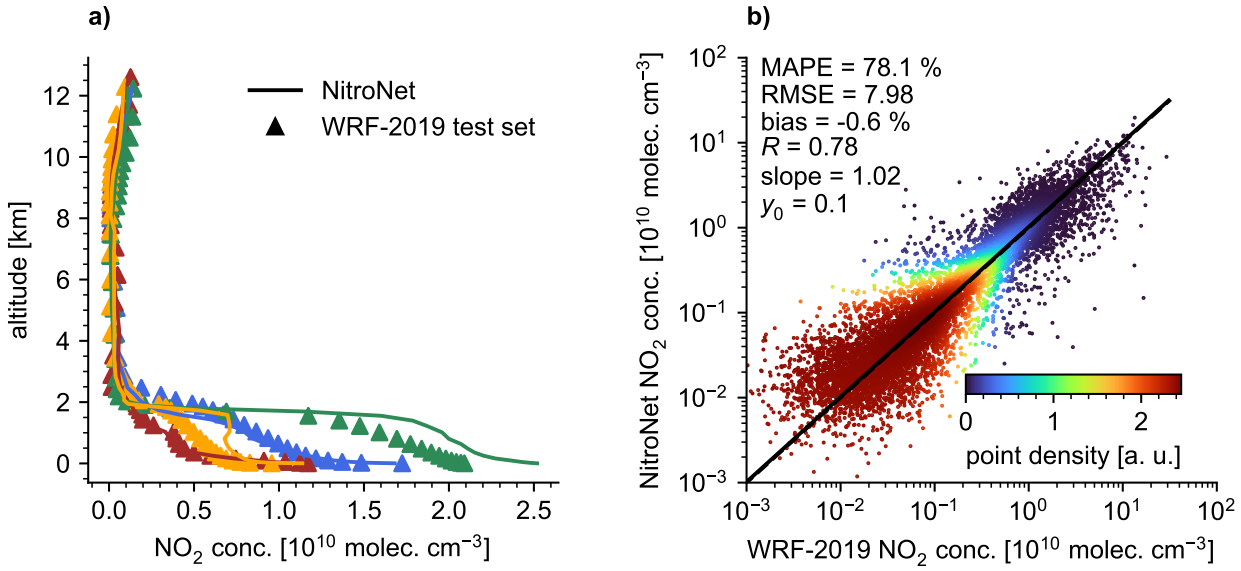

Figure 4.11: Exemplary NitroNet predictions on the full test set. Identical to Fig. 4.10, except that the full unfiltered test set and bias-corrected NitroNet predictions were used.

each vertical layer as

$$c(z) = \frac{1}{1 + \text{bias}(z)} \tag{4.8}$$

Thereby, Table 4.5 defines a look-up table by means of which the prediction biases resulting from the data filtering procedure can be compensated. $c(z)$ can be obtained for arbitrary altitudes by inter/extrapolation of Table 4.5. From hereon, all NitroNet predictions are bias-corrected, i.e. multiplied by the bias correction factors $c(z)$ corresponding to the queried altitude z.

Note, that $c(z)$ has been derived exclusively from the training set, and much like the neural network predictions themselves, must be validated on an independent dataset. In that pursuit, the evaluations of the previous sect. 4.2.7 are repeated, using the full test set and bias-corrected NitroNet predictions. Here, "full test set" means that the filter criteria from eqs. (4.3) and (4.4) were not applied. The qa-value criterion, defined in eq. (4.5) is applied throughout, following the recommendations of the TROPOMI PUM (see Eskes et al., 2022). The versions of Fig. 4.10 and Table 4.4 corresponding to this evaluation are shown as Fig. 4.11 and Table 4.6.

The main observation to be made here are the low prediction biases of below 1 % achieved in almost all vertical layers (with the exception of layers 26 – 31, where the mean prediction bias reaches up to +9 %). This means, that the bias correction works as intended, and generalizes well to previously unseen data. On the other hand, evaluation on the full test set results in significantly worse performance metrics. For example, the MAPE ranges from

WRF-2019 layer	altitude [m]	bias [%]	RMSE	MAPE [%]	slope	intercept	R
1	4	+0.0	18.6	40.5	0.8	3.5	0.71
2	20	+0.0	16.9	41.9	0.8	3.6	0.69
3	48	+0.0	15.8	42.0	0.8	2.7	0.67
4	93	+0.0	15.1	42.3	0.8	3.1	0.66
5	161	+0.0	14.4	42.4	0.8	2.8	0.66
6	251	+0.0	13.7	42.7	0.9	1.9	0.66
7	358	+0.0	13.0	43.1	0.9	2.1	0.66
8	487	+0.0	12.3	43.9	0.9	1.9	0.66
9	638	+0.0	11.4	45.2	0.9	1.7	0.66
10	812	+0.1	10.5	47.6	0.9	1.4	0.66
11	1025	+0.0	9.5	52.0	0.9	1.1	0.65
12	1236	+0.3	8.3	75.5	0.9	1.1	0.64
13	1415	+0.2	7.3	79.8	1.0	0.7	0.62
14	1575	+0.2	6.3	85.4	0.9	0.7	0.60
15	1718	−0.3	5.4	88.1	0.9	0.4	0.57
16	1835	−0.5	4.4	87.3	1.0	0.3	0.55
17	1930	−0.3	3.6	82.2	1.0	0.2	0.53
18	2050	−0.4	2.8	76.6	0.9	0.2	0.49
19	2216	+0.2	2.0	73.1	0.8	0.2	0.42
20	2450	+0.5	1.5	67.8	0.8	0.2	0.36
21	2826	+0.1	1.0	63.1	0.7	0.1	0.35
22	3436	+0.3	0.6	66.4	0.4	0.2	0.34
23	4231	+0.8	0.4	78.0	0.2	0.2	0.30
24	5087	+0.5	0.3	92.3	0.1	0.2	0.19
25	5957	+0.9	0.2	100.8	0.1	0.1	0.12
26	6815	+2.7	0.2	108.0	0.1	0.1	0.12
27	7519	+9.0	0.2	130.5	0.0	0.1	0.10
28	8035	−3.9	0.2	112.6	0.2	0.1	0.23
29	8455	−4.2	0.2	116.6	0.3	0.1	0.34
30	8806	−2.3	0.3	129.2	0.3	0.1	0.38
31	9109	−1.2	0.3	138.3	0.4	0.1	0.41
32	9369	−0.2	0.3	143.2	0.5	0.1	0.42
33	9603	+0.1	0.4	137.0	0.5	0.2	0.43
34	9858	−0.2	0.4	115.8	0.6	0.2	0.45
35	10196	−0.8	0.5	86.6	0.6	0.2	0.45
36	10685	−0.3	0.5	58.4	0.7	0.3	0.45
37	11359	+0.0	0.4	33.5	0.6	0.4	0.43
38	12271	+0.0	0.4	21.8	0.4	0.8	0.32

Table 4.6: NitroNet performance metrics on the full test set. Identical to Table 4.4, except that the full unfiltered test set and bias-corrected NitroNet predictions were used. RMSE values and intercepts are given in units of 10^9 molec. cm^{-3}.

approx. 40 % to to 145 %, although it only exceeds 100 % in the altitude range from approx. $6 - 10$ km. Likewise, the RMSE values are larger (up to $18.6 \cdot 10^9$ molec. cm^{-3}), and the correlation coefficients smaller (between 0.10 and 0.71) than on the filtered test set. This is also reflected in Fig. 4.11, showing significantly larger deviations between the NO_2 targets and the corresponding NitroNet predictions. Versions of Fig. 4.11 and Table 4.6 without application of the empirical bias correction are found in Fig. C.1 and Table C.2. However, these contain no significant additions to the combined results presented so far and are thus not further discussed.

Concluding remarks on the evaluation of NitroNet on the filtered and full test set

With the two presented evaluations of NitroNet, one on the filtered and one on the full test set, a few preliminary conclusions can be drawn:

- (on the filtered test set): There are no signs of significant overfitting. This means, that the amount and variety of training data is sufficient, and the neural network is not overcomplex.

- (on the filtered test set): The performance metrics such as the MAPE in Table 4.4 are satisfactory. This indicates, that NitroNet's input variables and network capacity are sufficient for the reconstruction of NO_2 profiles.

- (on the full test set): The reduced performance on the full test set has two possible explanations: Either NitroNet generalizes poorly from the filtered training set to the full test set, or the full test set is plagued by errors/inconsistencies (e.g. high tropospheric NO_2 VCDs from TROPOMI, paired with low-amplitude NO_2 profiles from WRF-Chem) which NitroNet does not reproduce as a result of being trained on filtered data. The two effects do not exclude each other, and cannot be separately quantified based on the presented data.

- (on the full test set): The empirical bias correction obtained from the full training set generalizes well to unseen data and may be used throughout the following sections of this thesis.

Note, that the evaluation on the test set only reveals, how well NitroNet can reproduce the NO_2 profiles from WRF-Chem. Any differences between the two are manifestations of the machine learning uncertainties discussed in sect. 2.5.7, whose individual contributions cannot be determined. However, WRF-Chem is not a perfect representation of the real world, and NitroNet was trained on a filtered subset of data that was shown to favour specific physical scenarios. Therefore, no claims can be made about NitroNet's performance with respect to

the real world. In fact, any of the supposed "errors" showing in Fig. 4.11 and Table 4.6 might represent cases where NitroNet disagrees with WRF-Chem, precisely because training on the filtered dataset – and thereby preventing NitroNet from adopting WRF-Chem's errors – had the intended effect. In consequence, NitroNet must be validated against observational data, in analogy to how the WRF-Chem simulation was evaluated in Chapter 3. The results of this validation study are presented in sect. 4.3. Before that, a feature relevance analysis (sect. 4.2.9) and an investigation of the NO_2 gradients in the lowest 100 m above ground (sect. 4.2.10) are conducted.

4.2.9 Feature relevance analysis

This section presents a feature relevance analysis, addressing the question of how much each input variable contributes to NitroNet's prediction quality based on the *Shapley scores*. The procedure follows sect. 2.5.6, including the two approaches to reducing the computational burden described there. This includes treating some input variables as groups. For example, the group "viewing geometry" mentioned in Table 4.2 consists of two variables, the viewing zenith angle and the azimuth angle. The same holds for vertically resolved groups, e.g. the "tropospheric averaging kernels" (9 individual variables, representing the averaging kernels at individual TM5-MP layers), and one-hot encoded variables such as the ternary surface classification. In addition, the five variables referring to NO_x emissions (total, SNAP 1, SNAP 3, SNAP 4, surface emissions) were grouped under the term "NO_x emission data".

Figure 4.12a shows the resulting feature relevances as a bar chart. As expected, the tropospheric NO_2 VCD is by far the most important input feature (relevance: 30.9 %), followed by the NO_x emission data (relevance: 8.9 %) and the PBLH (relevance: 6.9 %). A more detailed picture is given in Fig. 4.12b, where the feature relevances were computed individually for each vertical layer of WRF-2019, resulting in vertical feature relevance profiles (restricted to the lowest 2800 m above ground for easier readability). The feature relevances are normalized *per layer*, i.e. the sum over the entire coalition in each layer equates to 1. The sum of a feature's relevance along the vertical axis, however, is not normalized. The vertically resolved analysis reveals that the feature relevances vary significantly with altitude. The main findings corresponding to Fig. 4.12 can be summarized as follows:

1. The relevance of the tropospheric NO_2 VCD dominates in the lowest 1500 m of the troposphere, but decreases towards the top of the boundary layer. Within the boundary layer, the NO_2 VCD is the most important input variable.

2. The relevance of the PBLH increases between 1000 m and 2300 m, which corresponds to the typical PBLH values in the WRF-2019 dataset. This, too, is expected, seeing that the PBLH informs NitroNet about the approximate altitude at which a sudden

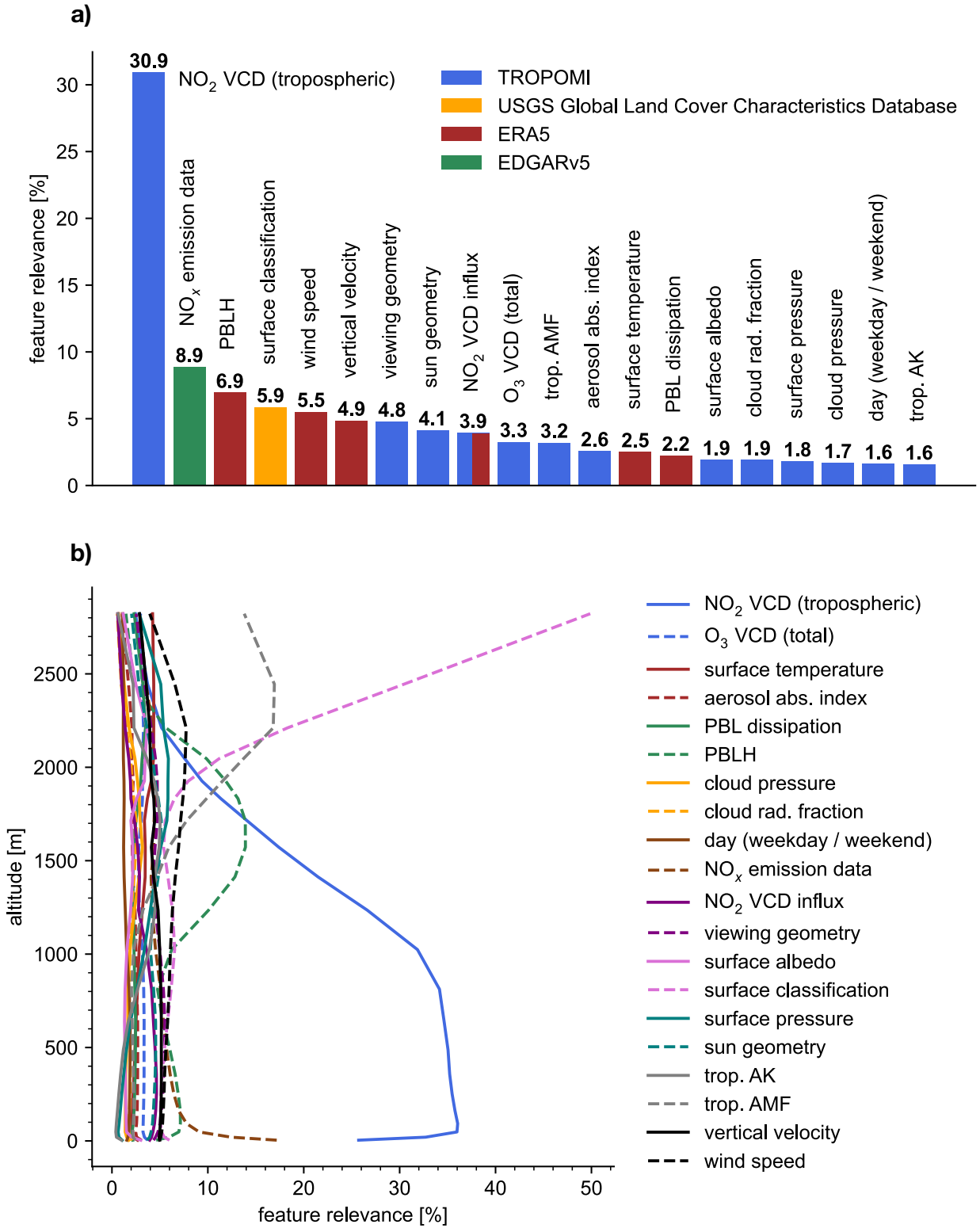

Figure 4.12: Feature relevances of NitroNet's input variables. a) Feature relevances, computed regardless of altitude. The bold numbers above each bar denote the percentual feature relevance of each feature (or feature group). The bars' colours represent the source of each feature, see the legend in the top-right corner. For a reference to the USGS Global Land Cover Characteristics Database, see Eskes et al. (2022). **b)** Feature relevance profiles, obtained from computing the feature relevances at each layer of WRF-2019 individually.

drop-off in NO_2 concentration (corresponding to the transition to the free troposphere) is expected to occur.

3. In the free troposphere, where the NO_2 concentrations are usually low, and much weaker correlated to the satellite observations, NitroNet's predictions rely on other features, such as the surface classification and the tropospheric air mass factor. How these bits of information are processed or interpreted by the neural network remains up for speculation. Most likely, NitroNet uses them as coarse identifiers of certain retrieval scenarios (e.g. "over water", "over land", etc.) and simply returns the average corresponding NO_2 concentrations learned from WRF-2019.

4. Some features, e.g. the tropospheric averaging kernels or the total O_3 VCD, show low feature relevance throughout. This indicates that they are either not informative, meaning they contain no helpful information for the reconstruction of the NO_2 profiles, or they are redundant to other input variables. For example, many of the variables that determine the tropospheric air mass factor (viewing geometry, cloud pressure, surface pressure, etc.) are available to NitroNet themselves.

5. There is a trade-off between the relevances of the NO_x emission data and the NO_2 VCD at the surface. Here, the relevance of the NO_2 VCD drops abruptly from approximately 36 % to 25 %, and the relevance of the NO_x emissions increases from approximately 8 % to 18 %. This demonstrates, that particularly at the surface, emission data yields an important contribution to the prediction quality. The NO_x emissions are expected to be most important near strong sources (such as power plants or highways), but this is not verified here due to the high computational expense of resolving the Shapley scores horizontally.

The influence of the NO_x emission data on NitroNet's surface predictions is further demonstrated in Fig. 4.13, visualizing the differences between predictions with and without NO_x emission data. For this experiment, NitroNet was supplied with NO_2 VCDs from May 2022, as shown in Figure 4.13a. Subfigures 4.13b-c show the corresponding predictions of the NO_2 surface concentrations, with all NO_x emission data set to 0 in Fig. 4.13b. The surface concentrations shown in Fig. 4.13b (without emission data) are essentially proportional to the NO_2 VCDs shown in Fig. 4.13a. Figure 4.13c, on the other hand, shows far more variability and small-scale structures corresponding to car highways between major cities, or shipping tracks over the sea. Note, that the NO_2 concentrations in Fig. 4.13c are considerably larger than in Fig. 4.13b. This is caused by NitroNet partially scaling its predictions based on the NO_x emission data (as opposed to the tropospheric NO_2 VCD alone, as one might think). This would not be expected to occur, had the NO_x emission data already been omitted during training, but it has no further relevance for the points made here.

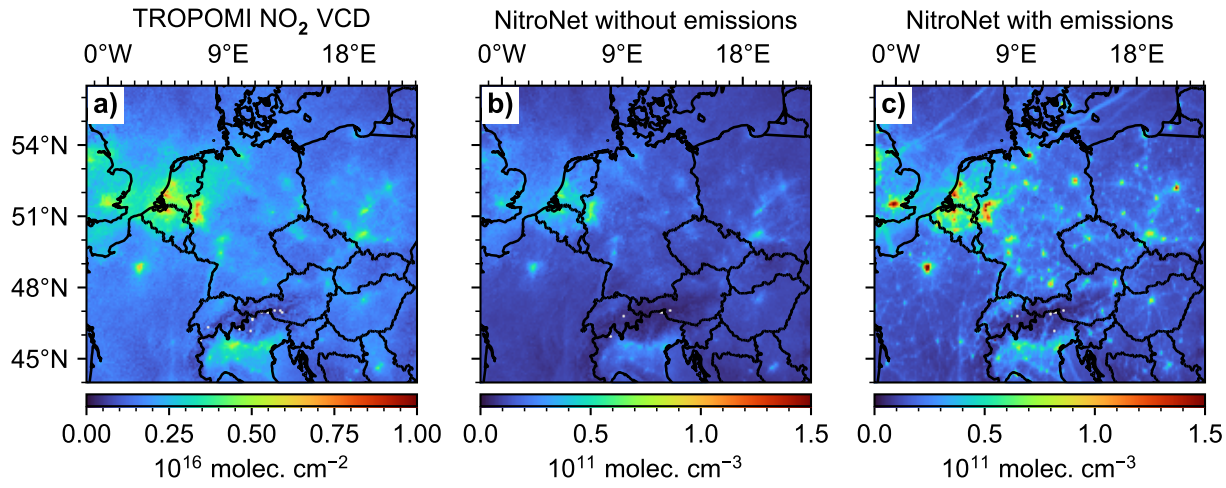

Figure 4.13: NitroNet predictions of the NO_2 surface concentration with and without emission data. a) Exemplary input NO_2 VCDs from TROPOMI. **b)** NitroNet predictions at the surface, with all emission variables set to 0. **c)** NitroNet predictions at the surface, with unchanged emission variables. Shown here are monthly means for the month of May 2022.

4.2.10 NO_2 gradients in the lowest 100 meters above ground

As explained in sect. 2.4.3, WRF-Chem operates on a vertical grid of fixed η-layers with different vertical extent depending on location and time. For example, the lowest layer of the WRF-Chem simulations discussed in Chapter 3 ranges from $0-8$ m on average (see e.g. Table 4.6). Thereby, the η-levels define the scale on which a WRF-Chem simulation can resolve NO_2 gradients. Although the number of η-levels can theoretically be increased, the randomness in their vertical extent and the additional computational burden remain major obstacles for the determination of NO_2 gradients at finer resolutions. As explained in sect. 4.1.2, NitroNet can predict NO_2 concentrations at arbitrary tropospheric altitudes. From the standpoint of statistical learning the neural network is expected to extract the relevant information from the variations of the layers' center altitudes, which stems from the conversion of the fixed η-levels to altitudes in meters. This results in a spread of the "altitude" variable, which is large enough to yield a continuous vertical coverage in specific altitude ranges (e.g. from $1500-2500$ m, see Fig. 4.5a).

In this context, two questions are adressed for the vertical range from $0-100$ m in the following:

1. How well can NitroNet reproduce/extrapolate the NO_2 gradients of WRF-Chem, particularly at sub-layer resolution (e.g. on the meter scale within the lowest 10 m above ground)?

2. Are the observed NO_2 gradients on these scales large enough to deem this a valuable

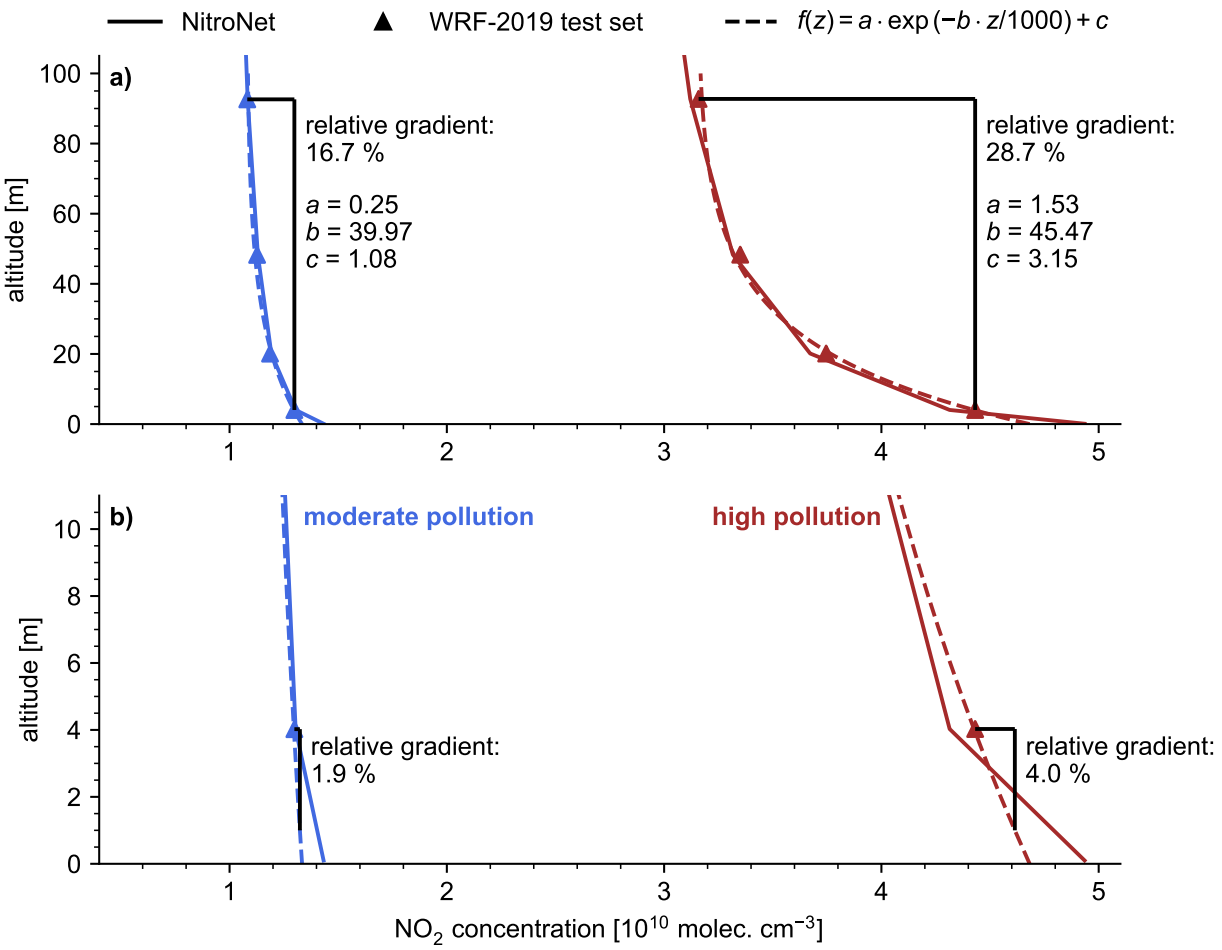

Figure 4.14: Average NO_2 profiles from NitroNet and WRF-Chem (filtered test set) in the lowest 100 m above ground. a) Vertical range from 0 m – 100 m. **b)** Vertical range from 0 m – 10 m. The blue lines refer to "moderate pollution" scenarios (0 % – 75 % quantile of the WRF-Chem NO_2 concentration in the lowest layer). The red lines refer to "high pollution" scenarios (75 % – 100 % quantile). The dashed lines show exponential fits through the NO_2 concentrations of the lowest four WRF-Chem layers. The fit function is given in the figure's legend.

benefit of NitroNet over classic RCT simulations? For example, would the ability to distinguish NO_2 concentrations at 1 m (the approximate size of children) and 4 m (the typical inlet altitude of in situ measurements) above ground be significant for studies on the medical impact of NO_2 exposure?

Figure 4.14 gives an overview of the average NO_2 profiles from NitroNet and WRF-Chem (taken from the filtered test set) in different vertical ranges. The figure distinguishes between two scenarios: "moderate pollution" (blue, identified as the 0 % – 75 % quantile of the NO_2 concentration in the lowest layer) and "high pollution" (red, identified as the 75 % – 100 % quantile). Subfigure 4.14a shows the average NO_2 profiles in the lowest 100 m, with a characteristic exponential shape and relative average gradients of 16.7 % (moderate pollution)

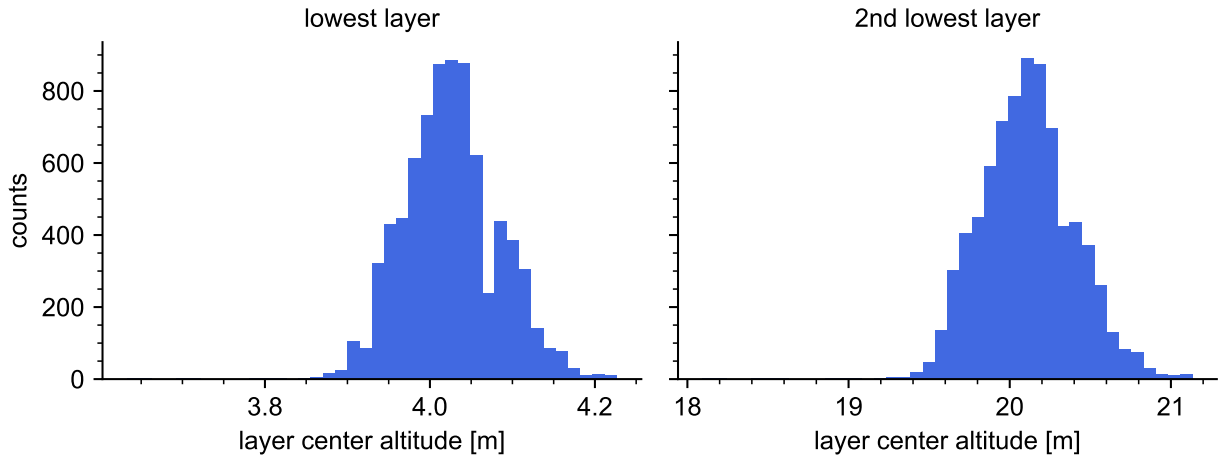

Figure 4.15: Histogram of the layers' center altitudes in the lowest two WRF-Chem simulation layers (filtered test set).

and 28.7 % (high pollution). These gradients were computed based on the NO_2 concentrations from WRF-Chem in the lowest layer (approx. $0 - 8$ m) and the fourth layer (approx. $66 - 125$ m). WRF-Chem and NitroNet are in excellent qualitative and quantitative agreement, as already documented in Table 4.4. Additionally, exponential fits with the fit function

$$f(z) = a \cdot \exp\left(-b \cdot \frac{z}{1000}\right) + c \tag{4.9}$$

where:

z : altitude in meters

$f(z)$: NO_2 concentration in 10^{10} molec. cm^{-3}

a, b, c : the free fit parameters

were computed, which are used in the next step to estimate the relative NO_2 gradients in the lowest 10 m, shown in 4.14b. It can be attempted to compute average relative gradients in a sub-layer vertical range (e.g. here: between 4 m and 1 m). However, because WRF-Chem does not resolve the NO_2 gradients at this scale, it is required to compute the gradients using either the previously computed exponential fits or the NO_2 profiles from NitroNet as extrapolations of the WRF-Chem data. Figure 4.14b shows the relative NO_2 gradients obtained from the exponential fits, which are 1.9 % (moderate pollution) and 4.0 % (high pollution). NitroNet predicts larger surface concentrations in both pollution scenarios, leading to larger relative gradients of 7.0 % (moderate pollution) and 10.1 % (high pollution), omitted from Fig. 4.14b for better readability. These results must be interpreted in the context of the available training data. As revealed by Fig. 4.15, the spread of the lowest two layers' center altitudes is quite

narrow. The figure shows data from the filtered test set, whose distribution is identical to that of the training set. The center altitudes of the lowest layer only reach from approx. $3.9 - 4.2$ m (second lowest layer: approx. $19.5 - 21.0$ m). This spread is obviously too small to enable genuine predictions in-between layers (e.g. at 1 m above ground) based on actual information extracted from the training set. In conclusion, NitroNet's ability to make sub-layer predictions has no significant benefit over simple extrapolation from the (approximate) vertical grid of the WRF-Chem training data. However, as mentioned before, this does not apply to the entire troposphere. For example, in the range from $1500 - 2500$ m the spread of the altitude variable is large enough to result in an effectively gap-free vertical coverage (see Fig. 4.5a).

Ideally, these findings should be validated by independent measurements. The existing literature on near-surface NO_2 gradients is quite sparse, unfortunately. Alicke (2000) presents DOAS measurements from Milan, Italy (May 1998) in the lowest 5 m above ground. Nighttime gradients of up to 20 % were found in this shallow vertical range, but no significant daytime gradients. Similar measurements in a vertical range of up to 110 m are reported by Stutz et al. (2004). The daytime NO_2 gradients in this vertical range were found to be as large as 20 ppb ($\approx 5 \cdot 10^{11}$ molec. cm^{-3} under standard conditions), although only for short periods in which the measurements appear to have been influenced by strong local NO_2 emissions (e.g. from a shipping channel and a nearby airport). Due to the presentation of the results in graphical form only, it is hardly possible to quantify relative gradients. Data from the recently conducted CINDI-3 measurement campaign in Cabauw, Netherlands (see `https://frm4doas.aeronomie.be/index.php/news?view=article&id=23`) might provide further insight into the near-surface NO_2 gradients and, in contrast to the aforementioned historical data, could be directly compared to NitroNet predictions once published.

With this, the two questions formulated earlier (referring to the NO_2 gradients in the lowest 100 m above ground) can be answered as follows:

1. NitroNet can reproduce the NO_2 concentrations from WRF-Chem with relative errors of < 5 % (see also Table 4.4). NitroNet can also predict NO_2 concentrations on the sub-layer resolution (i.e. "in-between" the layers it was trained on). However, as the training dataset shows significant gaps in the vertical coverage, such predictions should not be treated as genuinely well-informed, but rather as extrapolations from the approximate vertical grid of the WRF-Chem simulation.

2. Over the lowest 100 m, the NO_2 profiles from WRF-Chem and NitroNet show steep exponential vertical gradients of up to 30 % in polluted regions. Extrapolations of the NO_2 profiles towards the surface indicate relative NO_2 gradients of up to 10 % within the lowest 5 m above ground. The few published studies on the topic report significant

NO_2 gradients of up to 20 ppb in the lowest 100 m with large temporal inconsistencies and no conclusive results in the lowest 5 m.

In conclusion, the magnitude of the NO_2 gradients in the lowest few meters above ground can currently not be assessed with confidence. Furthermore, NitroNet's nominal ability to predict NO_2 concentrations at sub-layer resolution appears to have no clear advantage over simple extrapolation of the WRF-Chem training data. Besides, the evaluation of NitroNet and WRF-Chem against observational in situ data (see sect. 4.3.1) reveals that the influence of the Mo-CL bias is a far more dominant factor than the small NO_2 gradients expected on the (sub)meter scale.

4.3 Validation of NitroNet using observational data

After the formal validation of NitroNet on the test set NitroNet is now evaluated against various observational datasets. Here, the same reference data as in Chapter 3 are used, i.e. tropospheric NO_2 VCDs from TROPOMI, NO_2 surface concentrations from AirBase in situ measurements (using only the background instruments), and NO_2 profiles from the FRM_4DOAS MAX-DOAS measurements. Like during training, only data with a TROPOMI quality score of $f_{QA} > 0.75$ are used. As a first step in this validation study, an intercomparison between WRF-Chem, NitroNet, TROPOMI NO_2 VCDs and AirBase NO_2 surface concentrations is presented in section 4.3.1 with the intent to demonstrate the benefits of the data curation procedure described in sect. 4.2.2. This analysis is based on the comparison of monthly average data, using NitroNet predictions from May 2019, as seen during the training. Then, in section 4.3.2, NitroNet is evaluated on previously unseen data featuring comparisons to TROPOMI, AirBase and FRM_4DOAS, with and without monthly averaging. Many qualitative descriptions of the data shown here (e.g. on the geographic distribution of NO_2 VCDs or MAX-DOAS profile shapes) were already given in Chapter 3 and are not repeated. The statistical diagnostics mentioned in the following main text refer to the monthly means (e.g. RMSE values computed on pairs of monthly mean values), unless specified otherwise. The uncertainty of the NitroNet predictions was estimated as described in sect. 2.5.7 and amounts to approx. $30\ \% - 60\ \%$ (depending on the uncertainties of the input variables, particularly those of the NO_2 VCDs used as input), but is not shown in order to maintain easier readability.

4.3.1 Comparison to WRF-Chem simulation results (May 2019)

Remarks on vertical interpolation

The intercomparison of simulated VCDs from WRF-Chem and NitroNet to TROPOMI observations requires a mutual scheme for vertical interpolation. As explained in sect. 2.4.3, the barometric height formula is used to convert the vertical pressure levels of the TROPOMI retrieval (at which the averaging kernels are defined) to altitude in meters. This conversion also considers the vertical temperature profiles (see eq. 2.2.3 in sect. 2.2). The WRF-Chem simulation results shown in Chapter 3 used WRF-Chem's simulated temperature profiles for this purpose. NitroNet cannot predict temperature profiles, and operates on the premise of the fixed dry-adiabatic lapse rate ($\Gamma = -0.0098$ K m^{-1}) and the surface temperature of the international standard atmosphere ($T_0 = 288.15$ K). In the pursuit of intercomparing NitroNet and WRF-Chem, this method was applied to WRF-Chem as well, so that both models use the same vertical interpolation scheme. Note, that this study uses data from May 2019, like during training. Nonetheless, approximately 93 % of the input data are entirely new to NitroNet as a consequence of the rigorous filtering that had been previously applied to the training data.

Comparison to TROPOMI satellite measurements

Figure 4.16a shows the comparison of tropospheric NO_2 VCDs from WRF-Chem and TROPOMI for the month of May 2019. The air mass factors of the TROPOMI VCDs were re-computed using the NO_2 profiles from WRF-Chem, as described in sect. 2.3.4. The WRF-Chem results shown here are almost identical to those from Chapter 3, see Fig. 3.16g, due to the minimal differences in the vertical interpolation procedure. The aforementioned adjustments to the vertical interpolation scheme can therefore be considered viable.

Figure 4.16b shows the same comparison, except that the simulated NO_2 VCDs and

	bias	RMSE	R	reference
WRF-Chem vs. TROPOMI	-2.9 %	$6.7 \cdot 10^{14}$ molec. cm^{-2}	0.88	Fig. 4.16a
NitroNet vs. TROPOMI	-8.1 %	$3.8 \cdot 10^{14}$ molec. cm^{-2}	0.97	Fig. 4.16b
WRF-Chem vs. AirBase	-11.7 %	3.4 µg m^{-3}	0.69	Fig. 4.17a
NitroNet vs. AirBase	-16.0 %	3.2 µg m^{-3}	0.67	Fig. 4.17b
WRF-Chem vs. AirBase (w/o urban)	-3.5 %	3.1 µg m^{-3}	0.67	Fig. C.2a
NitroNet vs. AirBase (w/o urban)	-9.1 %	2.9 µg m^{-3}	0.64	Fig. C.2b

Table 4.7: Statistical summary of Fig. 4.16 and Fig. 4.17 based on monthly-mean data of May 2019. "w/o urban" refers to an evaluation without urban background stations, see the discussions in sect. 4.3.2.

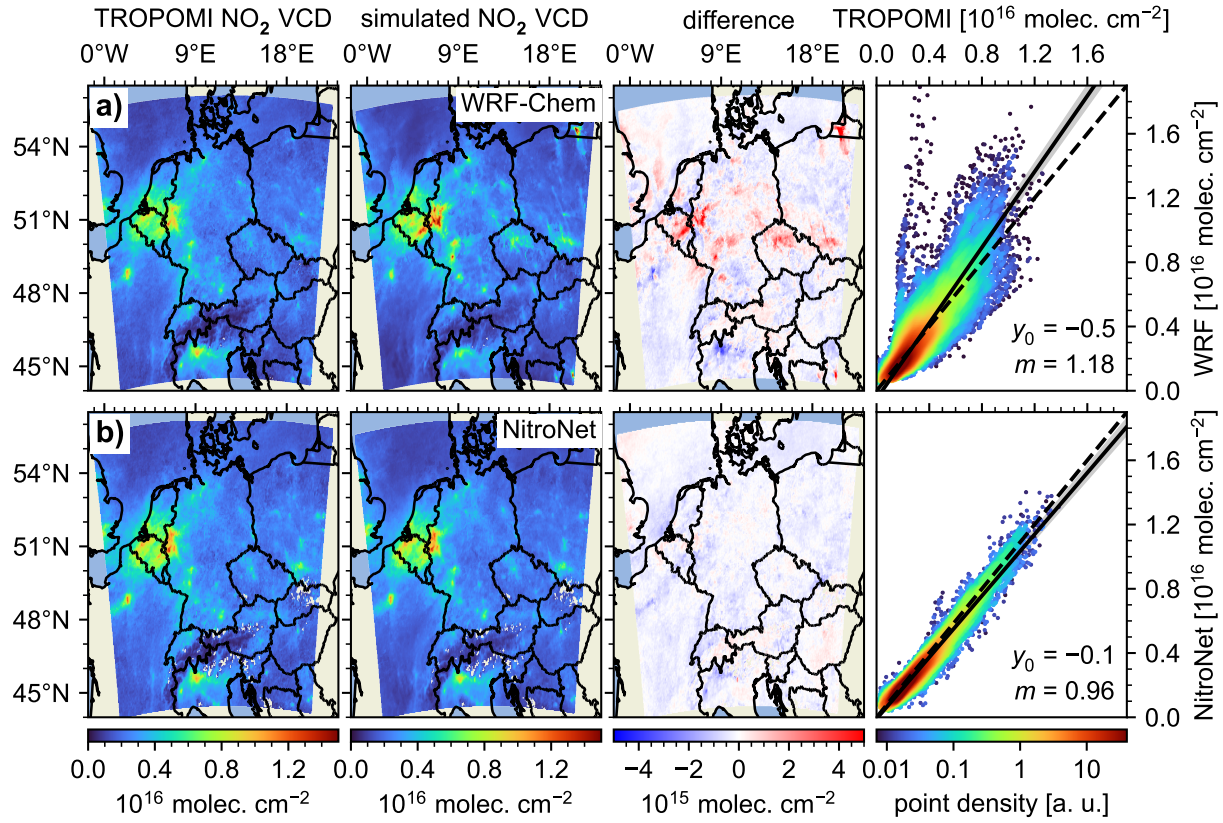

Figure 4.16: Intercomparison of monthly-mean tropospheric NO_2 VCDs from NitroNet, WRF-Chem, and TROPOMI satellite measurements, May 2019. **a)** WRF-Chem vs. TROPOMI. **b)** NitroNet vs. TROPOMI. Intercepts (y_0) are given in units of 10^{15} molec. cm^{-2}. The air mass factors of the TROPOMI reference data were re-computed using the NO_2 profiles from WRF-Chem and NitroNet, respectively. Further statistical diagnostics are given in Table 4.7.

air mass factors were computed using the NO_2 profiles from NitroNet, resulting in significantly better agreement to the TROPOMI NO_2 VCDs. A summary is given in Table 4.7. Most notably, the RMSE of NitroNet is almost halved (from $6.7 \cdot 10^{14}$ molec. cm^{-2} to $3.8 \cdot 10^{14}$ molec. cm^{-2}), and the correlation coefficient increased (from $R = 0.88$ to $R = 0.97$). This marks fundamental improvements, achieved by training NitroNet on a filtered dataset. In some regions of the domain (e.g. near large cities such as Frankfurt or Mannheim), these improvements relate to a significant reduction of the simulated NO_2 VCDs. Likewise, underestimations in other places are reduced, where NitroNet produces larger tropospheric columns (e.g. in the Lombardy region in northern Italy). In other cases the improvements must instead be (partially) attributed to larger TROPOMI reference VCDs, resulting from the use of presumably more realistic NO_2 a priori profiles (e.g. at the border between Belgium, the Netherlands, and Germany). Overall, the NO_2 VCDs from NitroNet exhibit a slight low bias throughout, while the VCDs from WRF-Chem are characterized by severe negative and positive biases in different subregions of the domain.

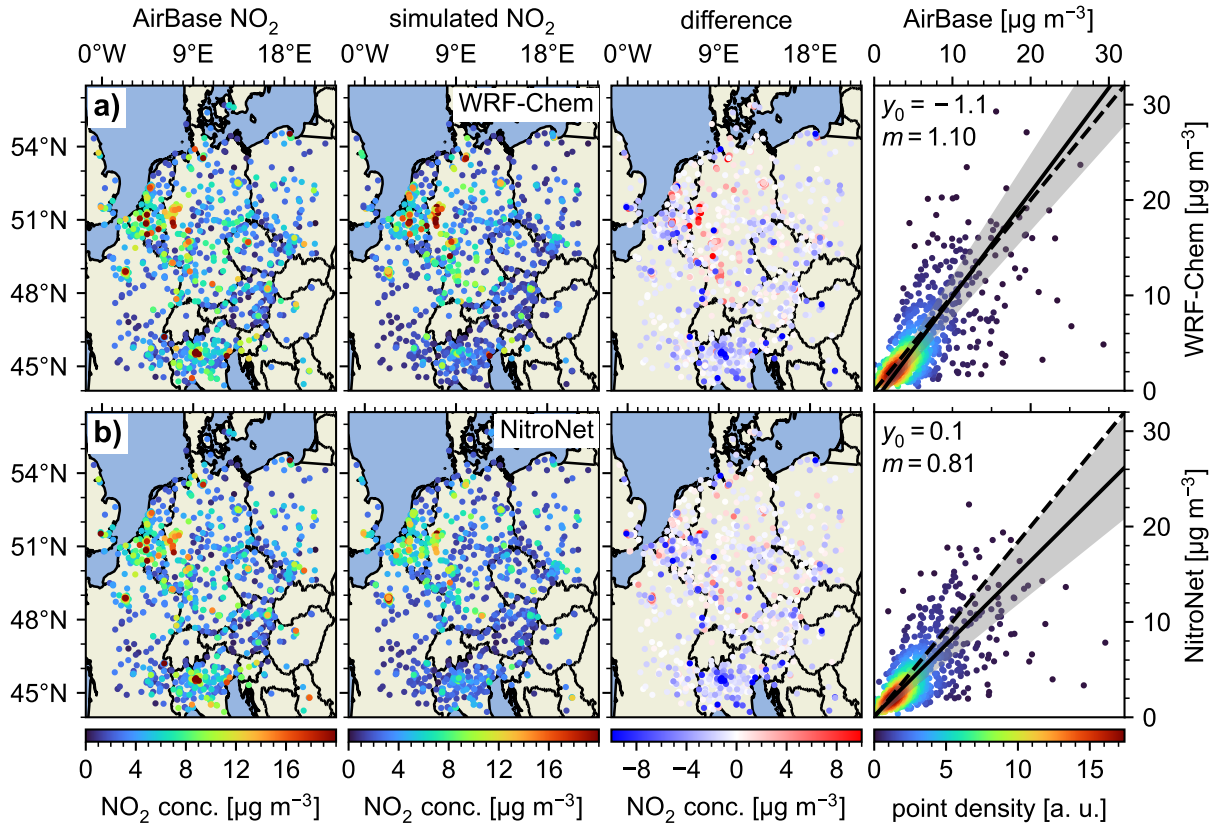

Figure 4.17: Intercomparison of monthly-mean NO_2 surface concentrations from NitroNet, WRF-Chem, and AirBase in situ measurements, May 2019. **a)** WRF-Chem vs. AirBase. **b)** NitroNet vs. AirBase. Intercepts (y_0) are given in units of µg m^{-3}. Further statistical diagnostics are given in Table 4.7.

Comparison to AirBase in situ measurements

Figure 4.17 shows the comparison to NO_2 surface concentrations from the AirBase dataset in analogy to Fig. 4.16. Like in Chapter 3 only background instruments were used and the Mo-CL biases were corrected using the correction factors F from WRF-Chem and NitroNet's F-network, respectively. Here, the affected AirBase measurements were divided by F. The NitroNet predictions were generated in half-meter vertical steps between $0 - 10$ % above ground, and interpolated to the inlet height of the AirBase instruments (typically at 4 – 5 m above ground). The WRF-Chem results shown here differ from those seen in Chapter 3, because they were interpolated to the NitroNet prediction grid with the quality criterium $f_{\mathrm{QA}} > 0.75$ applied. In (rare) cases in which the F-network predicted a Mo-CL bias correction factor of below 0 the prediction was dismissed altogether. The entries of Table 4.7 referring to an evaluation without urban background stations ("w/o urban") can be ignored for now, and are only presented for completeness with regards to the discussions in sect. 4.3.2 later on.

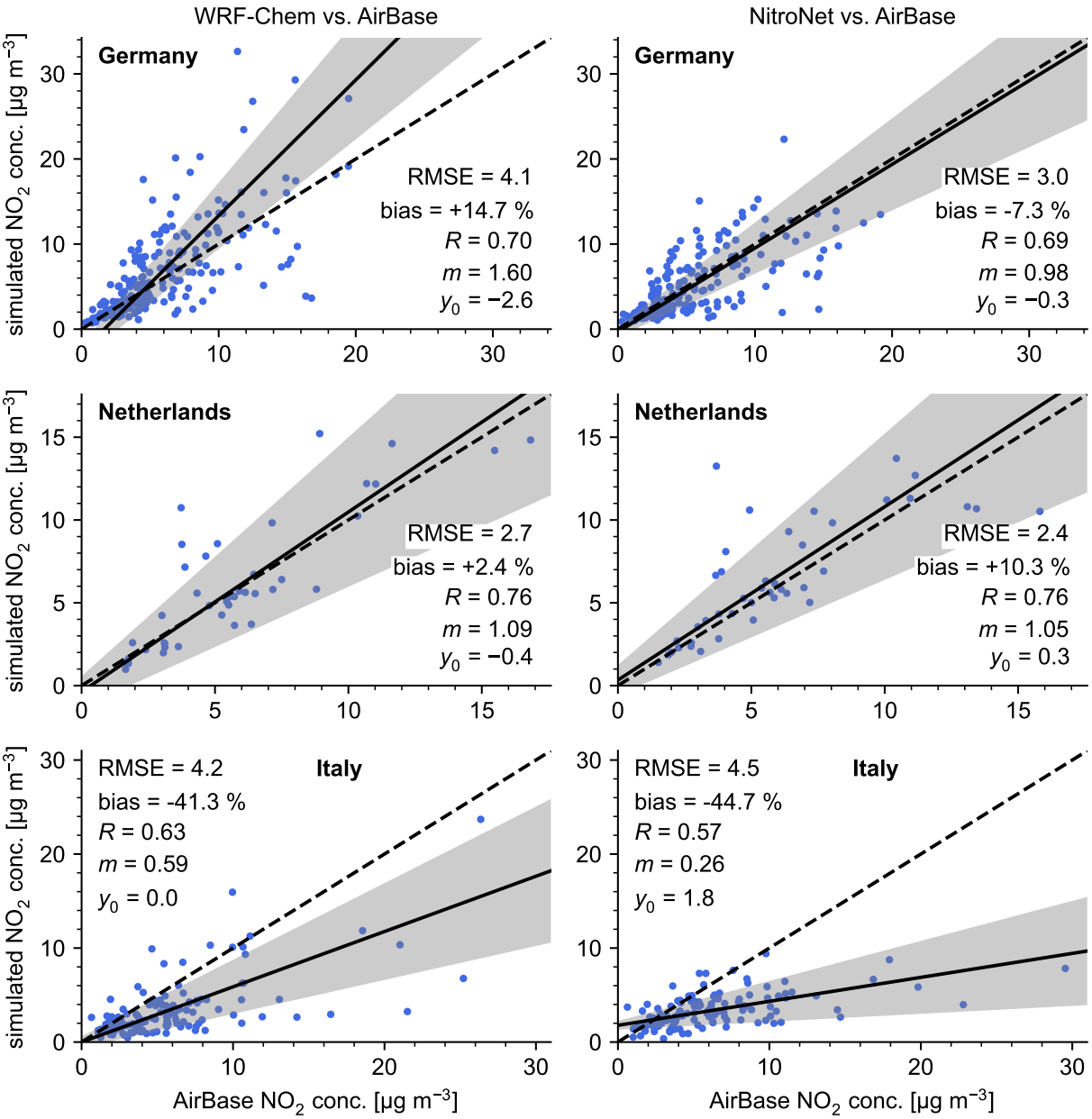

Figure 4.18: Intercomparison of monthly-mean NO_2 surface concentrations from NitroNet, WRF-Chem, and AirBase in situ measurements for specific European countries, May 2019. Each row corresponds to a specific European country (top to bottom: Germany, Netherlands, Italy). RMSE values and intercepts (y_0) are given in units of µg m^{-3}.

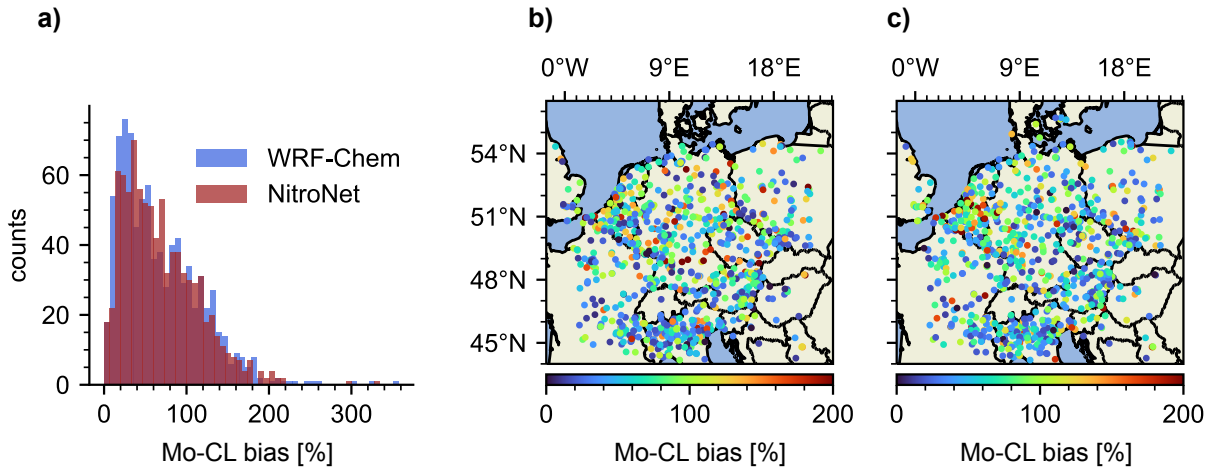

Figure 4.19: Comparison of the monthly-mean Mo-CL biases as estimated by WRF-Chem and NitroNet, May 2019. a) Histogram of the Mo-CL biases. **b)** Mo-CL biases estimated by WRF-Chem. **c)** Mo-CL biases estimated by NitroNet.

In some regions of the domain NitroNet shows less systematic overestimations in NO_2 surface concentration than WRF-Chem, similar to the improvements observed in the context of the NO_2 VCDs. However, in other regions such as in northern Italy, NitroNet shows similar underestimations as WRF-Chem, despite the aforementioned considerably better agreement with TROPOMI. This is further exemplified by Fig. 4.18, comparing NitroNet and WRF-Chem in individual European countries. NitroNet clearly outperforms WRF-Chem in Germany, but shows similar (or slightly worse) performance in the Netherlands and Italy. Overall, as listed in Table 4.7, NitroNet produces slightly lower RMSE values than WRF-Chem (3.2 µg m^{-3} vs. 3.4 µg m^{-3}), but weaker correlation to the AirBase reference data ($R = 0.67$ vs. $R = 0.69$). Furthermore, NitroNet's mean bias is larger than WRF-Chem's (-16.0 % vs. -11.7 %), although this should be interpreted with care: Figure 4.16 indicates that in the comparison to TROPOMI, WRF-Chem produces positive and negative biases in approximate balance, while NitroNet produces weaker negative, and almost no positive biases. In the comparison to AirBase, NitroNet also shows lower positive biases, but similar negative biases as WRF-Chem (hence the lower RMSE values). This results in a stronger overall low bias, but marks an improvement nonetheless. Under such circumstances the mean bias is obviously not a suitable measure for overall model skill, and veils the considerable improvements obtained e.g. in Germany.

Figure 4.19 gives an overview of the Mo-CL biases as estimated by WRF-Chem and NitroNet. The histogram in Fig. 4.19a reveals that the Mo-CL biases are mostly below 200 %, but can reach beyond 300 % in rare cases. The distributions of WRF-Chem and NitroNet peak at $\sim$ 30 %, with mean values at $\sim$ 65 %. The good agreement between WRF-Chem

NitroNet compared to	bias	RMSE	R	reference
TROPOMI	+6.7 %	$2.8 \cdot 10^{14}$ molec. cm^{-2}	0.95	Fig. 4.20a
AirBase	−10.5 %	1.7 µg m^{-3}	0.75	Fig. 4.20b
AirBase (w/o urban)	+2.2 %	1.2 µg m^{-3}	0.73	Fig. 4.20c

Table 4.8: Statistical summary corresponding to Fig. 4.20 based on monthly-mean data of May 2022. "w/o urban" refers to evaluations in which AirBase instruments classified as "urban background" were excluded.

NitroNet compared to	bias	RMSE	R	reference
TROPOMI	+7.0 %	$6.9 \cdot 10^{14}$ molec. cm^{-2}	0.83	Fig. 4.20a
AirBase	−10.4 %	2.8 µg m^{-3}	0.58	Fig. 4.20b
AirBase (w/o urban)	+2.6 %	2.2 µg m^{-3}	0.53	Fig. 4.20c

Table 4.9: Like Table 4.8, but based on individual orbits/hourly data.

and NitroNet aligns well with the low test errors of NitroNet's F-network (~ 6 %) reported in sect. 4.2.5.

4.3.2 Validation against observational reference data (May 2022)

This section addresses the main validation of NitroNet against the available reference data. Here, detailed quantitative comparisons are made to data from TROPOMI, AirBase and FRM$_4$DOAS. The main difference to sect. 4.3.1 is that this validation is conducted on entirely new data from May 2022. Furthermore, no WRF-Chem data are available from this point on. All statistical diagnostics referenced in the following are summarized in Table 4.8 (based on monthly-means) and Table 4.9 (based on individual orbits/hourly data).

Comparison to TROPOMI satellite measurements

Figure 4.20a shows the comparison of monthly-mean NO_2 VCDs from TROPOMI against corresponding NitroNet predictions. This evaluation on data from May 2022 is qualitatively similar to that from May 2019 (see Fig. 4.16), although a few differences can be identified. NitroNet produces even lower RMSE values ($2.8 \cdot 10^{14}$ molec. cm^{-2} vs. $3.8 \cdot 10^{14}$ molec. cm^{-2}, corresponding to a relative reduction by 26 %) and slightly overestimates the NO_2 VCDs (mean bias of +6.7 % vs. −8.1 %). A similarly high correlation coefficient (0.95 vs. 0.97) is achieved. Notice, that compared to May 2019 the NO_2 VCDs of May 2022 are on average 18 % smaller. This affects the RMSE in two ways: Firstly, the RMSE is not a relative

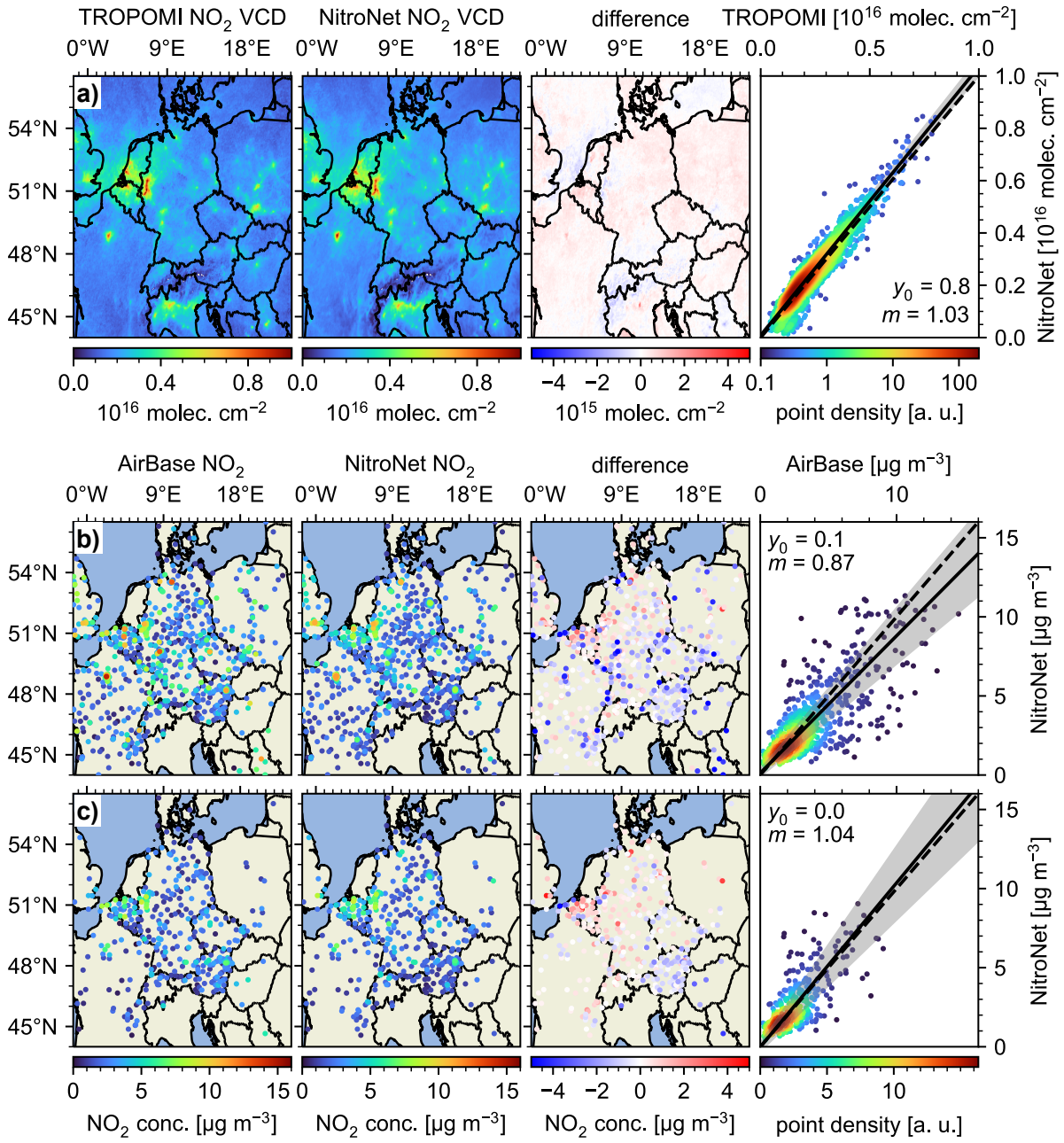

Figure 4.20: Comparison of monthly-mean tropospheric NO_2 VCDs from TROPOMI satellite measurements and surface NO_2 concentrations from AirBase in situ measurements to corresponding NitroNet predictions, May 2022. a) NitroNet vs. TROPOMI. The air mass factors of the TROPOMI reference data were re-computed using the NO_2 profiles from NitroNet. The right-side scatter plots were drawn from 10000 random samples. Intercepts (y_0) are given in units of 10^{15} molec. cm^{-2}. **b)** NitroNet vs. AirBase. The Mo-CL correction factors from NitroNet's *F*-network were applied. Intercepts (y_0) are given in units of µg m^{-3}. **c)** Like b), but without urban background instruments. A statistical summary is found in Tables 4.8 and 4.9.

error metric, and thus scales with its input data. Secondly, NitroNet might make more accurate predictions on cases of lower NO_2 pollution, because these are far more represented in the training set (see e.g. Fig. 4.4). Beyond that, the significantly higher biases observed across the entire domain indicate inherent differences between the joint input data from May 2022 (including the NO_2 VCD itself) and those from May 2019. Unfortunately, the neural network is essentially a black box in this regard, and therefore the said differences must remain unspecified. Nonetheless, NitroNet's overall agreement to the TROPOMI NO_2 VCDs of May 2022 is satisfactory, particularly compared to the performance of WRF-Chem in 2019, see Table 4.7. A version of Fig. 4.20 based on data of a single day is found in Fig. C.4, characterized by larger dispersion in the scatter plots due to larger random noise in the absence of averaging. Other than that, no significant differences to the evaluation based on monthly averages can be identified.

Comparison to AirBase in situ measurements

Figure 4.20b shows the comparison between monthly-mean NO_2 surface concentrations from AirBase and NitroNet. Only background measurements were used. Compared to May 2019, NitroNet shows a lower RMSE (1.7 µg m^{-3} vs. 3.2 µg m^{-3}), a lower mean bias (−10.5 % vs. −16.0 %), and increased correlation ($R = 0.75$ vs. $R = 0.67$). However, these supposed improvements cannot be attributed to a better prediction quality of NitroNet. Instead, they relate to the measurements in the Lombardy region in northern Italy, to which NitroNet showed strong underestimations in May 2019. In 2022, the corresponding instruments are missing entirely. This is also the case in other countries (e.g. in the Netherlands), but as revealed by comparison of Fig. 4.17b and Fig. 4.20b, the relevant loss of critical measurements occurs in Italy. According to the AirBase metadata, 92 % of the Italian AirBase instruments were flagged "valid", 5 % as "invalid", and 2 % as "below detection limit" in May 2019 (see the Eionet dictionary, accessible under `dd.eionet.europa.eu/vocabularies` for the EEA's definition of these terms). In May 2022, 48 % were flagged as "valid", 13 % as "invalid", and 39 % as "below detection limit". The increased portion of measurements below the detection limit might be a consequence of the steadily decreasing European NO_2 pollution levels over the past years (see e.g. Cooper et al., 2022; Anenberg et al., 2022). Additionally, the total number of Italian AirBase instruments has shrunk from 320 to just 69, for which the reason remains unclear.

Another important observation is the dependence of NitroNet's low bias on the reference instruments' types (see sect. 2.3.6 and Fig. 2.14). It was shown in sect. 3.2 that WRF-Chem (and by extension NitroNet) cannot reproduce traffic pollution accurately. For this reason the validation studies are restricted to background measurements. The EEA's definition of the term "background instrument" reads: "*Located such that its pollution levels are representative*

of the average exposure of the general population within the type of area under assessment. The pollution level should not be dominated by a single source type (e.g. traffic), unless that source type is typical within the area under assessment. The station should usually be representative of a wider area of at least several square kilometers" (see `dd.eionet.europa.eu/vocabularies`). In other words, even instruments dominated by traffic emissions are classified as "background" (instead of "traffic") if they are representative of a larger surrounding area. It is reasonable to assume, that this predominantly occurs in urban areas. In order to investigate the effects of this possibly misleading instrument classification, the preceding analysis is repeated under omission of the "urban background" stations, as shown in Fig. 4.20c. This results in a reduced mean bias (+2.2 % vs. −10.5 %), a reduced RMSE (1.2 µg m^{-3} vs. 1.7 µg m^{-3}), and a better linear fit (see rightmost panel of Fig. 4.20c). The correlation coefficient, however, is slightly reduced from 0.75 to 0.73. Nonetheless, this marks a significant improvement, which might reflect a tendency of NitroNet to underestimate NO_2 concentrations in certain urban areas. On the other hand, this is not supported by the comparison to TROPOMI data, which shows no significantly enhanced underestimations in most urban regions. Together this indicates that NitroNet predicts NO_2 profiles with realistic tropospheric columns, but (sometimes) faulty profile shapes with too little NO_2 at the surface. In other words, NitroNet might suffer from similar misrepresentations of vertical mixing as were discovered previously in the WRF-Chem model. However, the highly variable Mo-CL bias of the AirBase measurements and the vague information of their classification process add significant uncertainty to this consideration. Based on the available information no definite conclusions can be drawn on whether the inclusion of urban background stations is justified or not. Therefore all remaining evaluations are presented in both variants: with and without the urban background stations. For brevity, the discussions in the main text are limited to the case *without* urban background stations, but all corresponding figures and tables include the case with urban background stations (some are found in Appendix C.2).

Comparison to NO_2 profiles from FRM_4DOAS MAX-DOAS measurements

As a last step in this validation study, NO_2 profiles from FRM_4DOAS MAX-DOAS measurements are compared to corresponding NitroNet predictions. The evaluation method is identical to that described in sect. 3.7.3, including the application of the MMF averaging kernels, the FRM_4DOAS processor version (fv003) and the practice of averaging over all instruments deployed in each location. No data from Mainz is available in May 2022, but instead, data from San Pietro Capofiume (Italy) and Cabauw (Netherlands). A temporal colocation threshold of 60 minutes is used, meaning that each NitroNet NO_2 profile is associated with the average over all colocated MAX-DOAS profiles retrieved within ±60 minutes of the corresponding satellite overpass.

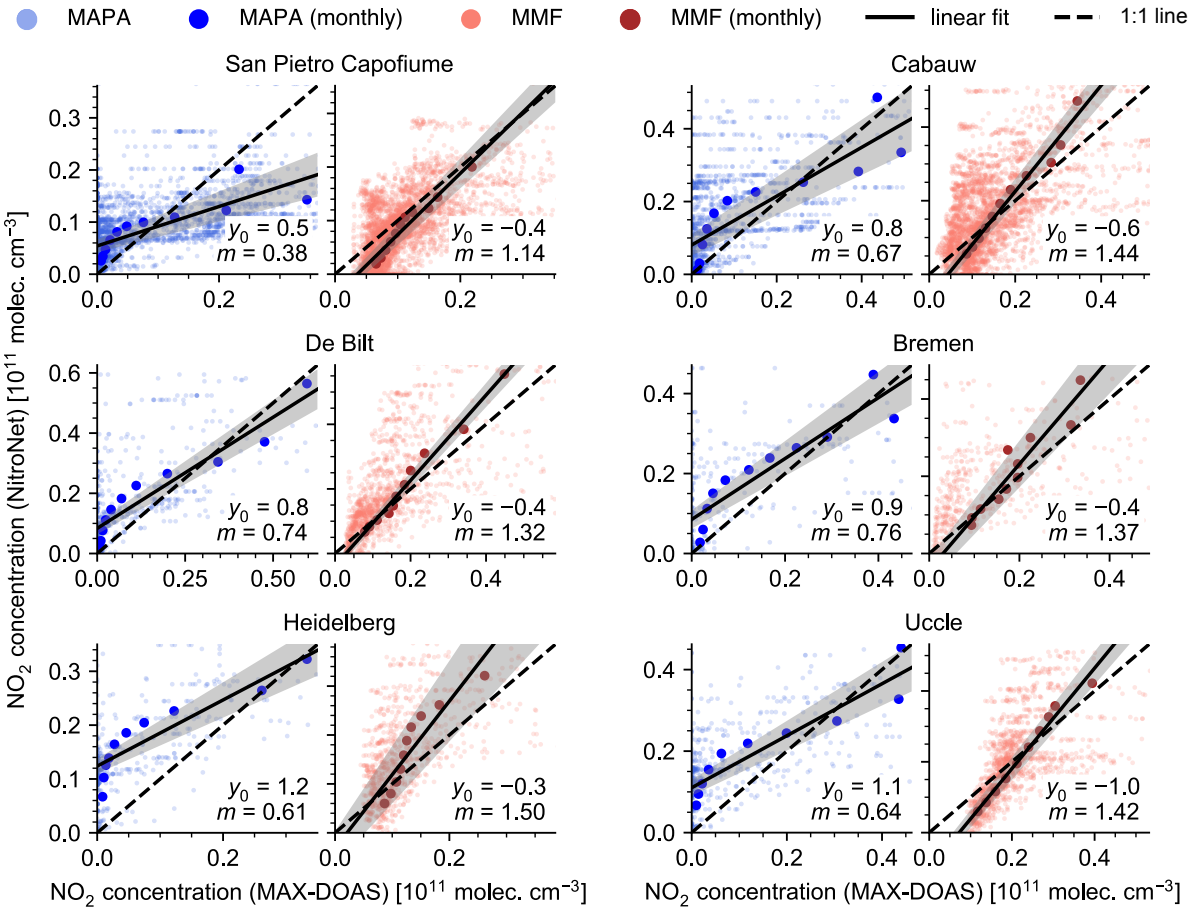

Figure 4.21: Comparison of NO_2 concentrations from FRM$_4$DOAS MAX-DOAS measurements to corresponding NitroNet predictions in the lowest 2 km, May 2022. MAPA results are drawn in blue, and MMF results in red. Thin scatter points represent a one-to-one comparison of NO_2 concentrations regardless of altitude. Thick scatter points represent the monthly-mean NO_2 concentrations of each retrieval layer. Intercepts (y_0) are given in units of 10^{10} molec. cm^{-3}. Further statistical diagnostics are given in Tables 4.10 and 4.11.

Figure 4.21 shows the results of this evaluation with the corresponding statistical diagnostics given in Tables 4.10 (based on monthly-mean data) and 4.11 (based on hourly data). The following text refers to the monthly-mean data given in Table 4.10. In Fig. 4.21 the thin scatter points represent a one-to-one comparison of hourly NO_2 concentration values from NitroNet, MAPA, and MMF (here: regardless of the altitude). The thick scatter points represent the monthly-mean NO_2 concentrations corresponding to each layer of the FRM$_4$DOAS profile retrieval (vertical layer extent: 200 m). The level of agreement between FRM$_4$DOAS and NitroNet varies, depending on the retrieval algorithm and location. MAPA shows significant differences in some locations, with biases ranging from -9.1 % to $+96.5$ %, RMSE values of $\sim 0.8 \cdot 10^{10}$ molec. cm^{-3}, and correlation coefficients ranging from $R = 0.84$ to $R = 0.96$. The agreement to MMF is significantly better, with biases in

	MMF			MAPA		
	bias	RMSE	R	bias	RMSE	R
San Pietro Capofiume	−20.6 %	0.3	0.98	−9.1 %	0.7	0.84
De Bilt	+12.6 %	0.5	0.99	+22.5 %	0.8	0.96
Heidelberg	+27.5 %	0.6	0.91	+96.5 %	1.0	0.92
Cabauw	+11.0 %	0.5	0.99	+12.3 %	0.9	0.89
Bremen	+14.4 %	0.5	0.95	+27.6 %	0.7	0.92
Uccle	−2.7 %	0.4	0.99	+30.9 %	0.9	0.94

Table 4.10: Validation of NitroNet against monthly-mean NO_2 profiles from FRM_4DOAS in the lowest 2 km, May 2022. RMSE values are given in units of 10^{10} molec. cm^{-3}.

	MMF			MAPA		
	bias	RMSE	R	bias	RMSE	R
San Pietro Capofiume	−20.6 %	0.7	0.63	−11.6 %	1.3	0.54
De Bilt	+12.6 %	2.0	0.54	+19.1 %	2.9	0.57
Heidelberg	+27.5 %	1.1	0.54	+74.3 %	1.8	0.48
Cabauw	+11.0 %	1.4	0.67	+21.7 %	1.7	0.73
Bremen	+14.4 %	2.3	0.25	+25.9 %	2.4	0.40
Uccle	−2.7 %	1.2	0.70	+23.6 %	2.2	0.61

Table 4.11: Like Table 4.10, but based on hourly data.

the range from −20.6 % to +27.5 %, RMSE values of $\sim 0.5 \cdot 10^{10}$ molec. cm^{-3} and correlation coefficients of $R > 0.90$. The linear regression through the point clouds in Fig. 4.21 (here drawn through the monthly-mean values) show steeper slopes and lower intercepts for MMF ($1.14 \leq m \leq 1.50$, and $-1.0 \cdot 10^{10}$ molec. $cm^{-3} \leq y_0 \leq -0.3 \cdot 10^{10}$ molec. cm^{-3}) than for MAPA ($0.38 \leq m \leq 0.76$, and $0.5 \cdot 10^{10}$ molec. $cm^{-3} \leq y_0 \leq 1.2 \cdot 10^{10}$ molec. cm^{-3}).

Figure 4.22 shows plots of the monthly-mean NO_2 profiles from MMF, MAPA, and NitroNet (with and without averaging kernels). The qualitative descriptions from the evaluation of WRF-Chem against FRM_4DOAS for May 2019 (see sect. 3.7.3 referring to Figs. 3.17 and 3.18) extend to the results shown here and are not repeated. Instead, the most important differences are briefly discussed:

- The AirBase instrument "NL00643" in De Bilt is not available in 2022.
- In 2019 the NO_2 concentrations from MAPA were either considerably larger or on the same scale as those from MMF (depending on location, see 3.17). In 2022 MAPA tends

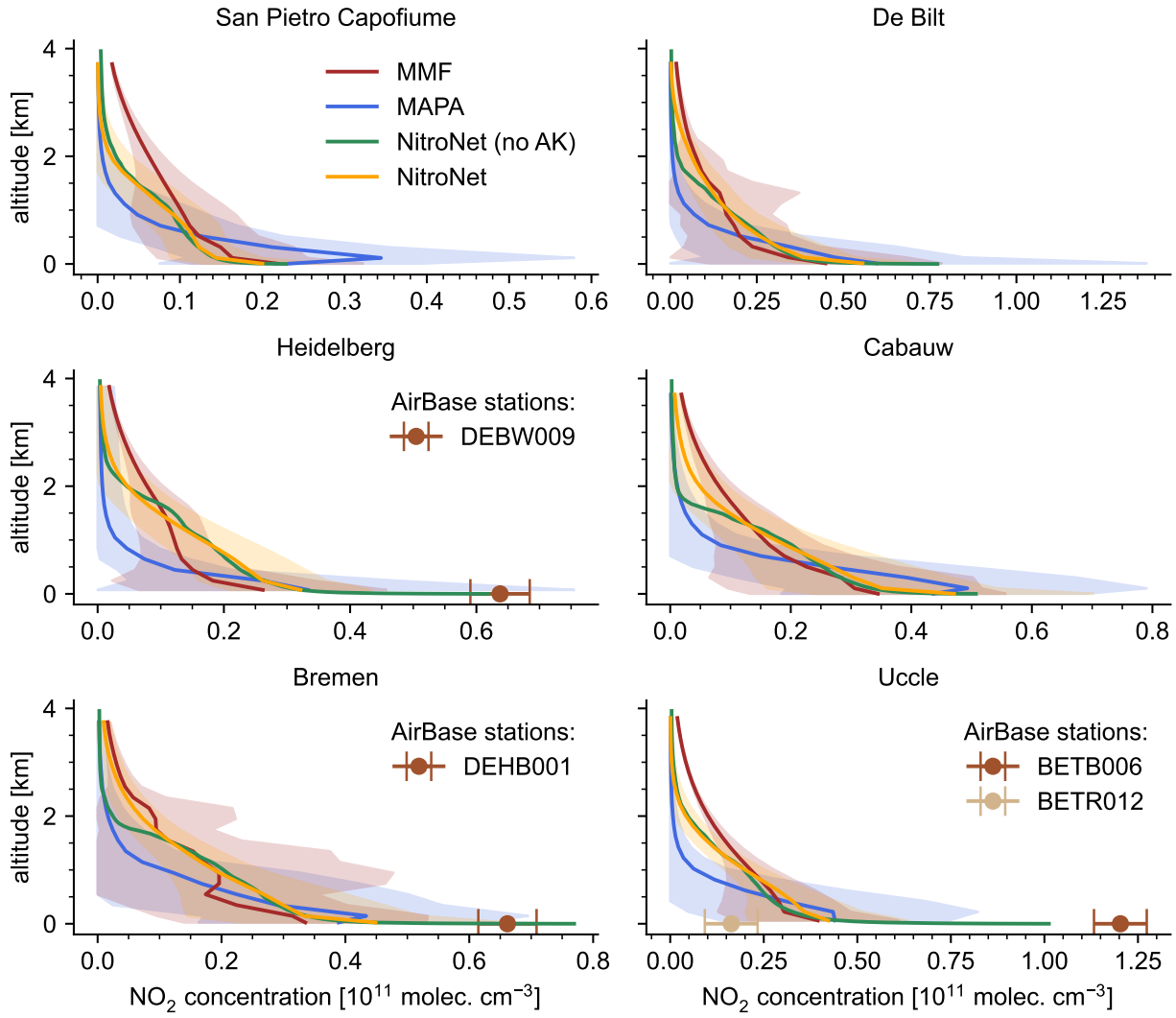

Figure 4.22: Comparison of full monthly-mean NO_2 profiles from FRM_4DOAS MAX-DOAS measurements to corresponding NitroNet predictions, May 2022. The monthly standard deviations are drawn as shaded regions. Colocated AirBase background measurements within a 5 km radius are drawn at an altitude of 0 m. The Mo-CL bias correction was applied.

to produce higher NO_2 concentrations than MMF in the lowest few hundred meters, but smaller concentrations above. The NitroNet predictions are in between, resulting in the "S"-shaped distribution of the scatter points in Fig. 4.21. Overall, as evident from Table 4.10, NitroNet's mean bias and RMSE are much lower with respect to MMF than to MAPA. This was also the case for the comparison of WRF-Chem and FRM_4DOAS in 2019 (see entry "YSU-2-5-B" in Table 3.9), although there NitroNet showed a negative mean bias towards MAPA. In 2022, NitroNet's mean bias towards MAPA is of similar magnitude, but positive. The RMSE values are lower than in the comparison between WRF-Chem and FRM_4DOAS in 2019, partly due to the lower magnitude of the NO_2 profiles in most locations (see also the point below).

- The MAPA results in the two new measurement locations, San Pietro Capofiume and Cabauw, as well as in Bremen, are characterized by a strong exponential gradient and an optional peak in the 2nd or 3rd layer above the ground. This profile shape was also discussed by Beirle et al. (2019), and might relate to the presence of elevated NO_2 layers or horizontal NO_2 gradients. WRF-Chem's incapability of reproducing most elevated layers was discussed in sect. 3.7.3, and appears to extend to NitroNet. Regardless, the concentration spikes produced by MAPA in 2022 are considerably larger than in 2019 and might reflect an incompatibility between the true NO_2 profile shape and MAPA's profile parametrization (technically a form of "model misspecification error").

Overall, much like in the validation of WRF-Chem in sect. 3.7.3, the evaluation of NitroNet against FRM_4DOAS MAX-DOAS measurements is associated with large retrieval uncertainties. Moreover, a validation in only six locations cannot be considered representative for extended spatial domains. Within these limitations, the evaluation reveals no fatal discrepancies, but also provides no more than a coarse validation of NitroNet's profile shapes and amplitudes.

4.3.3 Regional and seasonal validation study

NitroNet has shown convincing results in the evaluation studies of the previous two sections. This verifies that the model works as intended when used in a similar geospatial domain it was trained on (i.e. summertime in central Europe). Good generalization to other geospatial domains is not guaranteed, and requires that NitroNet manages to accurately extrapolate from the training set to entirely different physical scenarios. For example, the atmosphere is characterized by much longer NO_2 lifetimes of up to $\gtrsim 20$ h during winter, which are not necessarily reflected in NitroNet's training data. Nonetheless, it is plausible that NitroNet has learned some physical relations of general validity, which extend to other geospatial domains. This section presents a regional validation study on five additional geographic domains:

- UK, including the entire United Kingdom, Ireland, and the surrounding North Sea and Celtic Sea (contains AirBase instruments).

- ES + PT, including Portugal, Spain, and the Alboran Sea (contains AirBase instruments).

- US, covering approximately the eastern third of the United States of America and some of the Atlantic Ocean, excluding Florida.

- India, including parts of Pakistan, Nepal and Tibet, as well as the Arabian Sea and the Bay of Bengal.

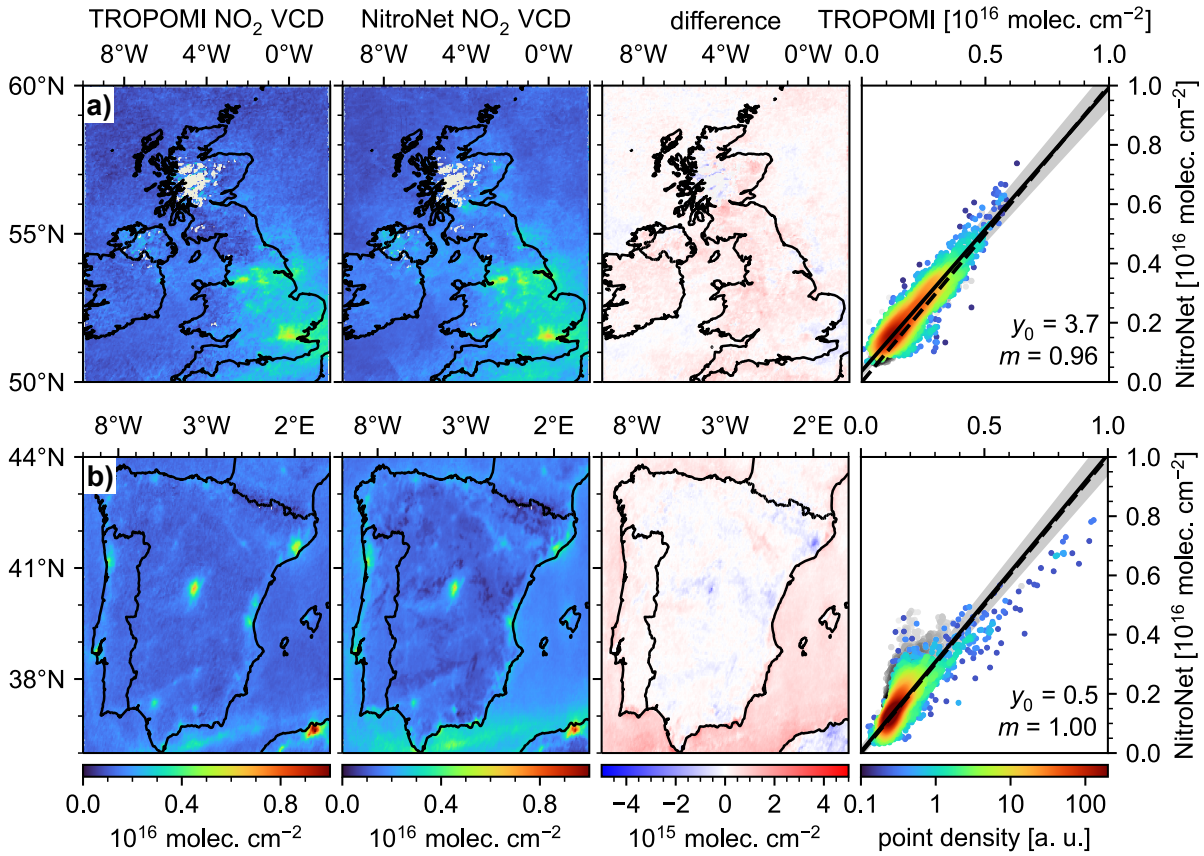

Figure 4.23: Evaluation of monthly-mean tropospheric NO_2 VCDs from NitroNet against TROPOMI satellite measurements, May 2022. Domains shown here: **a)** UK and **b)** ES + PT. Water-pixels are drawn as gray dots in the right-side scatter plots and excluded in the computation of the statistical analyses in Tables 4.12 and 4.13. The right-side scatter plots were drawn on the basis of 10000 random samples. Intercepts (y_0) are given in units of 10^{14} molec. cm^{-2}.

- China, covering approximately its eastern half from the southern border up to Beijing, including parts of the East China Sea.

This regional study is conducted for the month of May 2022. Additionally, a full-year seasonal evaluation (August 2021 to July 2022) is conducted on the central European domain used in the previous sections.

Regional validation study

The results of the regional validation study are presented first. Figure 4.23 shows the comparison between tropospheric NO_2 VCDs from NitroNet and TROPOMI satellite observations for the UK and the Mediterranean region of Spain and Portugal. The corresponding statistical diagnostics are found in Tables 4.12 (based on monthly mean data) and 4.13 (based on individual orbits/hourly measurements). As before, the following text refers to the evaluation on

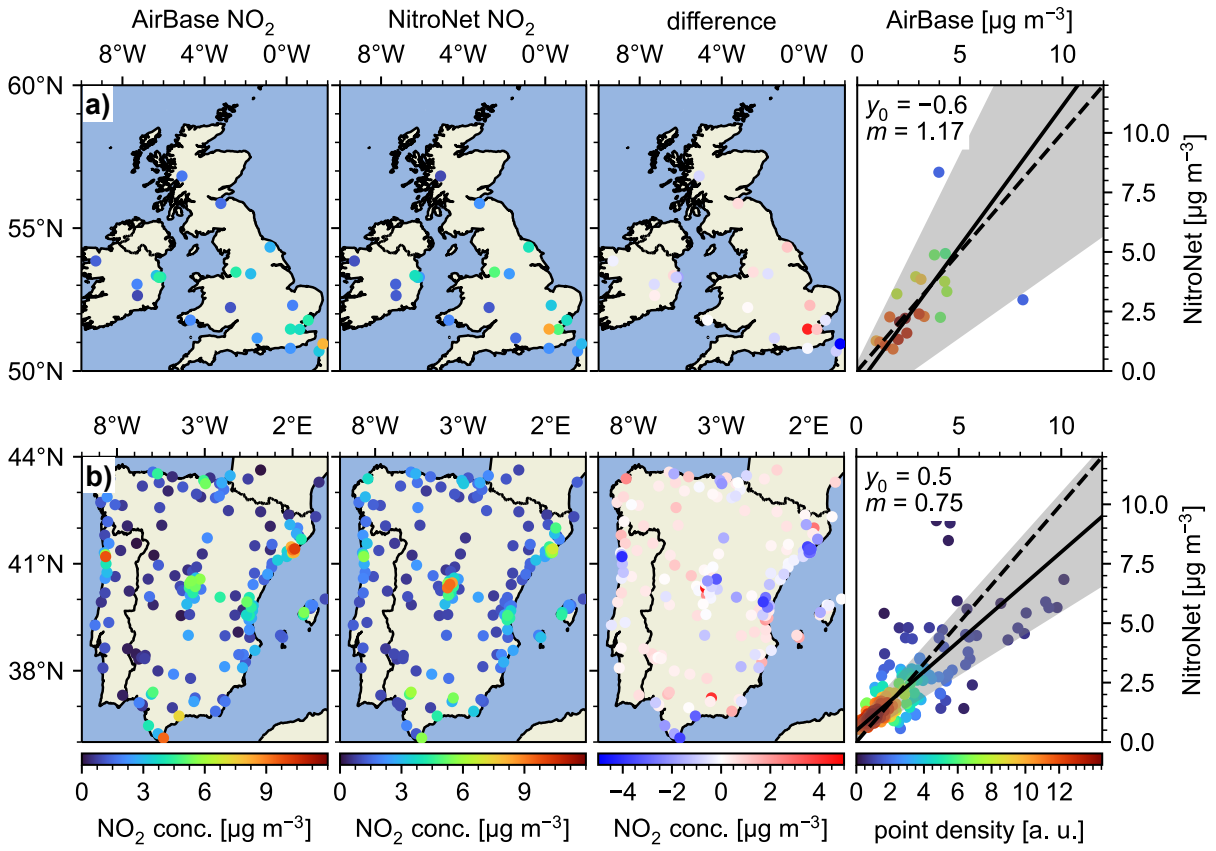

Figure 4.24: Evaluation of monthly-mean NO_2 surface concentrations from NitroNet against AirBase in situ measurements, May 2022. Urban background stations were excluded. Intercepts (y_0) are given in units of µg m^{-3}. Further statistical diagnostics are given in Tables 4.12 and 4.13.

monthly means, i.e. the entries of Table 4.12. In the validation against TROPOMI, the level of agreement is slightly reduced, but nonetheless comparable to that on the central European domain, as discussed in sect. 4.3.2. The RMSE values are increased from $2.8 \cdot 10^{14}$ molec. cm^{-2} (EU) to $4.3 \cdot 10^{14}$ molec. cm^{-2} (UK) and $3.1 \cdot 10^{14}$ molec. cm^{-2} (ES + PT), while the correlation coefficients are reduced from $R = 0.95$ (EU) to $R = 0.92$ (UK) and $R = 0.86$ (ES + PT). A noteworthy difference is NitroNet's tendency to overestimate the tropospheric NO_2 column over water by approximately 10^{15} molec. cm^{-2} in Fig. 4.23b. These overestimations might result from a lack of suitable training examples over water due to the data filtering based on relative criteria (see sect. 4.2.2), and could potentially be resolved in the future by the use of absolute filter criteria instead. As the evaluation over water significantly skews the overall results, it was decided to reduce its impact on the analysis by drawing the corresponding scatter points in the rightmost panel of Fig. 4.23 in gray colour. Additionally, all water pixels were excluded in the computation of the statistical diagnostics presented in Tables 4.12 and 4.13. A supplementary evaluation including the water pixels is found in Table C.3.

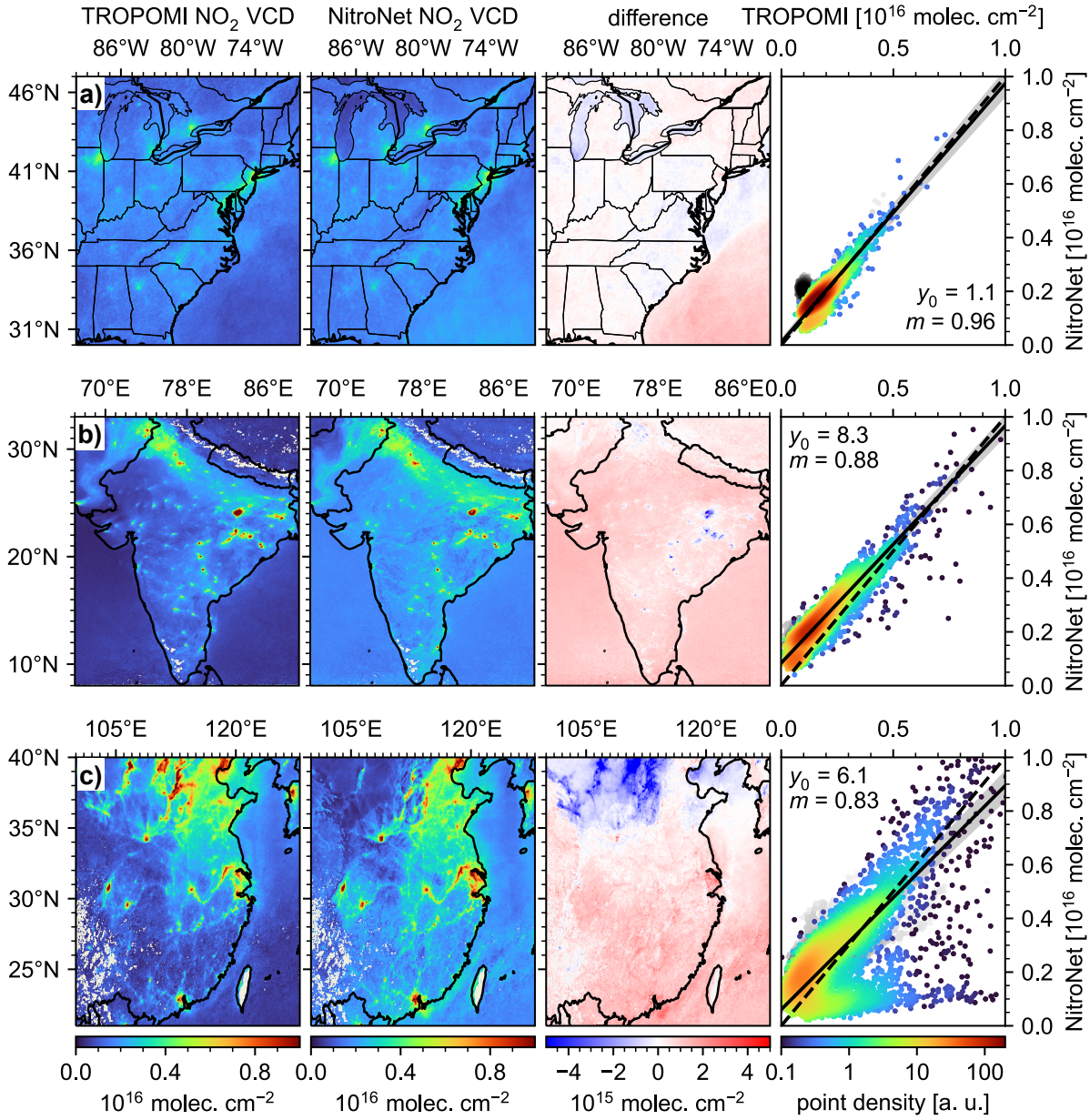

Figure 4.25: Like Fig. 4.23 but for the USA, India and China.

Figure 4.24 shows the corresponding evaluation against NO_2 surface concentrations from AirBase in situ measurements. For the reasons given in the previous sections, the urban background instruments were removed. A version of the figure with urban background stations included is found in Fig. C.5. Compared to the central European domain the RMSE is slightly increased from 1.2 µg m^{-3} to 1.8 µg m^{-3} (UK) and 1.6 µg m^{-3} (ES + PT), and the correlation coefficients reduced from $R = 0.73$ to $R = 0.45$ (UK) and $R = 0.71$ (ES + PT). The significantly weaker correlation on the UK domain can be attributed to the two outliers in its south-eastern corner combined with a relatively low amount of total instruments (21).

The domains shown so far are geographically quite close to NitroNet's original training domain. Figure 4.25 shows an evaluation against TROPOMI data for the three remaining, more distant domains. Good agreement is obtained for the US west coast (RMSE $= 2.7 \cdot 10^{14}$ molec. cm^{-2}, bias $= +2.7$ %, $R = 0.84$). The Indian domain shows even stronger correlation, but significant systematic overestimations (RMSE $= 8.0 \cdot 10^{14}$ molec. cm^{-2}, bias $= +41.5$ %, $R = 0.91$). The largest differences occur on the Chinese domain (RMSE $= 12.6 \cdot 10^{14}$ molec. cm^{-2}, bias $= +12.5$ %, $R = 0.70$). Although NitroNet captures a few pollution hotspots (e.g. Shanghai and Hong Kong), the polluted areas in the northern Chinese provinces Shanxi and Shaanxi are not reproduced well. These regions contain many facilities of the coal, steel, chemical and military industry and have undergone rapid economic and infrastructural development over the past years (see e.g. Peng et al., 2023). It is therefore plausible, that the EDGARv5 emission data of the year 2015 might already be outdated in these regions. Additionally, India and China are characterized by possibly different atmospheric conditions (mostly due to higher aerosol loads and larger NO_2 concentrations in the free troposphere), which NitroNet may struggle to interpret adequately.

Seasonal validation study (August 2021 – July 2022)

In this final section, a seasonal validation of NitroNet (August 2021 – July 2022) is conducted on the original central European domain. First, NitroNet is evaluated against tropospheric NO_2 VCDs from TROPOMI and NO_2 surface concentrations from AirBase measurements. Figure 4.26 shows the corresponding time series of the mean bias, RMSE, and correlation coefficient, as obtained from daily and monthly mean data. Additionally, the diagnostics obtained from the WRF-Chem simulation in May 2019 (see Table 4.7) are represented as dotted horizontal lines. The figure excludes all urban background AirBase instruments for the previously established reasons. A version of the figure with urban background stations included is found in Fig. C.6. The term "monthly mean" is used analogously to the previous sections of this thesis. For example, "monthly mean bias" refers to the *bias computed on monthly mean data*, not the monthly mean of daily biases (which can be obtained from averaging the daily values presented in Fig. 4.26). The same holds for the RMSE and the correlation coefficient. All text in the following refers to the monthly mean data, unless specified otherwise.

NitroNet's performance shows a clear annual cycle. From April to September, the results are comparable to those from May 2022 (see sect. 4.3.2). In all months except for November, December, and January, NitroNet' RMSE with respect to TROPOMI and AirBase is lower than the RMSE of WRF-2019 in May 2019. During winter NitroNet's NO_2 VCDs and surface concentrations show a distinct negative bias, peaking approximately in December with values of up to -22.6 % (vs. TROPOMI) and -49.6 % (vs. AirBase). Correspondingly, the

NitroNet compared to	domain	bias	RMSE	R	reference
TROPOMI	UK	+16.6 %	$4.3 \cdot 10^{14}$ molec. cm^{-2}	0.92	Fig. 4.23a
AirBase		−6.2 %	2.3 µg m^{-3}	0.57	Fig. C.5a
AirBase (w/o urban)		−4.5 %	1.8 µg m^{-3}	0.45	Fig. 4.24a
TROPOMI	ES + PT	+3.4 %	$3.1 \cdot 10^{14}$ molec. cm^{-2}	0.86	Fig. 4.23b
AirBase		−15.8 %	1.9 µg m^{-3}	0.80	Fig. C.5b
AirBase (w/o urban)		−5.4 %	1.6 µg m^{-3}	0.71	Fig. 4.24b
TROPOMI	USA	+2.7 %	$2.7 \cdot 10^{14}$ molec. cm^{-2}	0.84	Fig. 4.25a
TROPOMI	India	+41.5 %	$8.0 \cdot 10^{14}$ molec. cm^{-2}	0.91	Fig. 4.25b
TROPOMI	China	+12.5 %	$12.6 \cdot 10^{14}$ molec. cm^{-2}	0.70	Fig. 4.25c
TROPOMI (dec.)	EU	−21.8 %	$11.2 \cdot 10^{14}$ molec. cm^{-2}	0.89	Fig. 4.27a
AirBase (dec.)		−52.2 %	7.5 µg m^{-3}	0.81	Fig. 4.27b
AirBase (dec., no urban)		−49.6 %	6.4 µg m^{-3}	0.82	Fig. 4.27c
TROPOMI (dec., no wins.)	EU	−32.5 %	$14.1 \cdot 10^{14}$ molec. cm^{-2}	0.72	Fig. C.7a
AirBase (dec., no wins.)		−62.5 %	5.4 µg m^{-3}	0.77	Fig. C.7b
AirBase (dec., no wins., no urban)		−60.6 %	4.6 µg m^{-3}	0.74	Fig. C.7c

Table 4.12: Statistical summary of the regional-seasonal evaluation study based on monthly-mean data. All water pixels were excluded. “dec.” refers to an evaluation in December 2021. “no wins.” refers to evaluation without winsorization. “w/o urban” refers to evaluations in which AirBase instruments classified as “urban background” were excluded. All other entries refer to regular evaluations for the month of May 2022.

NitroNet compared to	domain	bias	RMSE	R	reference
TROPOMI	UK	+13.7 %	$6.7 \cdot 10^{14}$ molec. cm^{-2}	0.87	Fig. 4.23a
AirBase		−4.1 %	4.2 µg m^{-3}	0.36	Fig. C.5a
AirBase (w/o urban)		−12.8 %	3.6 µg m^{-3}	0.33	Fig. 4.24a
TROPOMI	ES + PT	+4.4 %	$6.6 \cdot 10^{14}$ molec. cm^{-2}	0.69	Fig. 4.23b
AirBase		−15.4 %	3.4 µg m^{-3}	0.59	Fig. C.5b
AirBase (w/o urban)		−3.2 %	2.6 µg m^{-3}	0.55	Fig. 4.24b
TROPOMI	USA	+3.1 %	$6.8 \cdot 10^{14}$ molec. cm^{-2}	0.72	Fig. 4.25a
TROPOMI	India	+40.8 %	$11.3 \cdot 10^{14}$ molec. cm^{-2}	0.82	Fig. 4.25b
TROPOMI	China	−7.1 %	$20.4 \cdot 10^{14}$ molec. cm^{-2}	0.63	Fig. 4.25c
TROPOMI (dec)	EU	−21.6 %	$18.5 \cdot 10^{14}$ molec. cm^{-2}	0.81	Fig. 4.27a
AirBase (dec)		−52.3 %	8.6 µg m^{-3}	0.70	Fig. 4.27b
AirBase (dec, no urban)		−49.9 %	7.8 µg m^{-3}	0.73	Fig. 4.27c
TROPOMI (dec., no wins.)	EU	−27.3 %	$30.6 \cdot 10^{14}$ molec. cm^{-2}	0.61	Fig. C.7a
AirBase (dec., no wins.)		−58.7 %	6.6 µg m^{-3}	0.71	Fig. C.7b
AirBase (dec., no wins., no urban)		−56.1 %	5.8 µg m^{-3}	0.71	Fig. C.7c

Table 4.13: Like Fig. 4.12, but based on individual orbits/hourly data.

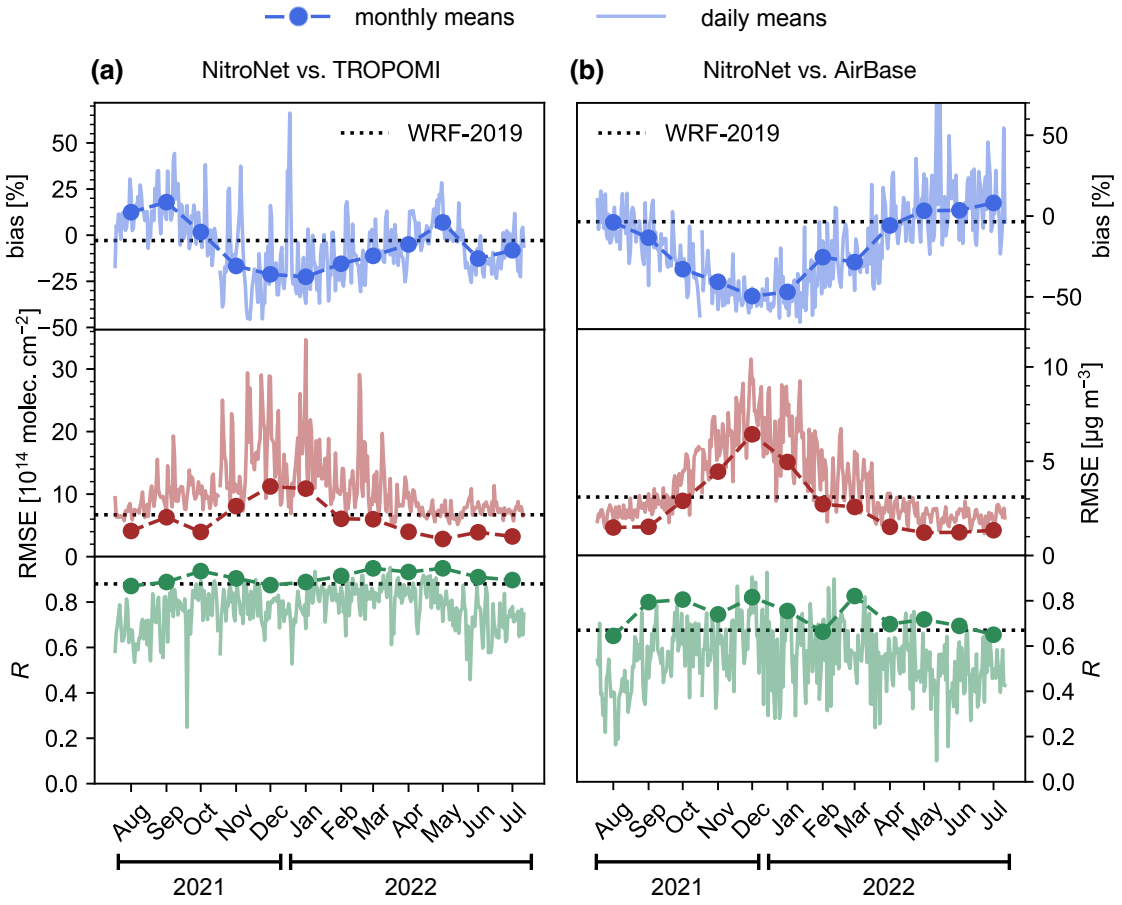

Figure 4.26: Seasonal evaluation of NitroNet on the central European domain from August 2021 – July 2022. **a)** Validation against tropospheric NO_2 VCDs from TROPOMI. **b)** Validation against NO_2 surface concentrations from AirBase. The dotted horizontal lines represent the monthly-mean statistical diagnostics (bias, RMSE, and R) of the WRF-2019 dataset (see Table 4.7). Urban background stations excluded.

RMSE is higher, reaching values of $11.2 \cdot 10^{14}$ molec. cm^{-2} (vs. TROPOMI) and 6.4 µg m^{-3} (vs. AirBase). The correlation coefficients, on the other hand, are stable around $R \approx 0.90$ (vs. TROPOMI) and $R \approx 0.70$ (vs. AirBase), and even slightly larger in wintertime. The diagnostics obtained from daily data are more noisy, resulting in larger RMSE values and lower correlation coefficients, following similar annual patterns as the monthly diagnostics. Because NitroNet is trained exclusively on summertime data, the lapse in wintertime prediction accuracy is not unexpected. In particular, NitroNet is uninformed about the significantly lower oxidative capacity and the correspondingly longer NO_2 lifetimes of $\gtrsim$ 20 h in winter (Liu et al., 2016; Shah et al., 2020; Lange et al., 2022). Figure 4.27 shows exemplary map plots against TROPOMI and AirBase data for December 2021. Figure C.7 shows a version of Fig. 4.27, but without application of the winsorization method described in sect. 4.2.3. As expected, the results are significantly worse (e.g. the RMSE of the monthly-mean NO_2 VCDs is increased from $11.2 \cdot 10^{14}$ molec. cm^{-2} to $14.1 \cdot 10^{14}$ molec. cm^{-2}, and the correlation

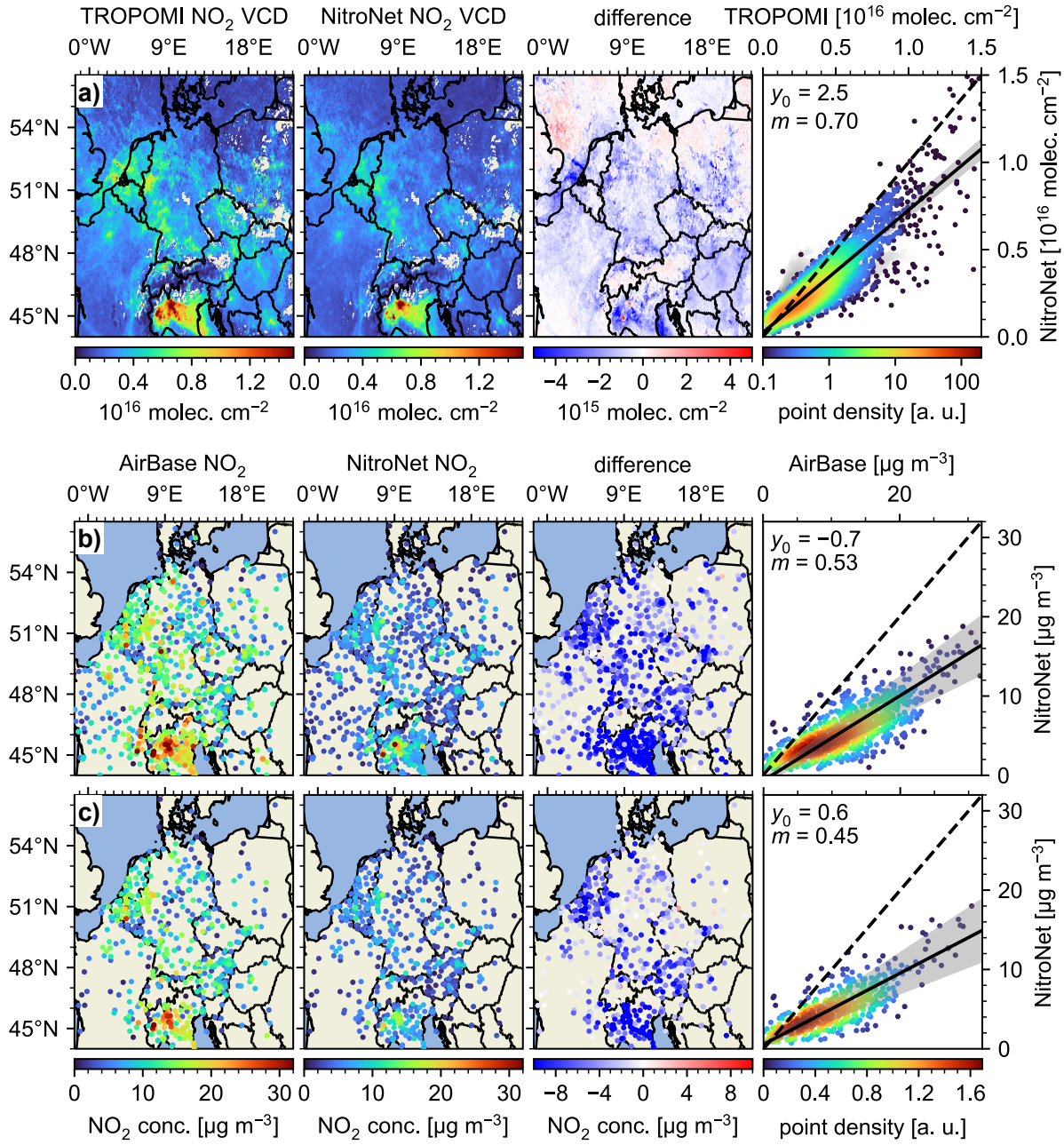

Figure 4.27: Like Fig. 4.20, but for December 2021. A statistical summary is found in Tables 4.12 and 4.13.

coefficient reduced from $R = 0.89$ to $R = 0.72$). Further diagnostics are found in Table 4.12. For reference, with winsorization turned on, approximately 8 % of the wintertime input data are affected (i.e. replaced by samples from the prior). On the other spatiotemporal domains shown in this study, the proportion of affected input data is considerably smaller ($\lesssim 3$ %).

Figure 4.28 shows a full-year evaluation of NitroNet against FRM_4DOAS data. This evaluation follows the descriptions given in the previous sect. 4.3.2, except that the mean bias (left column) and absolute error (middle column) of NitroNet are averaged over all instrument

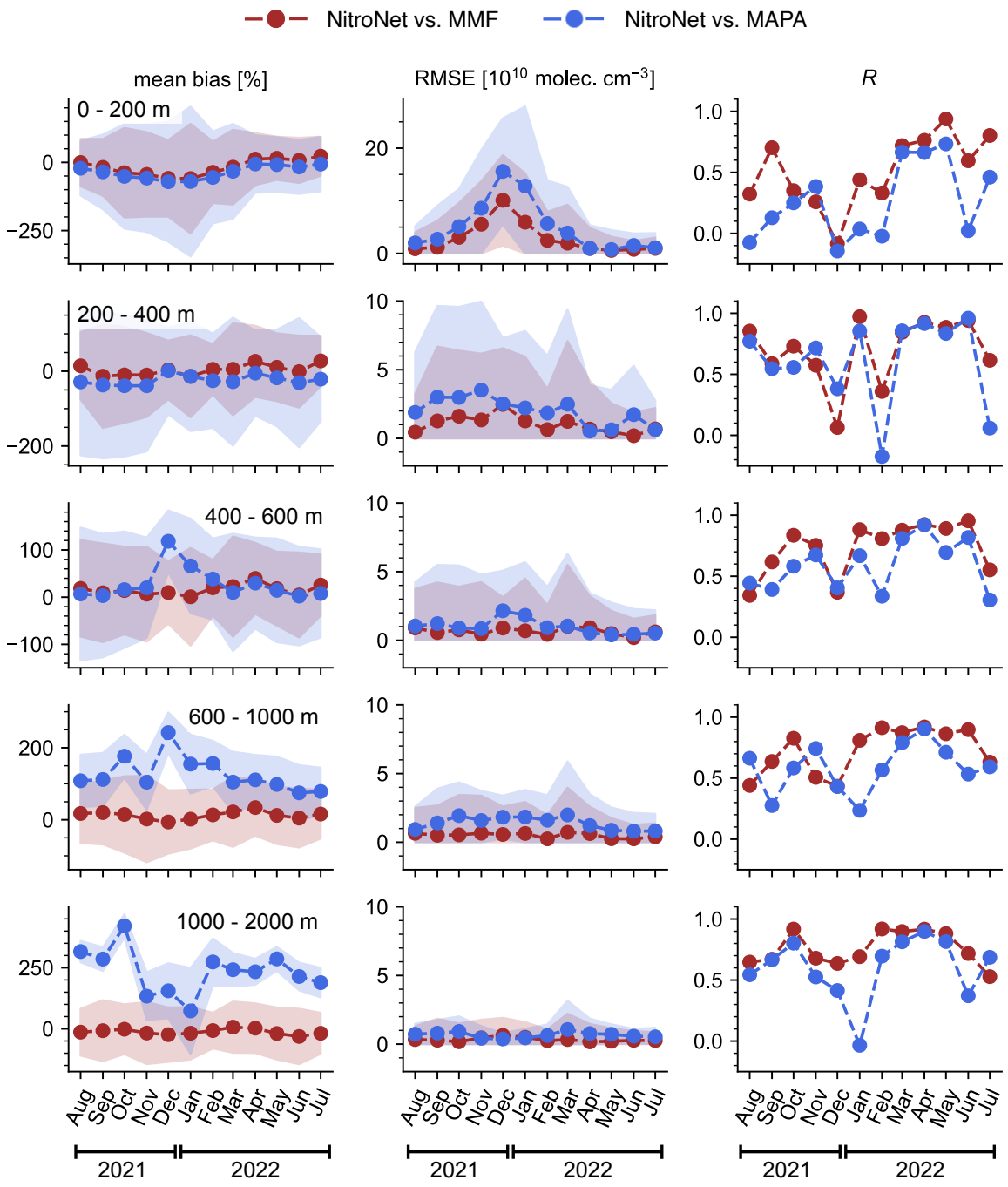

Figure 4.28: Seasonal evaluation of NitroNet on the central European domain against NO_2 concentrations from the FRM_4DOAS dataset. Shown here are NitroNet's monthly-mean bias, RMSE, and *R*, averaged over all available MAX-DOAS instruments in selected altitude ranges. The shaded background regions indicate the uncertainty of the displayed monthly-mean diagnostics.

locations. The shaded background regions indicate the uncertainty of the displayed monthly mean values (not the range of daily values, like e.g. in Fig. 4.26). A version without error bands of the mean bias was produced for easier readability (see Fig. C.8). A figure showing daily values is discussed later. The lowest evaluation layer (0 – 200 m) is characterized by particularly good agreement between MAPA and MMF, and large NitroNet biases between approx. −70 % and +20 %. Similar to the seasonal evaluation against TROPOMI and AirBase, the largest negative biases and RMSE values occur during winter. In some months, the correlation coefficient with respect to MAPA even drops to 0, or slightly below. Note, that NitroNet's correlation to MAPA was found to be considerably weaker than that to MMF in previous parts of the analysis (see e.g. Table 4.10). The seasonal trends described earlier are represented well in this comparison to MAX-DOAS data, although only in the lowest layer ($0 - 200$ m), which indicates that they relate to the lower regions of the troposphere. In the higher evaluation layers, particularly above 600 m, the comparison to MMF yields moderate biases (mostly between −30 % and +30 %), while the comparison to MAPA results in biases of up to $\gtrsim 100$ % ($600 - 1000$ m) and $\gtrsim 200$ % ($1000 - 2000$ m). This aligns well with the profile shapes observed in sect. 4.3.2. Although mean biases of 200 % or more may appear concerning, they should be put into perspective based on the following considerations:

1. As demonstrated e.g. by Fig. 2.13 in sect. 2.3.5, the sensitivity of the MAX-DOAS measurements is significantly reduced beyond ~ 1500 m, increasing the retrievals' dependence on a priori assumptions. Although MAPA does not depend on explicit a priori information in the form of a priori profiles (see Beirle et al., 2019), it nonetheless prescribes a mostly exponential profile shape at higher altitudes, resulting in correspondingly low NO_2 concentrations.

2. The mean (relative) bias as defined in eq. (3.1) can be a misleading error metric, if the reference values are sufficiently small (see the low RMSE values for MAPA at $1000-2000$ m). Here, NitroNet's high relative biases with respect to MAPA correspond to rather moderate absolute errors of $\sim 10^{10}$ molec. cm^{-3}.

3. The differences between MAPA and NitroNet are approximately on the same scale as the differences between MAPA and MMF. This highlights the considerable retrieval uncertainties, in regard of which the intercomparison between NitroNet, MMF, and MAPA should be given correspondingly little weight at higher altitudes.

Figure 4.29 shows a combined evaluation in the lowest 2 km, including daily diagnostics. As in Fig. 4.28, the values given here represent averages with respect to all MAX-DOAS instruments operated by FRM_4DOAS. The results are qualitatively similar to those from the seasonal evaluation against TROPOMI and AirBase in Fig. 4.26. In particular, the

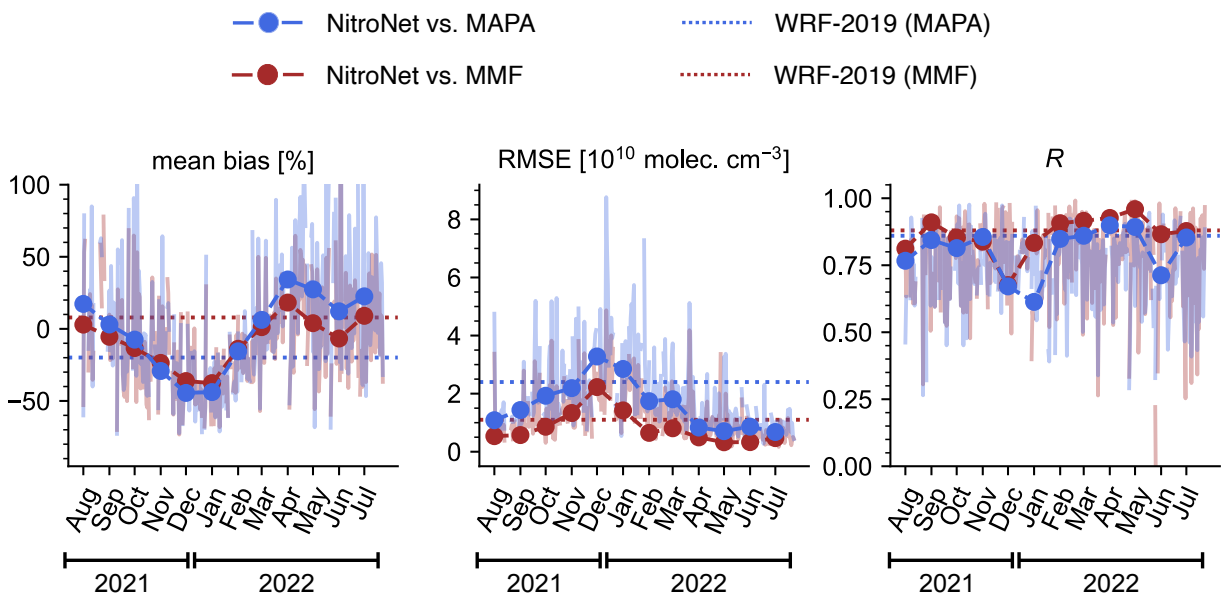

Figure 4.29: Seasonal evaluation of NitroNet on the central European domain against NO_2 concentrations from the FRM_4DOAS dataset in the lowest 2 km. The dotted horizontal lines represent the monthly-mean statistical diagnostics (bias, RMSE, and *R*) of the WRF-2019 dataset (see Table 3.12).

annual cycles of the mean bias (monthly means reaching down to −44.4 % during winter) and the RMSE (monthly means peaking at up to $3.3 \cdot 10^{10}$ molec. cm^{-3}) are in good qualitative agreement to the corresponding cycles from the evaluation against AirBase. Note, that the correlation coefficients displayed in Fig. 4.29 are higher than those in Fig. 4.28 due to the inclusion of data from overall more FRM_4DOAS retrieval layers.

This concludes the regional-seasonal evaluation of NitroNet. The obtained results are summarized in a final concluding overview in the following section, see Table 4.14.

4.4 Summary, discussion, and conclusions

This chapter has presented NitroNet, a feed-forward neural network for the prediction of NO_2 concentration profiles based on TROPOMI satellite data and other ancillary input variables. NitroNet was implemented in the Python programming language, and designed to operate on high performance computing clusters, particularly with parallel usage of multiple GPUs. It computes NO_2 profiles on arbitrary user-chosen vertical grids in the troposphere.

Training

NitroNet was trained on a synthetic dataset of NO_2 profiles from a WRF-Chem simulation ("WRF-2019") on a central European domain for the month of May 2019. Prior to training, the dataset was filtered based on the agreement to reference tropospheric NO_2 VCDs from TROPOMI (relative errors < 20 % allowed) and the planetary boundary layer heights from the ERA5 reanalysis (relative errors < 10 % allowed). As a result of this data filtering, only ~ 7 % of the original training data came into use. The neural network's hyperparameters were optimized using the random search strategy, in the scope of which over 100 network configurations were intercompared. The formal evaluation on the test set revealed an excellent prediction accuracy (relative errors of $\lesssim 10$ % within the planetary boundary layer) with no significant biases if the same data filters were applied as during training. Based on the combined results of Chapter 3, this indicates that the errors between NitroNet and WRF-Chem are far lower than those between WRF-Chem and the observational reference data. Without application of the data filters much larger errors occured on the test set, although these partly relate to NitroNet not fully reproducing WRF-Chem's systematic errors, which is overall favourable. Additionally, a slight prediction bias of approx. -10 % was identified, based on which an altitude-dependent bias correction look-up table was implemented. The feature relevances of NitroNet's input variables were determined by approximating their Shapley scores. The tropospheric NO_2 VCD from TROPOMI was found to be the most important input variable (average relevance: 30.9 %, but 25 % at the surface), followed by the EDGARv5 NO_x emission data (average relevance: 8.9 %, but 18 % at the surface) and the ERA5 boundary layer height (average relevance: 6.9 %, but up to 16 % at ~ 1700 m). Previous studies had estimated the feature relevance of the tropospheric NO_2 VCD for the prediction of surface NO_2 concentrations in the range of 15 % – 30 % (see Kang et al., 2021; Chan et al., 2021; Ghahremanloo et al., 2021; Zhang et al., 2022).

NitroNet vs. WRF-Chem

NitroNet and WRF-Chem were intercompared on the central European training domain of May 2019. The NO_2 profiles of both models were used to compute simulated tropospheric NO_2 VCDs and NO_2 surface concentrations, which were then compared to reference data from TROPOMI and background instruments of the AirBase network. In this process, the air mass factors of the TROPOMI NO_2 VCDs were re-computed using the horizontally higher resolved NO_2 profiles from WRF-Chem and NitroNet, respectively. The Mo-CL biases of AirBase instruments, using the molybdenum-based chemiluminescence method, were corrected using the volume mixing ratios of PAN and HNO_3 from WRF-Chem and corresponding correction factors predicted by NitroNet. The statistical diagnostics given in the following were computed

on monthly-mean data. The NO_2 VCDs from NitroNet were found to agree significantly better with the TROPOMI reference data than those from WRF-Chem (mean bias: -8.1 % vs. -2.9 %, RMSE: $6.7 \cdot 10^{14}$ molec. cm^{-2} vs. $3.8 \cdot 10^{14}$ molec. cm^{-2}, R: 0.97 vs. 0.88). In particular, NitroNet did not reproduce the considerable high biases of WRF-Chem in highly polluted regions. These improvements are a consequence of the aforementioned training data filtering and demonstrate its essential benefits in the training of NitroNet. Furthermore, the use of the NO_2 VCD as the main input variable is informative of the total tropospheric NO_2 load, thereby constraining the amplitude of the predicted NO_2 profiles. The intercomparison to the AirBase in situ measurements was less conclusive, with negligible performance differences between NitroNet and WRF-Chem. Another major advantage of NitroNet over WRF-Chem are the significantly shorter runtimes. To put this into perspective: The computation of one month of training data with WRF-Chem took ~ 5 days with ~ 800 CPUs. NitroNet can process the same amount of data in just ~ 20 minutes using 31 GPUs, with obvious operational advantages. Of course this functionality is limited to the prediction of NO_2 profiles and the Mo-CL bias correction factors, hence why NitroNet cannot fully replace WRF-Chem.

Regional-seasonal validation study

A detailed validation study of NitroNet was conducted on data from May 2022 on various geographic domains (central Europe, the UK, Spain and Portugal, India, and parts of the US and China). For that purpose, NitroNet was evaluated against tropospheric NO_2 VCDs from TROPOMI, background in situ measurements of surface NO_2 concentrations from AirBase, and NO_2 profiles from FRM$_4$DOAS MAX-DOAS measurements (only available in central Europe). The results varied strongly across the individual evaluation domains. In central Europe, the UK, Spain and Portugal, and the US, NitroNet showed similar prediction accuracy with respect to TROPOMI as on the original training domain (central Europe, 2019). In India and China, NitroNet's prediction errors were significantly larger (e.g. RMSE values of up to $12.6 \cdot 10^{14}$ molec. cm^{-2}). On all domains except for the UK, NitroNet showed significant overestimations of the tropospheric NO_2 VCD over water ($\sim 10^{15}$ molec. cm^{-2}). This is most likely caused by a low amount of suitable training data due to the use of relative filter criteria in the curation of the training set. In the evaluation against AirBase in situ measurements, NitroNet's surface predictions were identified as low biased by approx. -10 %. This bias was attributed to in situ measurements classified as "urban background". The vague definition of this group of instruments hints towards a somewhat misleading classification terminology, and implies the inclusion of traffic-dominated instruments. Other than that, it can not be excluded that NitroNet's predicted NO_2 concentrations are too low in certain urban areas, although the comparison to TROPOMI data did not support this claim. In lack of more detailed information on the classification of the AirBase instruments, no definite conclusion

on this matter can be drawn.

In order to assess the model's seasonal generalization capabilities, the study was extended to a full year of data (August 2021 – July 2022), evaluated on the central European domain. Although NitroNet's prediction accuracy remained roughly stable between April and September, severe underestimations of the tropospheric NO_2 VCDs and surface concentrations were found during winter (up to -22.6 % wrt. TROPOMI, -49.6 % wrt. AirBase, -44.4 % wrt. MAX-DOAS measurements from FRM_4DOAS). The correlation coefficients remained stable throughout the entire year ($R \approx 0.90$ wrt. TROPOMI, $R \approx 0.70$ wrt. AirBase, $0.60 \lesssim R \lesssim 0.90$ wrt. MAX-DOAS measurements from FRM_4DOAS). A similar seasonal trend was identified in the comparison to MAX-DOAS measurements at low altitudes ($0 - 200$ m). These underestimations cannot be ignored and indicate that a neural network like NitroNet, trained on summertime data only, is limited in its capability to generalize to other seasons. However, RCT simulations, which are the only alternative source of NO_2 profiles with dense coverage on extended domains, were found to suffer from biases of similar magnitude. A selection of relevant studies on European domains includes:

- Blechschmidt et al. (2020), who validate the CAMS regional ensemble model against tropospheric NO_2 VCDs from MAX-DOAS measurements in Bremen, De Bilt, Uccle, and the *Observatoire de Haute-Provence* in France for the years 2010 – 2012. The individual sub-models were found to under/overestimate the reference NO_2 VCDs by approx. -50 % to $+100$ % or more, depending on the location, with a tendency towards overestimation in winter, and underestimation in summer. In some cases, even the ensemble prediction showed biases of the same magnitude.

- Douros et al. (2023), who evaluate the CAMS regional ensemble model for the years 2018 and 2019. The simulated NO_2 VCDs in winter were found to be high biased by up to $+50$ % compared to TROPOMI.

- Huijnen et al. (2010), who present a comparison study between simulated NO_2 VCDs from an ensemble of chemistry and transport models (MOZART, MOCAGE, and TM5) and the OMI satellite instrument for the years 2008 and 2009. The simulated NO_2 VCDs were found to be low biased by approx. -40 % compared to OMI during summer, but nearly unbiased in winter.

Within the scope of this thesis, it was also attempted to produce wintertime WRF-Chem simulation results (February 2019) with the simulation setup S-YSU-2-5-B described in Table 3.4. A validation of this simulation run against NO_2 VCDs from TROPOMI is found in Fig. C.9 and shows severe overestimations of the tropospheric columns by a factor ~ 2. However, not too much weight should be assigned to these preliminary results, because the simulation

NitroNet compared to	domain	bias [%]	RMSE	R
TROPOMI	central EU April – September	−12.7 **(−17.5)** to +17.9 **(+18.8)**	2.8 **(6.3)** to 6.3 **(10.6)**	0.87 **(0.61)** to 0.95 **(0.87)**
AirBase		−23.7 **(−30.4)** to −8.3 **(+6.1)**	1.7 **(2.3)** to 2.3 **(3.8)**	0.65 **(0.38)** to 0.76 **(0.63)**
AirBase (w/o urban)		−13.4 **(−20.0)** to +8.1 **(+24.2)**	1.2 **(1.7)** to 1.5 **(2.9)**	0.64 **(0.34)** to 0.79 **(0.64)**
FRM_4DOAS (MMF)		−6.7 **(−36.3)** to +18.2 **(+40.4)**	0.3 **(0.4)** to 0.6 **(2.1)**	0.81 **(0.55)** to 0.96 **(0.93)**
FRM_4DOAS (MAPA)		+3.0 **(−32.0)** to +34.2 **(+74.6)**	0.7 **(0.7)** to 1.4 **(3.4)**	0.71 **(0.54)** to 0.89 **(0.89)**
TROPOMI	central EU October – March	−22.6 **(−32.3)** to +1.7 **(+11.3)**	4.0 **(8.3)** to 11.2 **(20.8)**	0.87 **(0.71)** to 0.95 **(0.88)**
AirBase		−52.2 **(−58.2)** to −30.7 **(−25.2)**	3.4 **(4.0)** to 7.5 **(8.6)**	0.69 **(0.47)** to 0.81 **(0.73)**
AirBase (w/o urban)		−49.6 **(−55.8)** to −25.4 **(−17.1)**	2.6 **(3.1)** to 6.4 **(8.0)**	0.66 **(0.41)** to 0.82 **(0.75)**
FRM_4DOAS (MMF)		−37.6 **(−60.4)** to +1.2 **(+21.9)**	0.6 **(0.8)** to 2.2 **(3.7)**	0.68 **(0.59)** to 0.92 **(0.95)**
FRM_4DOAS (MAPA)		−44.4 **(−56.7)** to +6.2 **(+19.2)**	1.7 **(1.3)** to 3.3 **(5.5)**	0.61 **(0.62)** to 0.86 **(0.88)**
TROPOMI	other regions May	+2.7 **(−7.1)** to +41.5 **(+40.8)**	2.7 **(6.6)** to 12.6 **(20.4)**	0.70 **(0.63)** to 0.92 **(0.87)**

Table 4.14: Final summary of NitroNet's performance against reference data from TROPOMI, AirBase, and FRM_4DOAS. The black numbers represent the results obtained from monthly-mean data. The coloured bold numbers represent the results obtained without averaging (e.g. individual orbits from TROPOMI). For the central European domain, the lower and upper limits of the ranges given here were determined from the minimum and maximum values (for monthly means), or the 10 % – 90 % inter-quantile range (for unaveraged data), both along the time axis. Data for TROPOMI and AirBase were taken from Figs. 4.26 and C.6. Data for FRM_4DOAS were taken from Fig. 4.29 and refer to the lowest 2 km. For the evaluation of TROPOMI in all other regions, the lower and upper limits were estimated based on the minimum and maximum values across the different evaluation domains, with data taken from Tables 4.12 and 4.13. RMSE values are given in units of 10^{14} molec. cm^{-2} for TROPOMI, µg m^{-3} for AirBase, and 10^{10} molec. cm^{-3} for FRM_4DOAS.

setup was not yet optimized (e.g. with respect to the vertical mixing parametrization or the temporal emission profiles).

Based on the combined results of this regional-seasonal validation study, it can be concluded that NitroNet is principally capable of generalizing to new geographical domains and seasons. However, a loss of precision and/or considerable biases were found on some evaluation domains (e.g. in India and China) and on the central European domain during winter. Particularly in India and China such prediction errors might relate to inaccuracies in the NO_x emission data, for which a high feature relevance was found near the surface. Furthermore, much larger free tropospheric NO_2 concentrations must be expected in China. These are not represented in the "NO_2 influx" variable of NitroNet, which is computed based on surface wind speeds only. A final quantitative summary of the regional-seasonal validation study is given in Table 4.14, with data taken from Figs. 4.26, C.6, and 4.29, and Tables 4.12, and 4.13. For more details, refer to the table's caption. A visualization is found in Fig. 4.30.

Comparison to other machine learning models

The prediction of near-surface NO_2 concentrations by means of machine learning has become an active area of research over the past decade. Examples are found in Gardner and Dorling

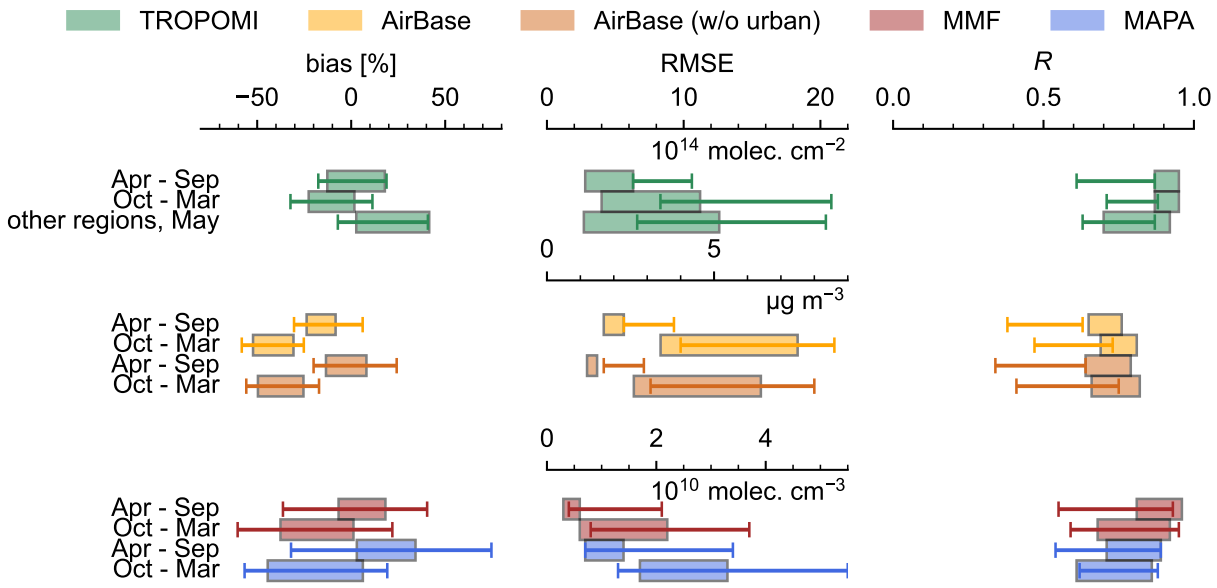

Figure 4.30: Visualization of Table 4.14. The thick bars/thin lines represent the diagnostics (bias, RMSE, *R*) computed based on monthly-mean/unaveraged data.

study	model type	domain	reference type	*R*	RMSE [µg m^{-3}]	FI[1]	Mo-CL bias	full NO_2 profiles
Gardner and Dorling (1999)	FF-NN[2]	London	all	0.69	34.2	N/A	✗	✗
Kang et al. (2021)	regression tree[3]	eastern Asia	all	0.84	9.1	30 %	✗	✗
Chan et al. (2021)	FF-NN	Germany	bgr.[5]	0.80	6.3	27 %	✗	✗
Ghahr.[7] et al. (2021)	CNN[4]	Texas	all	0.91	4.6	15 %	✗	✗
Zhang et al. (2022)	CNN	northern China	all	0.94	5.8	18 %	✗	✗
Cao (2023)	CNN	contiguous US	all	0.94	4.3	N/A	✗	✗
NitroNet	FF-NN	central EU	bgr.	0.38 – 0.63	2.3 – 3.8	31 %	✓	✓
NitroNet	FF-NN	central EU	bgr. w/o[6]	0.34 – 0.64	1.7 – 2.9	31 %	✓	✓

Table 4.15: Comparison of published machine learning models for the prediction of surface-level NO_2 concentrations against in situ reference data. All models use tropospheric NO_2 VCDs from TROPOMI as input data, except for Gardner and Dorling (1999), who use no satellite data and predict NO_2 concentrations at a single location. For publications in which different model types were tested, the best results are shown here. RMSE and *R* were computed on unaveraged data. For NitroNet, the results from Table 4.14 (central EU, April – September) are shown.

(1) feature importance of the tropospheric NO_2 VCD

(2) feed-forward neural network

(3) see Breiman et al. (2017)

(4) convolutional neural network

(5) background instruments

(6) background instruments without urban background stations

(7) Ghahremanloo et al. (2021)

(1999), Kang et al. (2021), Chan et al. (2021), Ghahremanloo et al. (2021), Zhang et al. (2022), and Cao (2023), see also sect. 2.5.8. The models proposed in these publications were not designed to predict full NO_2 profiles, and are limited to predictions of surface-level NO_2 concentrations learned from in situ measurements. An overview of the aforementioned literature and a comparison to NitroNet is given in Table 4.15. With the exception of Gardner and Dorling (1999), the referenced studies present models that can predict surface NO_2 concentrations with RMSE values of ~ 5 µg m^{-3} and correlation coefficients of $R \sim 0.90$. NitroNet, here evaluated on the central European domain from April to September using unaveraged data, shows significantly weaker correlation ($R = 0.34$ to $R = 0.64$, w/o urban background stations), but also lower RMSE values (1.7 µg m^{-3} to 2.9 µg m^{-3}). A similar correlation strength was already identified in the validation of WRF-2019 against AirBase in situ data ($R = 0.44$ for unaveraged data of May 2019, see Table 3.12). Therefore, NitroNet's reduced correlation to the AirBase in situ measurements must be attributed to the use of model-based training data, and does not indicate a deficiency of the neural network. This marks an important difference to the other models mentioned here which were trained on observational data alone, thereby evading the intermediate RCT simulation as a critical error source. The differences in RMSE must be interpreted under consideration of the different evaluation approaches among the referenced studies. Most importantly, except for Chan et al. (2021), none of the studies mention any filtering of the in situ instruments by type (background, traffic, industrial), thus it must be assumed that instruments of all types, possibly placed in heavily polluted traffic regions, were used. Furthermore, Kang et al. (2021) and Zhang et al. (2022) evaluated their models in highly polluted regions of China, which may lead to increased RMSE values. Therefore, the lower RMSE of NitroNet does not substantiate superiority over the competing models. Instead, the unique benefit of NitroNet lies in its ability to predict full NO_2 profiles and Mo-CL bias correction factors for molybdenum-based in situ measurements. Particularly the latter is important here, because the other models, trained on observational data, inevitably reproduce the measurement biases contained within their training data (often on the scale of 20 % to $\gtrsim 300$ %). The reported RMSE values do not account for this bias, as they were obtained by comparing bias-contaminated model predictions to bias-contaminated reference data. This marks a principal limitation of the purely observational training datasets of previous models, with little to no prospect of improvement.

Chapter 5

Summary, conclusion, and outlook

Summary

This thesis has presented NitroNet, a neural network for the prediction of tropospheric NO_2 concentration profiles on arbitrary vertical grids at a horizontal resolution of up to 3.5 km × 5.5 km. NitroNet uses TROPOMI satellite data and other ancillary variables as input and was trained on a dataset of modelled NO_2 profiles from the regional chemistry and transport (RCT) model WRF-Chem. The statistical diagnostics given in the following were computed on monthly-mean data. More detailed quantitative summaries are found in the conclusion sections of Chapters 3 and 4.

The first task in the development of NitroNet was to produce model-based training data using WRF-Chem on a central European domain for the month of May 2019. The simulation had a horizontal resolution of 3 km × 3 km. In the scope of a model evaluation study the WRF-Chem simulation results were validated against three observational datasets: tropospheric NO_2 VCDs from TROPOMI (representing the vertical integral of the concentration profile), in situ measurements of NO_2 surface concentrations from AirBase, and vertical NO_2 profiles from FRM_4DOAS MAX-DOAS measurements using the retrieval algorithms MAPA and MMF. Only AirBase instruments classified as "background" were used. Simulation results obtained with an original model configuration showed significant deviations from the observational data. Subsequently, a model optimization was conducted on the German model subdomain, where emission data of particularly high resolution ($0.01\,° \times 0.01\,°$) were available, with the aim to identify the underlying causes. Three modelling aspects were found to be crucial in this regard:

1. Exchange of the NO_2 a priori profiles provided by the TM5-MP model (horizontal resolution of $1\,° \times 1\,°$, used to compute the air mass factors of TROPOMI) by the far better resolved NO_2 profiles from WRF-Chem. This resulted in significantly larger reference NO_2 VCDs (as much as +30 % in polluted regions), and aligned well with previous

literature results.

2. The "Mo-CL bias" correction for molybdenum-based chemiluminescence in situ measurements. Previous literature had estimated these biases in the range of approximately +20 % to +300 % and attributed them to cross-sensitivities to other atmospheric trace gases such as PAN and HNO_3. Based on the simulated mixing ratios of these gases and estimates of the conversion rates in the widely used molybdenum cartridges, an average Mo-CL bias correction factor of $F = 1.22$ was determined on the German subdomain.

3. The treatment of vertical mixing in the boundary layer schemes of WRF-Chem. An inspection of the corresponding model code revealed a suspicious manipulation of the turbulent diffusion coefficients by clipping them to values of 1 $m^2\ s^{-1}$ and 2 $m^2\ s^{-1}$ in rural and urban model regions, respectively. Following the developers' recommendation, these clipping thresholds were iteratively tuned to the simulation domain. It was found that changing the rural clipping threshold to 5 $m^2\ s^{-1}$ resulted in far better overall agreement to the used reference data, indicating that the simulation had previously underestimated vertical mixing in rural regions.

With these corrections applied, the simulation results were evaluated on the full model domain. The simulation's RMSE values with respect to the three reference datasets were $7.9 \cdot 10^{14}$ molec. cm^{-2} (vs. TROPOMI), 5.1 µg m^{-3} (vs. AirBase at noontime) and up to $2.4 \cdot 10^{10}$ molec. cm^{-3} (vs. FRM_4DOAS from 11 AM – 2 PM, evaluated in the lowest 2 km). Furthermore, a mean noontime low bias of approx. −20 % remained with respect to the AirBase measurements. This bias was attributed to the model regions outside of Germany, on which emission data of far lower resolution ($0.1\,° \times 0.1\,°$) were used, hence why they were excluded from the aforementioned optimization study.

In the next step, the training dataset for NitroNet was assembled. For that purpose the NO_2 profiles from WRF-Chem (the targets) were colocated with TROPOMI observations (with a quality score of $f_{QA} > 0.75$) and further ancillary data variables from the ERA5 atmospheric reanalysis and the EDGARv5 emission inventory (the predictors). This dataset was then randomly split into a training set (80 %), a validation set (15 %), and a test set (5 %). Prior to training, a data filtering scheme was implemented, by means of which erroneous training examples were dismissed. More specifically, all NO_2 profiles whose simulated NO_2 VCD and boundary layer height disagreed with TROPOMI and ERA5 reference data by more than 20 % and 10 % were filtered out. This reduced the number of training profiles from $\sim$ 1800000 to $\sim$ 130000 ($\approx$ 7 %). Within the scope of a hyperparameter optimization, over 100 variants of feed-forward neural networks were trained (on the training set) and evaluated (on the validation set) with the aim to identify the ideal network configuration with respect to hyperparameters such as the network size, the learning rate, etc. Additionally, an ancillary

neural network (the "F-network") was trained to predict multiplicative correction factors for the aforementioned Mo-CL biases. Thereby, NitroNet's predictions at the surface could later be compared to the bias-contaminated in situ observations from AirBase. An assessment of how much the individual input variables contributed to the performance of the neural network was made based on estimates of their Shapley scores. The tropospheric NO_2 VCD from TROPOMI was identified as the most informative input variable (feature importance: 30.1 %), followed by NO_x emission data from EDGARv5 (8.9 %) and the planetary boundary layer height from ERA5 (6.9 %).

Lastly, the NitroNet model was extensively validated. A formal evaluation on the independent test set revealed far lower errors between the neural network's predictions and the WRF-Chem ground truth ($\lesssim 10$ % within the planetary boundary layer, with the aforementioned data filters applied), than between WRF-Chem and the independent reference data in the previously conducted model evaluation study. This substantiated two main conclusions:

1. The choice of input variables is adequate, and their combined information content is sufficient for the reconstruction of tropospheric NO_2 profiles.

2. The bottleneck of NitroNet's prediction accuracy is not the predictive capability of the neural network (determined by its type, size, etc.), but the errors inherent to the model-based training data. These can be reduced by the aforementioned data filters prior to training, although increasingly stricter filter criteria result in rapid loss of training examples.

The agreement of NitroNet and WRF-Chem with reference data from TROPOMI and AirBase was investigated on the central European training domain (May 2019). The NO_2 profiles from NitroNet were found to be in significantly better agreement with the NO_2 VCDs from TROPOMI (RMSE between NitroNet and TROPOMI: $3.8 \cdot 10^{14}$ molec. cm^{-2}, RMSE between WRF-Chem and TROPOMI: $6.7 \cdot 10^{14}$ molec. cm^{-2}). These improvements were attributed to the fact that NitroNet was trained on filtered data and thereby prevented from fully adopting the systematic prediction errors of WRF-Chem. Additionally, based on the NO_2 VCD received as input, NitroNet is informed of the total tropospheric NO_2 load, which constrains the amplitudes of the predicted NO_2 profiles. The assimilation of satellite data into chemistry and transport models is possible, but considered a highly complex endeavor and therefore not supported by WRF-Chem out-of-the-box. In the comparison to AirBase data, NitroNet and WRF-Chem showed similar prediction accuracy, with RMSE values of 3.2 µg m^{-3} and 3.4 µg m^{-3}. Furthermore, an average Mo-CL bias of ~ 65 % was determined, with good agreement between WRF-Chem and NitroNet. Overall, this intercomparison study was conducted with the aim to obtain a broad overview of the performance differences between NitroNet and WRF-Chem. Because it inevitably included a small fraction of data that NitroNet had already

seen during training (≈ 7 %, see above), it was not used to reason further about NitroNet's standalone performance. Instead, a validation study on entirely new data was conducted, featuring summertime evaluations (May 2022) on different geographic domains in Europe, the US, and Asia, and a full year evaluation (August 2021 – July 2022) on a central European domain. This also included validations against FRM_4DOAS MAX-DOAS data. The main findings of this regional-seasonal study can be summarized as follows:

1. NitroNet is capable of generalizing to new geographic domains. This was particularly well demonstrated over two domains in Europe covering the UK, Spain and Portugal, and another domain covering the eastern US. On two other domains over India and eastern China, significantly larger prediction errors were observed. These could partially relate to different atmospheric composition (e.g. higher aerosol loads and free tropospheric NO_2 concentrations) or to inaccuracies in the available input data (e.g. outdated emission data in China and India).

2. NitroNet can be used in other seasons as well, as demonstrated on the central European domain. Over a large period of the year (approximately from April – September), the prediction errors remained mostly stable (RMSE wrt. TROPOMI: $\sim 4.5 \cdot 10^{14}$ molec. cm^{-2}, RMSE wrt. AirBase: ~ 1.4 µg m^{-3}, w/o urban background stations, as described in the point below). Between October and March, significantly larger prediction errors and negative biases were identified, (bias wrt. TROPOMI: up to -22.6 %, bias wrt. AirBase: up to -49.6 %), while the correlation strength remained stable throughout ($R \approx 0.90$ wrt. TROPOMI, $R \approx 0.70$ wrt. AirBase). Nonetheless, even such severe biases are still comparable to those reportedly produced by RCT models during winter.

3. NitroNet's predictions are low biased compared to the AirBase in situ measurements classified as "urban background". The AirBase documentation mentions, that under specific circumstances, even measurements dominated by car traffic can be classified as "background" (as opposed to "traffic"). Traffic measurements had previously been verified to be poorly represented in the WRF-Chem simulation. Therefore, NitroNet cannot be expected to reproduce them well. Omission of the urban background stations was found to alleviate NitroNet's low bias entirely during summertime (from -10.5 % to $+2.2$ %). It remains unclear whether the observed low bias should be attributed to a misleading classification of the in situ instruments or a tendency of NitroNet to underestimate NO_2 concentrations in certain urban environments.

The unique benefits of NitroNet

NitroNet is conceptually novel, because, at the time of writing, all previously published neural networks for the prediction of NO_2 were exclusively trained on surface in situ measurements. The fact that NitroNet was trained on synthetic data from an RCT simulation results in three main benefits:

1. NitroNet can not only predict NO_2 surface concentrations, but full tropospheric NO_2 profiles at TROPOMI overpass time. These give a more comprehensive insight into tropospheric air chemistry, allow for validation against satellite observations, and may be used as a priori information in future satellite retrievals (see outlook below). The only other viable source of tropospheric NO_2 profiles with similar spatial resolution and coverage are chemistry and transport models like WRF-Chem. However, NitroNet is orders of magnitude faster and showed overall smaller errors in the validation against observational reference data. These improvements in prediction accuracy must be attributed to the aforementioned data filtering routine, which partly prevents NitroNet from reproducing WRF-Chem's systematic errors.

2. NitroNet can predict correction factors for the Mo-CL bias of the widely used molybdenum-based chemiluminescence in situ measurements. Other neural networks, trained on such measurements, inevitably reproduce this bias. NitroNet does not suffer from this issue and can be used to correct the biases of the affected instruments.

3. NitroNet's training dataset has dense spatial coverage. In situ instruments rather represent point measurements at selected locations, and are often not representative of extended domains due to their sparse coverage. Synthetic training data from RCT simulations are unaffected by this issue, and thus more diverse. In fact, training data over inaccessible terrains (e.g. on water, mountainous regions, and dense forest) can currently only be obtained in sufficient amounts within a model-based approach. Another benefit of training data with dense spatial coverage (here: more relevant in polluted regions) is that unlike in situ measurements they are certainly unaffected by "strategic placement". The training data of NitroNet are not entirely contiguous, due to the data filtering described above. Particularly the filtering based on the relative errors between the modelled NO_2 VCD and TROPOMI satellite observations leads to a significant loss of training examples with low NO_2 VCDs, which concerns mostly remote and/or over-water regions. Putting these remarks aside, a neural network trained on NO_2 data with such dense spatial coverage is a novelty.

Based on its capabilities, NitroNet can be viewed as a first NO_2 profile retrieval for the TROPOMI satellite instrument. However, its limitations should be kept in mind. Given

its considerable wintertime biases, NitroNet should be considered a *prototype* for now. Furthermore, NitroNet cannot replace classic RCT models such as WRF-Chem entirely, because they can simulate various other trace gases and meteorological variables without restriction to colocated satellite observations.

Outlook – Future improvements of NitroNet

More diverse training data — The most urgent step forward in the development of NitroNet is to obtain more diverse training data, possibly over water, from other geographic domains, and the winter season. This could substantially reduce the discussed generalization errors of the model, making it more broadly usable. Particularly wintertime RCT simulations have posed a serious hurdle and will require additional methodical refinements in order to achieve good agreement to independent measurements. An outreach to the international RCT modelling community could be the most effective way to obtain more abundant training data.

Validation against further observational data — Additionally, further observational data should be used for the validation of NitroNet. Evaluation studies based on different effective penetration depths into the atmosphere (e.g. above optically thick clouds) could be used to assess NitroNet's prediction accuracy in the higher layers of the atmosphere. The recently conducted CINDI-3 measurement campaign might provide helpful insight into the NO_2 profiles within the lowest $\sim$ 100 m above ground. Community efforts to produce a harmonized global dataset of surface NO_2 observations have been newly published (the GHOST dataset, see Bowdalo et al., 2024), and could be used to extend the validation of NitroNet at the surface to other geographical domains, such as the US and China. With the advent of geostationary Earth observation satellites such as GEMS in east Asia (Kim et al., 2020), Sentinel-4 in Europe (Stark et al., 2013), and TEMPO in the US (Naeger et al., 2021), hourly daytime satellite observations of tropospheric NO_2 VCDs will soon be available, with GEMS already in operation. Data from geostationary satellites will play a pivotal role in NitroNet's future development by extending its predictions to multiple hours of the day and providing an additional means of validation.

Other types of neural networks — Although the feed-forward neural network on which NitroNet is based was found to be adequate for the accurate reconstruction of model-based NO_2 profiles, more complex neural network topologies can be explored in the future. In particular, convolutional and invertible neural networks (see e.g. Ardizzone et al., 2019) would enable NitroNet to process input data in awareness of spatial context or within a Bayesian framework for better assessment of prediction uncertainties.

Public release of NitroNet — A community version of the NitroNet model is planned to

be released after a thorough revision of its source code. This concerns many of NitroNet's design decisions, such as the choice of data transformations, the use of relative (as opposed to absolute) filter criteria for the training set, and the use of surface wind speeds (as opposed to average values in the planetary boundary layer) in the computation of the NO_2 influx variable. NitroNet's public release will use the Python programming language and will require no advanced knowledge of machine learning on behalf of the user.

Outlook – Future use cases of NitroNet

In closing this thesis, a perspective on two future use cases of NitroNet is given.

Revision of NO_2 surface pollution estimates — The combined results of in situ measurements, RCT simulations, and previous neural network models give comprehensive insight into the surface NO_2 pollution levels in many parts of the world. Together with medical impact models derived from epidemiological studies this allows e.g. for the estimation of related mortality rates. However, there still exist considerable uncertainties. On one hand, the Mo-CL bias of the in situ instruments leads to an overestimation of the true NO_2 concentration. On the other hand, potentially strategic instrument placement in unrepresentative regions might lead to underestimations. So far it is unknown, in what balance these opposing effects stand. NitroNet's ability to predict NO_2 concentrations with dense spatial coverage, along with estimates of the Mo-CL bias, may provide answers in that regard and, subsequently, more realistic assessments of the associated medical impacts.

Implementation of improved regional satellite products — Satellite retrievals of the tropospheric NO_2 VCD require a priori information for the computation of the air mass factors that convert the observed slant column densities to vertical column densities. In the case of TROPOMI, these a priori profiles are taken from the coarsely resolved TM5-MP model ($1\,^\circ \times 1\,^\circ$ horizontal resolution). It has been shown in various studies and this thesis, that using higher resolved NO_2 a priori profiles from RCT models results in significantly larger tropospheric columns of +15 % on average and up to +30 % in strongly polluted regions (see e.g. Liu et al., 2021 using profiles with a horizontal resolution of $0.3\,^\circ \times 0.2\,^\circ$, or Douros et al., 2023 with a horizontal resolution of $0.1\,^\circ \times 0.1\,^\circ$). The main limitation in the development of such improved satellite products is the computational burden of running RCT models at higher resolutions. NitroNet could be used to generate NO_2 a priori profiles on the horizontal resolution of the TROPOMI observations (3.5 km × 5.5 km $\approx 0.03\,^\circ \times 0.05\,^\circ$ at nadir) in order to obtain an improved regional satellite product, e.g. for the central European domain. The results can be expected to differ from previous studies due to the improved horizontal resolution.

Author's publications

The following peer-reviewed articles were published during the research phase of this thesis:

Kuhn L., Kuhn J., Wagner T., Platt U. (2022). The NO_2 camera based on gas correlation spectroscopy. *Atmospheric Measurement Techniques*, 15(5):1395–1414.
DOI: 10.5194/amt-15-1395-2022.

Kuhn L., Beirle, S., Kumar, V., Osipov, S., Pozzer, A., Bösch, T., Kumar, R., and Wagner T. (2024). On the influence of vertical mixing, boundary layer schemes, and temporal emission profiles on tropospheric NO_2 in WRF-Chem – comparisons to in situ, satellite, and MAX-DOAS observations. *Atmospheric Chemistry and Physics*, 24(1):185–217.
DOI: 10.5194/acp-24-185-2024.

Kuhn, L., Beirle, S., Osipov, S., Pozzer, A., and Wagner, T. (2024). NitroNet – a machine learning model for the prediction of NO_2 profiles from TROPOMI observations. *Atmospheric Measurement Techniques*, 17(21):6485–6516. DOI: 10.5194/amt-17-6485-2024.

Lange, K., Richter, A., Bösch, T., Zilker, B., Latsch, M., Behrens, L. K., Okafor, C. M., Bösch, H., Burrows, J. P., Merlaud, A., Pinardi, G., Fayt, C., Friedrich M. M., Dimitropoulou, E., Van Roozendael, M., Ziegler, S., Ripperger-Lukosiunaite, S., **Kuhn L.**, Lauster, B., Wagner, T., Hong, H., Kim, D., Chang, L.-S., Bae, K., Song, C.-K., Park, J.-U., and Lee, H. (2024). Validation of GEMS tropospheric NO_2 columns and their diurnal variation with ground-based DOAS measurements. *Atmospheric Measurement Techniques*, 17(21):6315–6344. DOI: 10.5194/amt-17-6315-2024.

Acknowledgements

First and foremost, thank *you* for reading my thesis!

This work was enabled by the invaluable help of many colleagues and friends. Besides the professors who supervised and graded my thesis, I would like to express my sincere gratitude to the following people:

Thomas Wagner and Steffen Beirle for supervising my thesis from start to end. I have thoroughly enjoyed our cooperative work and appreciate your invested input, even when time was short. In particular, I am most thankful for the large amount of freedom I was granted in developing and exploring my own research ideas, and the possibility to participate in countless scientific conferences, workshops, and even a measurement campaign in South Korea. Our group has always made a unique impression to me for its culture of genuine appreciation of each member and their interests.

Andrea Pozzer, Sergey Osipov, and Vinod Kumar for their indispensable contribution to setting up, debugging, and understanding the WRF-Chem model. Without your help it would have been impossible to obtain training data for NitroNet, and the project could not have been continued. Thank you for selflessly investing weeks of your personal time into my cause.

Andreas Richter and Tim Bösch for sharing their expertise on satellite and MAX-DOAS retrievals and inviting me to present my work at the University of Bremen.

My colleagues at the Max-Planck Institute for Chemistry in Mainz and the Institute for Environmental Physics in Heidelberg for providing a friendly and diverse work environment. The meetings of Thomas Wagner's satellite group and Ulrich Platt's group for the troposphere and volcanoes have lead to countless peoples' ideas shaping and directing my work. I would like to thank my peers Eva, Eva, and Karolin for the good times in office 308 and wish you the best on your future path in science.

The proof-readers Thomas Wagner, Steffen Beirle, Andrea Pozzer, Sergey Osipov, Waltraud

Ulshöfer, and Thomas Gaskin, for providing their helpful corrections to this thesis.

My family and dear friends for their unfailing help. I would like to thank my friends Leo, Hanno, Thomas, Eva, Jonas, and all others for their company, be it from nearby or far away. From the bottom of my heart, I thank my family – Fritz, Waltraud, and Mario – and most of all Kim for their unwavering love and support at every stage of this journey.

Bibliography

Alicke, B. (2000). *The role of nitrous acid in the boundary layer*. PhD thesis, Universität Heidelberg. `http://katalog.ub.uni-heidelberg.de/cgi-bin/redir.cgi?typ=gnd&u=http%3A%2F%2Fd-nb.info%2Fgnd%2F122808266`.

Anderson, G. (1976). Error propagation by the Monte Carlo method in geochemical calculations. *Geochimica et Cosmochimica Acta*, 40(12):1533–1538.

Anenberg, S. C., Mohegh, A., Goldberg, D. L., Kerr, G. H., Brauer, M., Burkart, K., Hystad, P., Larkin, A., Wozniak, S., and Lamsal, L. (2022). Long-term trends in urban NO_2 concentrations and associated paediatric asthma incidence: estimates from global datasets. *The Lancet Planetary Health*, 6(1):E49–E58.

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U. (2019). Analyzing Inverse Problems with Invertible Neural Networks. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1808.04730.

Atkinson, R., Baulch, D. L., Cox, R. A., Crowley, J. N., Hampson, R. F., Hynes, R. G., Jenkin, M. E., Rossi, M. J., and Troe, J. (2004). Evaluated kinetic and photochemical data for atmospheric chemistry: Volume I - gas phase reactions of O_x, HO_x, NO_x and SO_x species. *Atmospheric Chemistry and Physics*, 4(6):1461–1738.

Barron, J. T. (2017). Continuously Differentiable Exponential Linear Units. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1704.07483.

Beirle, S., Borger, C., Dörner, S., Eskes, H., Kumar, V., de Laat, A., and Wagner, T. (2021). Catalog of NO_x emissions from point sources as derived from the divergence of the NO_2 flux for TROPOMI. *Earth System Science Data*, 13(6):2995–3012.

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y., and Wagner, T. (2019). The Mainz profile algorithm (MAPA). *Atmospheric Measurement Techniques*, 12(3):1785–1806.

Beirle, S., Lampel, J., Lerot, C., Sihler, H., and Wagner, T. (2017). Parameterizing the instrumental spectral response function and its changes by a super-Gaussian and its derivatives. *Atmospheric Measurement Techniques*, 10(2):581–598.

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(10):281–305.

Berkhout, A., Gast, L., van der Hoff, G., Swart, D., Hoed, M., and Allaart, M. (2018). Atmospheric NO_2 profiles measured with lidar during the CINDI-2 campaign. *EPJ Web of Conferences*, 176:10002.

Bieser, J., Aulinger, A., Matthias, V., Quante, M., and Denier van der Gon, H. (2011). Vertical emission profiles for Europe based on plume rise calculations. *Environmental Pollution*, 159(10):2935–2946.

Blechschmidt, A.-M., Arteta, J., Coman, A., Curier, L., Eskes, H., Foret, G., Gielen, C., Hendrick, F., Marécal, V., Meleux, F., Parmentier, J., Peters, E., Pinardi, G., Piters, A. J. M., Plu, M., Richter, A., Segers, A., Sofiev, M., Valdebenito, A. M., Van Roozendael, M., Vira, J., Vlemmix, T., and Burrows, J. P. (2020). Comparison of tropospheric NO_2 columns from MAX-DOAS retrievals and regional air quality model simulations. *Atmospheric Chemistry and Physics*, 20(5):2795–2823.

Bösch, T. (2019). *Detailed analysis of MAX-DOAS measurements in Bremen: spatial and temporal distribution of aerosols, formaldehyde and nitrogen dioxide*. PhD thesis, Universität Bremen. `http://nbn-resolving.de/urn:nbn:de:gbv:46-00107093-11`.

Bougeault, P. and Lacarrère, P. (1989). Parameterization of Orography-Induced Turbulence in a Mesobeta-Scale Model. *Monthly Weather Review*, 117(8):1872–1890.

Bourgeois, I., Peischl, J., Neuman, J. A., Brown, S. S., Allen, H. M., Campuzano-Jost, P., Coggon, M. M., DiGangi, J. P., Diskin, G. S., Gilman, J. B., Gkatzelis, G. I., Guo, H., Halliday, H. A., Hanisco, T. F., Holmes, C. D., Huey, L. G., Jimenez, J. L., Lamplugh, A. D., Lee, Y. R., Lindaas, J., Moore, R. H., Nault, B. A., Nowak, J. B., Pagonis, D., Rickly, P. S., Robinson, M. A., Rollins, A. W., Selimovic, V., St. Clair, J. M., Tanner, D., Vasquez, K. T., Veres, P. R., Warneke, C., Wennberg, P. O., Washenfelder, R. A., Wiggins, E. B., Womack, C. C., Xu, L., Zarzana, K. J., and Ryerson, T. B. (2022). Comparison of airborne measurements of NO, NO_2, HONO, NO_y, and CO during FIREX-AQ. *Atmospheric Measurement Techniques*, 15(16):4901–4930.

Bowdalo, D., Basart, S., Guevara, M., Jorba, O., Pérez García-Pando, C., Jaimes Palomera, M., Rivera Hernandez, O., Puchalski, M., Gay, D., Klausen, J., Moreno, S., Netcheva, S., and Tarasova, O. (2024). GHOST: A globally harmonised dataset of surface atmospheric composition measurements. *Earth System Science Data*, 16(3):4417–4495.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification And Regression Trees*. Routledge, 1st edition.

Brown, D., Brownrigg, R., Haley, M., and Huang, W. (2019). The NCAR Command Language (Version 6.6.2). DOI: 10.5065/D6WD3XH5.

Builtjes, P., van Loon, M., Schaap, M., Teeuwisse, S., Visschedijk, A., and Bloos, J. (2002). The development of an emission data base over Europe and further contributions of TNO-MEP. `https://www.umweltbundesamt.de/sites/default/files/medien/publikation/long/3607.pdf`, visited: 13 July 2021.

Burrows, J. P., Hölzle, E., Goede, A. P. H., Visser, H., and Fricke, W. (1995). SCIAMACHY scanning imaging absorption spectrometer for atmospheric chartography. *Acta Astronautica*, 35(7):445–451.

Burrows, J. P., Weber, M., Buchwitz, M., Rozanov, V., Ladstätter-Weißenmayer, A., Richter, A., DeBeek, R., Hoogen, R., Bramstedt, K., Eichmann, K.-U., Eisinger, M., and Perner, D. (1999). The Global Ozone Monitoring Experiment (GOME): Mission Concept and First Scientific Results. *Journal of the Atmospheric Sciences*, 56(2):151–175.

Callies, J., Corpaccioli, E., Eisinger, M., Hahne, A., and Lefebvre, A. (2000). GOME-2 Metop's Second-Generation Sensor for Operational Ozone Monitoring. *ESA Bulletin 102*, 102:28–36.

Cao, E. L. (2023). National ground-level NO_2 predictions via satellite imagery driven convolutional neural networks. *Frontiers in Environmental Science*, 11.

Chan, K. L., Khorsandi, E., Liu, S., Baier, F., and Valks, P. (2021). Estimation of Surface NO_2 Concentrations over Germany from TROPOMI Satellite Observations Using a Machine Learning Method. *Remote Sensing*, 13(5):969.

Chan, K. L., Wiegner, M., van Geffen, J., De Smedt, I., Alberti, C., Cheng, Z., Ye, S., and Wenig, M. (2020). MAX-DOAS measurements of tropospheric NO_2 and HCHO in Munich and the comparison to OMI and TROPOMI satellite observations. *Atmospheric Measurement Techniques*, 13(8):4499–4520.

Chin, M., Rood, R. B., Lin, S.-J., Müller, J.-F., and Thompson, A. M. (2000). Atmospheric sulfur cycle simulated in the global model GOCART: Model description and global properties. *Journal of Geophysical Research: Atmospheres*, 105(D20):24671–24687.

Compernolle, S. (2023). Quarterly Validation Report of the Sentinel-5 Precursor Operational Data Products #21: April 2018 - November 2023. `https://s5p-mpc-vdaf.aeronomie.be/ProjectDir/reports//pdf/S5P-MPC-IASB-ROCVR-21.01.00.pdf`, visited: 26 April 2024.

Cooper, M. J., Martin, R. V., Hammer, M. S., Levelt, P. F., Veefkind, P., Lamsal, L. N., Krotkov, N. A., Brook, J. R., and McLinden, C. A. (2022). Global fine-scale changes in ambient NO_2 during COVID-19 lockdowns. *Nature*, 601(7893):380–387.

Costantini, M. G., Khalek, I., McDonald, J. D., and van Erp, A. M. (2016). The Advanced Collaborative Emissions Study (ACES) of 2007- and 2010-Emissions Compliant Heavy-Duty Diesel Engines: Characterization of Emissions and Health Effects. *Emission Control Science and Technology*, 2:215–227.

Crippa, M., Guizzardi, D., Butler, T., Keating, T., Wu, R., Kaminski, J., Kuenen, J., Kurokawa, J., Chatani, S., Morikawa, T., Pouliot, G., Racine, J., Moran, M. D., Klimont, Z., Manseau, P. M., Mashayekhi, R., Henderson, B. H., Smith, S. J., Suchyta, H., Muntean, M., Solazzo, E., Banja, M., Schaaf, E., Pagani, F., Woo, J.-H., Kim, J., Monforti-Ferrario, F., Pisoni, E., Zhang, J., Niemi, D., Sassi, M., Ansari, T., and Foley, K. (2023). The HTAP_v3 emission mosaic: merging regional and global monthly emissions (20002018) to support air quality modelling and policies. *Earth System Science Data*, 15(6):2667–2694.

Crippa, M., Guizzardi, D., Oreggioni, G., Muntean, M., and Schaaf, E. (2020). EDGARv5.0 Air Pollutant Emissions. DOI: 10.1594/PANGAEA.921922.

Crippa, M., Solazzo, E., Huang, G., Guizzardi, D., Koffi, E., Muntean, M., Schieberle, C., Friedrich, R., and Janssens-Maenhout, G. (2020). High resolution temporal profiles in the Emissions Database for Global Atmospheric Research. *Scientific Data*, 7(121).

de Haan, J. F., Bosma, P. B., and Hovenier, J. W. (1987). The adding method for multiple scattering calculations of polarized light. *Astronomy and Astrophysics*, 183(2):371–391.

Deutschmann, T., Beirle, S., Frieß, U., Grzegorski, M., Kern, C., Kritten, L., Platt, U., Prados-Román, C., Puķīte, J., Wagner, T., Werner, B., and Pfeilsticker, K. (2011). The Monte Carlo atmospheric radiative transfer model McArtim: Introduction and validation of Jacobians and 3D features. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(6):1119–1137.

Douros, J., Eskes, H., van Geffen, J., Boersma, K. F., Compernolle, S., Pinardi, G., Blechschmidt, A.-M., Peuch, V.-H., Colette, A., and Veefkind, P. (2023). Comparing Sentinel-5P TROPOMI NO_2 column observations with the CAMS regional air quality ensemble. *Geoscientific Model Development*, 16(2):509–534.

Dozat, T. (2017). Incorporating Nesterov Momentum into *Adam*. `https://cs229.stanford.edu/proj2015/054_report.pdf`, visited: 18 May 2024.

Du, Q., Zhao, C., Zhang, M., Dong, X., Chen, Y., Liu, Z., Hu, Z., Zhang, Q., Li, Y., Yuan, R., and Miao, S. (2020). Modeling diurnal variation of surface $PM_{2.5}$ concentrations over East China with WRF-Chem: impacts from boundary-layer mixing and anthropogenic emission. *Atmospheric Chemistry and Physics*, 20(5):2839–2863.

Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., San Martini, F., Sheehy, P. M., Zahniser, M. S., Shorter, J. H., Wormhoudt, J. C., Lamb, B. K., Allwine, E. J., Gaffney, J. S., Marley, N. A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Ramos Villegas, C. R., Kolb, C. E., Molina, L. T., and Molina, M. J. (2007). Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment. *Atmospheric Chemistry and Physics*, 7(10):2691–2704.

Emmons, L. K., Schwantes, R. H., Orlando, J. J., Tyndall, G., Kinnison, D., Lamarque, J., Marsh, D., Mills, M. J., Tilmes, S., Bardeen, C., Buchholz, R. R., Conley, A., Gettelman, A., Garcia, R., Simpson, I., Blake, D. R., Meinardi, S., and Pétron, G. (2020). The Chemistry Mechanism in the Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(4).

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster, S. (2010). Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4). *Geoscientific Model Development*, 3(1):43–67.

Eskes, H., Tsikerdekis, A., Ades, M., Alexe, M., Benedictow, A. C., Bennouna, Y., Blake, L., Bouarar, I., Chabrillat, S., Engelen, R., Errera, Q., Flemming, J., Garrigues, S., Griesfeller, J., Huijnen, V., Ilic, L., Inness, A., Kapsomenakis, J., Kipling, Z., Langerock, B., Mortier, A., Parrington, M., Pison, I., Pitkanen, M., Remy, S., Richter, A., Schoenhardt, A., Schulz, M., Thouret, V., Warneke, T., Zerefos, C., and Peuch, V.-H. (2024). Technical Note: Evaluation of the Copernicus Atmosphere Monitoring Service Cy48R1 upgrade of June 2023. *EGUsphere*, 2024:1–57. DOI: 10.5194/egusphere-2023-3129.

Eskes, H., van Geffen, J., Boersma, F., Eichmann, K.-U., Apituley, A., Pedergnana, M., Sneep, M., Veefkind, J. P., and Loyola, D. (2022). *Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Nitrogendioxide*. Royal Netherlands Meteorological Institute. `https://sentinel.esa.int/documents/247904/2474726/Sentinel-5P-Level-2-Product-User-Manual-Nitrogen-Dioxide.pdf/ad25ea4c-3a9a-3067-0d1c-aaa56eb1746b?t=1626258361795&download=true`, visited: 22 May 2024.

European Environment Agency (2023). EMEP/EEA air pollutant emission inventory guidebook 2023: Technical guidance to prepare national emission inventories. https://www.eea.europa.eu/publications/emep-eea-guidebook-2023, visited: 1 August 2024.

European Environment Agency (2024). Exceedance of air quality standards in Europe. https://www.eea.europa.eu/en/analysis/indicators/exceedance-of-air-quality-standards?activeAccordion=ecdb3bcf-bbe9-4978-b5cf-0b136399d9f8, visited: 23 May 2024.

Fayt, C., Friedrich, M., and Hendrick, F. (2021). *Fiducial Reference Measurements for Ground-Based DOAS Air-Quality Observations*. Royal Belgian Institute for Space Aeronomy. https://frm4doas.aeronomie.be/ProjectDir/FRM4DOAS_CCN02_D20_MAXDOAS_Network_Operational_Processing_System_Architecture_Design_Document__v2.0_20210903.pdf, visited: 2 May 2024.

Finlayson-Pitts, B. J. and Pitts Jr., J. N. (2000). *Chemistry of the Upper and Lower Atmosphere: Theory, Experiments, and Applications*. Elsevier, 1st edition.

Fontijn, A., Sabadell, A. J., and Ronco, R. J. (1970). Homogeneous chemiluminescent measurement of nitric oxide with ozone. Implications for continuous selective monitoring of gaseous air pollutants. *Analytical Chemistry*, 42(6):575–579.

Freitas, S. R., Longo, K. M., Chatfield, R., Latham, D., Silva Dias, M. A. F., Andreae, M. O., Prins, E., Santos, J. C., Gielow, R., and Carvalho Jr., J. A. (2007). Including the subgrid scale plume rise of vegetation fires in low resolution atmospheric transport models. *Atmospheric Chemistry and Physics*, 7(13):3385–3398.

Friedrich, M. M., Rivera, C., Stremme, W., Ojeda, Z., Arellano, J., Bezanilla, A., García-Reynoso, J. A., and Grutter, M. (2019). NO_2 vertical profiles and column densities from MAX-DOAS measurements in Mexico City. *Atmospheric Measurement Techniques*, 12(4):2545–2565.

Frieß, U., Beirle, S., Alvarado Bonilla, L., Bösch, T., Friedrich, M. M., Hendrick, F., Piters, A., Richter, A., van Roozendael, M., Rozanov, V. V., Spinei, E., Tirpitz, J.-L., Vlemmix, T., Wagner, T., and Wang, Y. (2019). Intercomparison of max-doas vertical profile retrieval algorithms: studies using synthetic data. *Atmospheric Measurement Techniques*, 12(4):2155–2181.

Frieß, U., Monks, P. S., Remedios, J. J., Rozanov, A., Sinreich, R., Wagner, T., and Platt, U. (2006). MAX-DOAS O_4 measurements: A new technique to derive information on atmospheric aerosols: 2. Modeling studies. *Journal of Geophysical Research: Atmospheres*, 111(D14).

Gardner, M. and Dorling, S. (1999). Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment*, 33(5):709–719.

Ge, B., Sun, Y., Liu, Y., Dong, H., Ji, D., Jiang, Q., Li, J., and Wang, Z. (2013). Nitrogen dioxide measurement by cavity attenuated phase shift spectroscopy (CAPS) and implications in ozone production efficiency and nitrate formation in Beijing, China. *Journal of Geophysical Research: Atmospheres*, 118(16):9499–9509.

Georgoulias, A. K., Boersma, K. F., van Vliet, J., Zhang, X., van der A, R., Zanis, P., and de Laat, J. (2020). Detection of NO_2 pollution plumes from individual ships with the TROPOMI/S5P satellite sensor. *Environmental Research Letters*, 15(12):124037.

Ghahremanloo, M., Lops, Y., Choi, Y., and Yeganeh, B. (2021). Deep Learning Estimation of Daily Ground-Level NO_2 Concentrations From Remote Sensing Data. *Journal of Geophysical Research: Atmospheres*, 126(21).

Girshick, R. (2015). Fast R-CNN. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1504.08083.

Graininger, J. F. and Ring, J. (1962). Anomalous Fraunhofer Line Profiles. *Nature*, 193(4817):762.

Grell, G. A. and Dévényi, D. (2002). A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophysical Research Letters*, 29(14).

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and Eder, B. (2005). Fully coupled "online" chemistry within the WRF model. *Atmospheric Environment*, 39(37):6957–6975.

Griffin, D., Zhao, X., McLinden, C. A., Boersma, F., Bourassa, A., Dammers, E., Degenstein, D., Eskes, H., Fehr, L., Fioletov, V., Hayden, K., Kharol, S. K., Li, S., Makar, P., Martin, R. V., Mihele, C., Mittermeier, R. L., Krotkov, N., Sneep, M., Lamsal, L. N., ter Linden, M., Geffen, J. v., Veefkind, P., and Wolde, M. (2019). HighResolution Mapping of Nitrogen Dioxide With TROPOMI: First Results and Validation Over the Canadian Oil Sands. *Geophysical Research Letters*, 46(2):1049–1060.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C. (2006). Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). *Atmospheric Chemistry and Physics*, 6(11):3181–3210.

Guevara, M., Tena, C., Porquet, M., Jorba, O., and Pérez García-Pando, C. (2019). HERMESv3, a stand-alone multi-scale atmospheric emission modelling framework - Part 1: global and regional module. *Geoscientific Model Development*, 12(5):1885–1907.

Haagen-Smit, A. J. (1952). Chemistry and Physiology of Los Angeles Smog. *Industrial & Engineering Chemistry*, 44(6):1342–1346.

Hausmann, K., Strogies, S., Boettcher, C., P., G., Kotzulla, M., Günther, D., Juhrich, K., Plickert, S., Dettling, F., Kludt, R., Kuntze, D., Reichart, A., Reichel, J., Gromke, U., Haenel, H.-D., Rösemann, C., Döring, U., Schiller, S., Oehmichen, K., and Stümer, W. (2020). German Informative Inventory Report. `http://iir-de-2020.wikidot.com/summary`, visited: 13 July 2021.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1502.01852.

Hendrick, F., Müller, J.-F., Clémer, K., Wang, P., De Mazière, M., Fayt, C., Gielen, C., Hermans, C., Ma, J. Z., Pinardi, G., Stavrakou, T., Vlemmix, T., and Van Roozendael, M. (2014). Four years of ground-based MAX-DOAS observations of HONO and NO_2 in the Beijing area. *Atmospheric Chemistry and Physics*, 14(2):765–781.

Hendrycks, D. and Gimpel, K. (2023). Gaussian Error Linear Units (GELUs). *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1606.08415.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Hong, S.-Y. (2010). A new stable boundary-layer mixing scheme and its impact on the simulated East Asian summer monsoon. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1481–1496.

Horbanski, M., Pöhler, D., Lampel, J., and Platt, U. (2019). The ICAD (iterative cavity-enhanced DOAS) method. *Atmospheric Measurement Techniques*, 12(6):3365–3381.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Huang, G., Brook, R., Crippa, M., Janssens-Maenhout, G., Schieberle, C., Dore, C., Guizzardi, D., Muntean, M., Schaaf, E., and Friedrich, R. (2017). Speciation of anthropogenic

emissions of non-methane volatile organic compounds: a global gridded data set for 1970-2012. *Atmospheric Chemistry and Physics*, 17(12):7683–7701.

Huangfu, P. and Atkinson, R. (2020). Long-term exposure to NO_2 and O_3 and all-cause and respiratory mortality: A systematic review and meta-analysis. *Environment International*, 144:105998.

Huijnen, V., Eskes, H. J., Poupkou, A., Elbern, H., Boersma, K. F., Foret, G., Sofiev, M., Valdebenito, A., Flemming, J., Stein, O., Gross, A., Robertson, L., D'Isidoro, M., Kioutsioukis, I., Friese, E., Amstrup, B., Bergstrom, R., Strunk, A., Vira, J., Zyryanov, D., Maurizi, A., Melas, D., Peuch, V.-H., and Zerefos, C. (2010). Comparison of OMI NO_2 tropospheric columns with an ensemble of global and European regional air quality models. *Atmospheric Chemistry and Physics*, 10(7):3273–3296.

Hussein, M. R. (2005). Ultraviolet radiation and skin cancer: molecular mechanisms. *Journal of Cutaneous Pathology*, 32(3):191–205.

Hönninger, G., von Friedeburg, C., and Platt, U. (2004). Multi axis differential optical absorption spectroscopy (MAX-DOAS). *Atmospheric Chemistry and Physics*, 4(1):231–254.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research*, 113(D13).

Ialongo, I., Virta, H., Eskes, H., Hovila, J., and Douros, J. (2020). Comparison of TROPOMI/Sentinel-5 Precursor NO_2 observations with ground-based measurements in Helsinki. *Atmospheric Measurement Techniques*, 13(1):205–218.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Denier van der Gon, H., Flemming, J., Forkel, R., Giordano, L., Jiménez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Makar, P. A., Manders-Groot, A., Neal, L., Pérez, J. L., Pirovano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R. S., Syrakov, D., Torian, A., Tuccella, P., Wang, K., Werhahn, J., Wolke, R., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., and Galmarini, S. (2015). Evaluation of operational online-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part II: Particulate matter. *Atmospheric Environment*, 115:421–441.

International Civil Aviation Organization (1993). *Manual of the ICAO Standard Atmosphere (extended to 80 kilometres (262 500 feet))*. International Civil Aviation Organization, 3rd edition.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Preprint uploaded to the arXiv preprint server.* DOI: 10.48550/arXiv.1502.03167.

Jacob, D. J. (1999). *Introduction to Atmospheric Chemistry.* Princeton University Press, Princeton, NJ.

Jadon, A., Patil, A., and Jadon, S. (2022). A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting. *Preprint uploaded to the arXiv preprint server.* DOI: 10.48550/arXiv.2211.02989.

Janjić, Z. I. (1994). The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes. *Monthly Weather Review*, 122(5):927–945.

Javed, U., Kubistin, D., Martinez, M., Pollmann, J., Rudolf, M., Parchatka, U., Reiffs, A., Thieser, J., Schuster, G., Horbanski, M., Pöhler, D., Crowley, J. N., Fischer, H., Lelieveld, J., and Harder, H. (2019). Laser-induced fluorescence-based detection of atmospheric nitrogen dioxide and comparison of different techniques during the PARADE 2011 field campaign. *Atmospheric Measurement Techniques*, 12(3):1461–1481.

Jiang, Z., Zhu, R., Miyazaki, K., McDonald, B. C., Klimont, Z., Zheng, B., Boersma, K. F., Zhang, Q., Worden, H., Worden, J. R., Henze, D. K., Jones, D. B. A., Denier van der Gon, H. A. C., and Eskes, H. (2022). Decadal Variabilities in Tropospheric Nitrogen Oxides Over United States, Europe, and China. *Journal of Geophysical Research: Atmospheres*, 127(3):e2021JD035872.

Jimenez, J. L., Mcrae, G. J., Nelson, D. D., Zahniser, M. S., and Kolb, C. E. (2000). Remote Sensing of NO and NO_2 Emissions from Heavy-Duty Diesel Trucks Using Tunable Diode Lasers. *Environmental Science & Technology*, 34(12):2380–2387.

Judd, L. M., Al-Saadi, J. A., Szykman, J. J., Valin, L. C., Janz, S. J., Kowalewski, M. G., Eskes, H. J., Veefkind, J. P., Cede, A., Mueller, M., Gebetsberger, M., Swap, R., Pierce, R. B., Nowlan, C. R., Abad, G. G., Nehrir, A., and Williams, D. (2020). Evaluating Sentinel-5P TROPOMI tropospheric NO_2 column densities with airborne and Pandora spectrometers near New York City and Long Island Sound. *Atmospheric Measurement Techniques*, 13(11):6113–6140.

Jung, J., Lee, J., Kim, B., and Oh, S. (2017). Seasonal variations in the NO_2 artifact from chemiluminescence measurements with a molybdenum converter at a suburban site in Korea (downwind of the Asian continental outflow) during 2015-2016. *Atmospheric Environment*, 165:290–300.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., and Kim, S. (2021). Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia. *Environmental Pollution*, 288:117711.

Karl, M. (2018). Development of the city-scale chemistry transport model CityChem-EPISODE and its application to the city of Hamburg. *Geoscientific Model Development Discussions [preprint]*. DOI: 10.5194/gmd-2018-8.

Kessinger, S., Minkos, A., Dauert, U., Feigenspann, S., Hellack, B., Moravek, A., Richter, S., and Wichmann-Fiebig, M. (2023). Luftqualität 2022. Technical report. `https://www.umweltbundesamt.de/publikationen/luftqualitaet-2022`, visited: 15 March 2024.

Khaniabadi, Y. O., Goudarzi, G., Daryanoosh, S. M., Borgini, A., Tittarelli, A., and De Marco, A. (2016). Exposure to PM_{10}, NO_2, and O_3 and impacts on human health. *Environmental Science and Pollution Research*, 24(3):2781–2789.

Kida, H., Koide, T., Sasaki, H., and Chiba, M. (1991). A New Approach for Coupling a Limited Area Model to a GCM for Regional Climate Simulations. *Journal of the Meteorological Society of Japan. Ser. II*, 69(6):723–728.

Kim, J., Jeong, U., Ahn, M.-H., Kim, J. H., Park, R. J., Lee, H., Song, C. H., Choi, Y.-S., Lee, K.-H., Yoo, J.-M., Jeong, M.-J., Park, S. K., Lee, K.-M., Song, C.-K., Kim, S.-W., Kim, Y. J., Kim, S.-W., Kim, M., Go, S., Liu, X., Chance, K., Miller, C. C., Al-Saadi, J., Veihelmann, B., Bhartia, P. K., Torres, O., Abad, G. G., Haffner, D. P., Ko, D. H., Lee, S. H., Woo, J.-H., Chong, H., Park, S. S., Nicks, D., Choi, W. J., Moon, K.-J., Cho, A., Yoon, J., Kim, S.-K., Hong, H., Lee, K., Lee, H., Lee, S., Choi, M., Veefkind, P., Levelt, P. F., Edwards, D. P., Kang, M., Eo, M., Bak, J., Baek, K., Kwon, H.-A., Yang, J., Park, J., Han, K. M., Kim, B.-R., Shin, H.-W., Choi, H., Lee, E., Chong, J., Cha, Y., Koo, J.-H., Irie, H., Hayashida, S., Kasai, Y., Kanaya, Y., Liu, C., Lin, J., Crawford, J. H., Carmichael, G. R., Newchurch, M. J., Lefer, B. L., Herman, J. R., Swap, R. J., Lau, A. K. H., Kurosu, T. P., Jaross, G., Ahlers, B., Dobber, M., McElroy, C. T., and Choi, Y. (2020). New Era of Air Quality Monitoring from Space: Geostationary Environment Monitoring Spectrometer (GEMS). *Bulletin of the American Meteorological Society*, 101(1):E1–E22.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1412.6980.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-Normalizing Neural Networks. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1706.02515.

Knote, C., Tuccella, P., Curci, G., Emmons, L., Orlando, J. J., Madronich, S., Baró, R., Jiménez-Guerrero, P., Luecken, D., Hogrefe, C., Forkel, R., Werhahn, J., Hirtl, M., Pérez, J. L., San José, R., Giordano, L., Brunner, D., Yahya, K., and Zhang, Y. (2015). Influence of the choice of gas-phase mechanism on predictions of key gaseous pollutants during the AQMEII phase-2 intercomparison. *Atmospheric Environment*, 115:553–568.

Kuenen, J., Dellaert, S., Visschedijk, A., Jalkanen, J.-P., Super, I., and Denier van der Gon, H. (2022). CAMS-REG-v4: a state-of-the-art high-resolution European emission inventory for air quality modelling. *Earth System Science Data*, 14(2):491–515.

Kuhn, L., Beirle, S., Kumar, V., Osipov, S., Pozzer, A., Bösch, T., Kumar, R., and Wagner, T. (2024a). On the influence of vertical mixing, boundary layer schemes, and temporal emission profiles on tropospheric NO_2 in WRF-Chem – comparisons to in situ, satellite, and MAX-DOAS observations. *Atmospheric Chemistry and Physics*, 24(1):185–217.

Kuhn, L., Beirle, S., Kumar, V., Osipov, S., Pozzer, A., Bösch, T., Kumar, R., and Wagner, T. (2023). Modelling of tropospheric NO_2 using WRF-Chem with optimized temporal NO_x emission profiles derived from in-situ observations Comparisons to in-situ, satellite, and MAX-DOAS observations over central Europe. *EGUsphere [preprint]*. DOI: 10.5194/egusphere-2022-1473.

Kuhn, L., Beirle, S., Osipov, S., Pozzer, A., and Wagner, T. (2024b). NitroNet – a machine learning model for the prediction of tropospheric NO_2 profiles from TROPOMI observations. *Atmospheric Measurement Techniques*, 17(21):6485–6516. DOI: 10.5194/amt-17-6485-2024.

Kuhn, L., Kuhn, J., Wagner, T., and Platt, U. (2022). The NO_2 camera based on gas correlation spectroscopy. *Atmospheric Measurement Techniques*, 15(5):1395–1414.

Kuik, F., Kerschbaumer, A., Lauer, A., Lupascu, A., von Schneidemesser, E., and Butler, T. M. (2018). Top-down quantification of NO_x emissions from traffic in an urban area using a high-resolution regional atmospheric chemistry model. *Atmospheric Chemistry and Physics*, 18(11):8203–8225.

Kuik, F., Lauer, A., Churkina, G., Denier van der Gon, H. A. C., Fenner, D., Mar, K. A., and Butler, T. M. (2016). Air quality modelling in the Berlin-Brandenburg region using WRF-Chem v3.7.1: sensitivity to resolution of model grid and input data. *Geoscientific Model Development*, 9(12):4339–4363.

Kumar, V., Beirle, S., Dörner, S., Mishra, A. K., Donner, S., Wang, Y., Sinha, V., and Wagner, T. (2020). Long-term MAX-DOAS measurements of NO_2, HCHO, and aerosols and evaluation of corresponding satellite data products over Mohali in the Indo-Gangetic Plain. *Atmospheric Chemistry and Physics*, 20(22):14183–14235.

Kumar, V., Remmers, J., Beirle, S., Fallmann, J., Kerkweg, A., Lelieveld, J., Mertens, M., Pozzer, A., Steil, B., Barra, M., Tost, H., and Wagner, T. (2021). Evaluation of the coupled high-resolution atmospheric chemistry model system MECO(n) using in situ and MAX-DOAS NO_2 measurements. *Atmospheric Measurement Techniques*, 14(7):5241–5269.

Lamsal, L. N., Martin, R. V., van Donkelaar, A., Steinbacher, M., Celarier, E. A., Bucsela, E., Dunlea, E. J., and Pinto, J. P. (2008). Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument. *Journal of Geophysical Research: Atmospheres*, 113(D16).

Lange, K., Richter, A., and Burrows, J. P. (2022). Variability of nitrogen oxide emission fluxes and lifetimes estimated from Sentinel-5P TROPOMI observations. *Atmospheric Chemistry and Physics*, 22(4):2745–2767.

Lange, K., Richter, A., Schönhardt, A., Meier, A. C., Bösch, T., Seyler, A., Krause, K., Behrens, L. K., Wittrock, F., Merlaud, A., Tack, F., Fayt, C., Friedrich, M. M., Dimitropoulou, E., Van Roozendael, M., Kumar, V., Donner, S., Dörner, S., Lauster, B., Razi, M., Borger, C., Uhlmannsiek, K., Wagner, T., Ruhtz, T., Eskes, H., Bohn, B., Santana Diaz, D., Abuhassan, N., Schüttemeyer, D., and Burrows, J. P. (2023). Validation of Sentinel-5P TROPOMI tropospheric NO_2 products by comparison with NO_2 measurements from airborne imaging DOAS, ground-based stationary DOAS, and mobile car DOAS measurements during the S5P-VAL-DE-Ruhr campaign. *Atmospheric Measurement Techniques*, 16(5):1357–1389.

Latza, U., Gerdes, S., and Baur, X. (2009). Effects of nitrogen dioxide on human health: Systematic review of experimental and epidemiological studies conducted between 2002 and 2006. *International Journal of Hygiene and Environmental Health*, 212(3):271–287.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leighton, P. (2012). *Photochemistry of Air Pollution*. Elsevier.

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

Levelt, P., van den Oord, G. H. J., Dobber, M. R., Mälkki, A., Visser, H., de Vries, J., Stammes, P., Lundell, J. O. V., and Saari, H. (2006). The Ozone Monitoring Instrument. *IEEE Transactions on Geoscience and Remote Sensing*, 44(5):1093–1101.

Liley, J. B., Johnston, P. V., McKenzie, R. L., Thomas, A. J., and Boyd, I. S. (2000). Stratospheric NO_2 variations from a long time series at Lauder, New Zealand. *Journal of Geophysical Research: Atmospheres*, 105(D9):11633–11640.

Lin, H., Long, M. S., Sander, R., Sandu, A., Yantosca, R. M., Estrada, L. A., Shen, L., and Jacob, D. J. (2023). An Adaptive AutoReduction Solver for Speeding Up Integration of Chemical Kinetics in Atmospheric Chemistry Models: Implementation and Evaluation in the Kinetic PreProcessor (KPP) Version 3.0.0. *Journal of Advances in Modeling Earth Systems*, 15(2).

Liu, F., Beirle, S., Zhang, Q., Dörner, S., He, K., and Wagner, T. (2016). NO_x lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations. *Atmospheric Chemistry and Physics*, 16(8):5283–5298.

Liu, S., Valks, P., Pinardi, G., Xu, J., Chan, K. L., Argyrouli, A., Lutz, R., Beirle, S., Khorsandi, E., Baier, F., Huijnen, V., Bais, A., Donner, S., Dörner, S., Gratsea, M., Hendrick, F., Karagkiozidis, D., Lange, K., Piters, A. J. M., Remmers, J., Richter, A., Van Roozendael, M., Wagner, T., Wenig, M., and Loyola, D. G. (2021). An improved TROPOMI tropospheric NO_2 research product over Europe. *Atmospheric Measurement Techniques*, 14(11):7297–7327.

Lorente, A., Folkert Boersma, K., Yu, H., Dörner, S., Hilboll, A., Richter, A., Liu, M., Lamsal, L. N., Barkley, M., De Smedt, I., Van Roozendael, M., Wang, Y., Wagner, T., Beirle, S., Lin, J.-T., Krotkov, N., Stammes, P., Wang, P., Eskes, H. J., and Krol, M. (2017). Structural uncertainty in air mass factor calculation for NO_2 and HCHO satellite retrievals. *Atmospheric Measurement Techniques*, 10(3):759–782.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. *Preprint uploaded to the arXiv preprint server.* DOI: 10.48550/arXiv.1711.05101.

Madronich, S. (1987). Photodissociation in the atmosphere: 1. Actinic flux and the effects of ground reflections and clouds. *Journal of Geophysical Research*, 92(D8):9740–9752.

Mar, K. A., Ojha, N., Pozzer, A., and Butler, T. M. (2016). Ozone air quality simulations with WRF-Chem (v3.5.1) over Europe: model evaluation and chemical mechanism comparison. *Geoscientific Model Development*, 9(10):3699–3728.

Marais, E. A., Roberts, J. F., Ryan, R. G., Eskes, H., Boersma, K. F., Choi, S., Joiner, J., Abuhassan, N., Redondas, A., Grutter, M., Cede, A., Gomez, L., and Navarro-Comas, M. (2021). New observations of NO_2 in the upper troposphere from TROPOMI. *Atmospheric Measurement Techniques*, 14(3):2389–2408.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133.

McPeters, R. D., Krueger, A. J., Bhartia, P., Herman, J. R., Oaks, A., Ahmad, Z., Cebula, R. P., Schlesinger, B. M., Swissler, T., Taylor, S. L., Torres, O., and Wellemeyer, C. G. (1993). *Nimbus-7 Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide*. National Aeronautics and Space Administration.

Mesinger, F. (2020). Forecasting upper tropospheric turbulence within the framework of the Mellor-Yamada 2.5 closure. *UCAR Visiting Research Program at the U.S. National Meteorological Center*, 18:428–429.

Misra, D. (2020). Mish: A Self Regularized Non-Monotonic Activation Function. *Preprint uploaded to the arXiv preprint server*. DOI: 10.48550/arXiv.1908.08681.

Monin, A. and Obukhov, S. (1954). Basic laws of turbulent mixing in the surface layer of the atmosphere. *Tr. Akad. Nauk. SSSR Geophiz. Inst.*, 24(151):163–187.

Munro, R., Lang, R., Klaes, D., Poli, G., Retscher, C., Lindstrot, R., Huckle, R., Lacan, A., Grzegorski, M., Holdak, A., Kokhanovsky, A., Livschitz, J., and Eisinger, M. (2016). The GOME-2 instrument on the Metop series of satellites: instrument design, calibration, and level 1 data processing - an overview. *Atmospheric Measurement Techniques*, 9(3):1279–1301.

Müller, J.-F. (1992). Geographical distribution and seasonal variation of surface emissions and deposition velocities of atmospheric trace gases. *Journal of Geophysical Research: Atmospheres*, 97(D4):3787–3804.

Naeger, A. R., Newchurch, M. J., Moore, T., Chance, K., Liu, X., Alexander, S., Murphy, K., and Wang, B. (2021). Revolutionary Air-Pollution Applications from Future Tropospheric Emissions: Monitoring of Pollution (TEMPO) Observations. *Bulletin of the American Meteorological Society*, 102(9):E1735–E1741.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, 116(D12).

Noxon, J. F. (1975). Nitrogen dioxide in the stratosphere and troposphere measured by ground-based absorption spectroscopy. *Science*, 189(4202):547–549.

Omrani, H., Drobinski, P., and Dubos, T. (2012). Spectral nudging in regional climate modelling: how strongly should we nudge? *Quarterly Journal of the Royal Meteorological Society*, 138(668):1808–1813.

Ortega, I., Berg, L. K., Ferrare, R. A., Hair, J. W., Hostetler, C. A., and Volkamer, R. (2016). Elevated aerosol layers modify the O_2-O_2 absorption measured by ground-based MAX-DOAS. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 176:34–49.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Peng, S., Giron, C., Liu, G., d'Aspremont, A., Benoit, A., Lauvaux, T., Lin, X., de Almeida Rodrigues, H., Saunois, M., and Ciais, P. (2023). High-resolution assessment of coal mining methane emissions by satellite in Shanxi, China. *iScience*, 26(12):108375.

Petetin, H., Beekmann, M., Colomb, A., Denier van der Gon, H. A. C., Dupont, J.-C., Honoré, C., Michoud, V., Morille, Y., Perrussel, O., Schwarzenboeck, A., Sciare, J., Wiedensohler, A., and Zhang, Q. J. (2015). Evaluating BC and NO_x emission inventories for the Paris region from MEGAPOLI aircraft measurements. *Atmospheric Chemistry and Physics*, 15(17):9799–9818.

Phillips, N. A. (1957). A coordinate system having some special advantages for numerical forecasting. *Journal of the Atmospheric Sciences*, 14(2):184–185.

Platt, U., Perner, D., and Pätz, H. W. (1979). Simultaneous measurement of atmospheric CH_2O, O_3, and NO_2 by differential optical absorption. *Journal of Geophysical Research: Oceans*, 84(C10):6329–6335.

Platt, U. and Stutz, J. (2008). *Differential Optical Absorption Spectroscopy*. Springer Berlin, Heidelberg, 1st edition.

Poraicu, C., Müller, J.-F., Stavrakou, T., Fonteyn, D., Tack, F., Deutsch, F., Laffineur, Q., Van Malderen, R., and Veldeman, N. (2023). Cross-evaluating WRF-Chem v4.1.2, TROPOMI, APEX, and in situ NO_2 measurements over Antwerp, Belgium. *Geoscientific Model Development*, 16(2):479–508.

Pozzer, A., Jöckel, P., and Van Aardenne, J. (2009). The influence of the vertical distribution of emissions on tropospheric chemistry. *Atmospheric Chemistry and Physics*, 9(24):9417–9432.

Puķīte, J., Kühl, S., Deutschmann, T., Platt, U., and Wagner, T. (2010). Extending differential optical absorption spectroscopy for limb measurements in the UV. *Atmospheric Measurement Techniques*, 3(3):631–653.

Raissi, M., Perdikaris, P., and Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

Reed, C., Evans, M. J., Di Carlo, P., Lee, J. D., and Carpenter, L. J. (2016). Interferences in photolytic NO_2 measurements: explanation for an apparent missing oxidant? *Atmospheric Chemistry and Physics*, 16(7):4707–4724.

Richmond-Bryant, J., Chris Owen, R., Graham, S., Snyder, M., McDow, S., Oakes, M., and Kimbrough, S. (2017). Estimation of on-road NO_2 concentrations, NO_2/NO_x ratios, and related roadway gradients from near-road monitoring data. *Air Quality, Atmosphere & Health*, 10:611–625.

Richter, A. and Wagner, T. (2011). *The Remote Sensing of Tropospheric Composition from Space - The Use of UV, Visible and Near IR Solar BackScattered Radiation to Determine Trace Gases*, chapter 2, pages 67–121. Springer Berlin Heidelberg.

Riess, T. C. V. W., Boersma, K. F., Van Roy, W., de Laat, J., Dammers, E., and van Vliet, J. (2023). To new heights by flying low: comparison of aircraft vertical NO_2 profiles to model simulations and implications for TROPOMI NO_2 retrievals. *Atmospheric Measurement Techniques*, 16(21):5287–5304.

Riess, T. C. V. W., Boersma, K. F., van Vliet, J., Peters, W., Sneep, M., Eskes, H., and van Geffen, J. (2022). Improved monitoring of shipping NO_2 with TROPOMI: decreasing NO_x emissions in European seas during the COVID-19 pandemic. *Atmospheric Measurement Techniques*, 15(5):1415–1438.

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice*. Series on atmospheric, oceanic and planetary physics. World Scientific.

Roedel, W. and Wagner, T. (2017). *Physik unserer Umwelt: Die Atmosphäre*. Springer Spektrum Berlin, Heidelberg, 5th edition.

Ruppert, D. (2014). *Trimming and Winsorization*. John Wiley & Sons, Ltd.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Seinfeld, J. H. and Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, 3rd edition.

Shah, V., Jacob, D. J., Li, K., Silvern, R. F., Zhai, S., Liu, M., Lin, J., and Zhang, Q. (2020). Effect of changing NO_x lifetime on the seasonality and long-term trends of satellite-observed tropospheric NO_2 columns over China. *Atmospheric Chemistry and Physics*, 20(3):1483–1495.

Shapley, L. S. (1951). *Notes on the N-Person Game - II. The Value of an N-Person Game*. RAND Corporation.

Sluis, W. W., Allaart, M. A. F., Piters, A. J. M., and Gast, L. F. L. (2010). The development of a nitrogen dioxide sonde. *Atmospheric Measurement Techniques*, 3(6):1753–1762.

Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., and Janssens-Maenhout, G. (2021). Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases. *Atmospheric Chemistry and Physics*, 21(7):5655–5683.

Solomon, S., Schmeltekopf, A. L., and Sanders, R. W. (1987). On the interpretation of zenith sky absorption measurements. *Journal of Geophysical Research: Atmospheres*, 92(D7):8311–8319.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Stark, H., Möller, H., Courrèges-Lacoste, G., Koopman, R., Mezzasoma, S., and Veihelmann, B. (2013). The Sentinel-4 mission, its components and implementation. Proceedings of the ESA Living Planet Symposium, Edinburgh.

Stauffer, D. and Seaman, N. (1990). Use of Four-Dimensional Data Assimilation in a Limited-Area Mesoscale Model. Part I: Experiments with Synoptic-Scale Data. *Monthly Weather Review*, 118(6):1250–1277.

Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., and Hueglin, C. (2007). Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques. *Journal of Geophysical Research: Atmospheres*, 112(D11).

Stolarski, R. S., Krueger, A. J., Schoeberl, M. R., McPeters, R. D., Newman, P. A., and Alpert, J. C. (1986). Nimbus 7 satellite measurements of the springtime Antarctic ozone decrease. *Nature*, 322(6082):801–811.

Štrumbelj, E. and Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665.

Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*. Atmospheric and Oceanographic Sciences Library. Springer Netherlands, Dordrecht, Netherlands.

Stutz, J., Alicke, B., Ackermann, R., Geyer, A., White, A., and Williams, E. (2004). Vertical profiles of NO_3, N_2O_5, O_3, and NO_x in the nocturnal boundary layer: 1. Observations during the Texas Air Quality Study 2000. *Journal of Geophysical Research: Atmospheres*, 109(D12).

Su, J., McCormick, M. P., Johnson, M. S., Sullivan, J. T., Newchurch, M. J., Berkoff, T. A., Kuang, S., and Gronoff, G. P. (2021). Tropospheric NO_2 measurements using a three-wavelength optical parametric oscillator differential absorption lidar. *Atmospheric Measurement Techniques*, 14(6):4069–4082.

Tack, F., Merlaud, A., Iordache, M.-D., Pinardi, G., Dimitropoulou, E., Eskes, H., Bomans, B., Veefkind, P., and Van Roozendael, M. (2021). Assessment of the TROPOMI tropospheric NO_2 product based on airborne APEX observations. *Atmospheric Measurement Techniques*, 14(1):615–646.

Terrenoire, E., Bessagnet, B., Rouïl, L., Tognet, F., Pirovano, G., Létinois, L., Beauchamp, M., Colette, A., Thunis, P., Amann, M., and Menut, L. (2015). High-resolution air quality simulation over Europe with the chemistry transport model CHIMERE. *Geoscientific Model Development*, 8(1):21–42.

The European Parliament and The European Council (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. `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32008L0050#d1e89-30-1`, visited: 22 April 2024.

Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D. (2008). Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. *Monthly Weather Review*, 136(12):5095–5115.

Tie, X., Madronich, S., Walters, S., Zhang, R., Rasch, P., and Collins, W. (2003). Effect of clouds on photolysis and oxidants in the troposphere. *Journal of Geophysical Research*, 108(D20).

Tirpitz, J.-L., Frieß, U., Hendrick, F., Alberti, C., Allaart, M., Apituley, A., Bais, A., Beirle, S., Berkhout, S., Bognar, K., Bösch, T., Bruchkouski, I., Cede, A., Chan, K. L., den Hoed, M., Donner, S., Drosoglou, T., Fayt, C., Friedrich, M. M., Frumau, A., Gast, L., Gielen, C., Gomez-Martín, L., Hao, N., Hensen, A., Henzing, B., Hermans, C., Jin, J., Kreher, K., Kuhn, J., Lampel, J., Li, A., Liu, C., Liu, H., Ma, J., Merlaud, A., Peters, E., Pinardi, G., Piters, A., Platt, U., Puentedura, O., Richter, A., Schmitt, S., Spinei, E., Stein Zweers, D., Strong, K., Swart, D., Tack, F., Tiefengraber, M., van der Hoff, R., van Roozendael, M., Vlemmix, T., Vonk, J., Wagner, T., Wang, Y., Wang, Z., Wenig, M., Wiegner, M., Wittrock, F., Xie, P., Xing, C., Xu, J., Yela, M., Zhang, C., and Zhao, X. (2021). Intercomparison of MAX-DOAS vertical profile retrieval algorithms: studies on field data from the CINDI-2 campaign. *Atmospheric Measurement Techniques*, 14(1):1–35.

University Corporation for Atmospheric Research (2024). WRF-Chem Version 4.4 User's Guide. `https://ruc.noaa.gov/wrf/wrf-chem/Users_guide.pdf`, visited: 13 March 2023.

U.S. Environmental Protection Agency (1997). National Air Pollutant Emission Trends, 1900-1996. EPA-454/R-97-011, `https://nepis.epa.gov/Exe/ZyNET.exe/2000D3SU.TXT?ZyActionD=ZyDocument&Client=EPA&Index=1995+Thru+1999&Docs=&Query=&Time=&EndTime=&SearchMethod=1&TocRestrict=n&Toc=&TocEntry=&QField=&QFieldYear=&QFieldMonth=&QFieldDay=&IntQFieldOp=0&ExtQFieldOp=0&XmlQuery=&File=D%3A%5Czyfiles%5CIndex%20Data%5C95thru99%5CTxt%5C00000009%5C2000D3SU.txt&User=ANONYMOUS&Password=anonymous&SortMethod=h%7C-&MaximumDocuments=1&FuzzyDegree=0&ImageQuality=r75g8/r75g8/x150y150g16/i425&Display=hpfr&DefSeekPage=x&SearchBack=ZyActionL&Back=ZyActionS&BackDesc=Results%20page&MaximumPages=1&ZyEntry=1&SeekPage=x&ZyPURL`, visited: 15 March 2024.

van Geffen, J., Boersma, K. F., Eskes, H., Sneep, M., ter Linden, M., Zara, M., and Veefkind, J. P. (2020). S5P TROPOMI NO_2 slant column retrieval: method, stability, uncertainties and comparisons with OMI. *Atmospheric Measurement Techniques*, 13(3):1315–1335.

van Geffen, J., Eskes, H., Compernolle, S., Pinardi, G., Verhoelst, T., Lambert, J.-C., Sneep, M., ter Linden, M., Ludewig, A., Boersma, K. F., and Veefkind, J. P. (2022a). Sentinel-5P TROPOMI NO_2 retrieval: impact of version v2.2 improvements and comparisons with OMI and ground-based data. *Atmospheric Measurement Techniques*, 15(7):2037–2060.

van Geffen, J., Eskes, H. J., Boersma, K. F., and Veefkind, J. P. (2022b). *TROPOMI ATBD of the total and tropospheric NO_2 data products.* Royal Netherlands Meteorological Institute. `https://sentinel.esa.int/documents/247904/2476257/Sentinel-5P-TROPOMI-ATBD-NO2-data-products`, visited: 22 May 2024.

Veefkind, J., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H., de Haan, J., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., Vink, R., Visser, H., and Levelt, P. (2012). TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120:70–83.

Villena, G., Bejan, I., Kurtenbach, R., Wiesen, P., and Kleffmann, J. (2012). Interferences of commercial NO_2 instruments in the urban atmosphere and in a smog chamber. *Atmospheric Measurement Techniques*, 5(1):149–159.

Visser, A. J., Boersma, K. F., Ganzeveld, L. N., and Krol, M. C. (2019). European NO_x emissions in WRF-Chem derived from OMI: impacts on summertime surface ozone. *Atmospheric Chemistry and Physics*, 19(18):11821–11841.

Volten, H., Brinksma, E. J., Berkhout, A. J. C., Hains, J., Bergwerff, J. B., Van der Hoff, G. R., Apituley, A., Dirksen, R. J., Calabretta-Jongen, S., and Swart, D. P. J. (2009). NO_2 lidar profile measurements for satellite interpretation and validation. *Journal of Geophysical Research: Atmospheres*, 114(D24301).

Wagner, T., Beirle, S., Benavent, N., Bösch, T., Chan, K. L., Donner, S., Dörner, S., Fayt, C., Frieß, U., García-Nieto, D., Gielen, C., González-Bartolome, D., Gomez, L., Hendrick, F., Henzing, B., Jin, J. L., Lampel, J., Ma, J., Mies, K., Navarro, M., Peters, E., Pinardi, G., Puentedura, O., Puķīte, J., Remmers, J., Richter, A., Saiz-Lopez, A., Shaiganfar, R., Sihler, H., Van Roozendael, M., Wang, Y., and Yela, M. (2019). Is a scaling factor required to obtain closure between measured and modelled atmospheric O_4 absorptions? An assessment of uncertainties of measurements and radiative transfer simulations for 2 selected days during the MAD-CAT campaign. *Atmospheric Measurement Techniques*, 12(5):2745–2817.

Wagner, T., Chance, K., Frieß, U., Gil, M., Goutail, F., Hönninger, G., Johnston, P., Karlsen-Tørnkvist, K., Kostadinov, I., Leser, H., Petritoli, A., Richter, A., Van Roozendael, M., and Platt, U. (2001). Correction of the Ring effect and I_0-effect for DOAS observations of scattered sunlight. Proceedings of the 1st DOAS Workshop.

Warnach, S. (2022). *Bromine monoxide in volcanic plumes - A global survey of volcanic plume composition and chemistry derived from Sentinel-5 Precursor/TROPOMI data.* PhD thesis, Universität Heidelberg. DOI: 10.11588/HEIDOK.00031910.

Wesely, M. (1989). Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmospheric Environment*, 23(6):1293–1304.

Wiedinmyer, C., Akagi, S. K., Yokelson, R. J., Emmons, L. K., Al-Saadi, J. A., Orlando, J. J., and Soja, A. J. (2011). The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning. *Geoscientific Model Development*, 4(3):625–641.

Wikipedia contributor SkywalkerPL (2023). Sentinel-5 Precursor - Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/w/index.php?title=Sentinel-5_Precursor&oldid=1169059861`, visited: 25 April 2024.

Wild, R. J., Dubé, W. P., Aikin, K. C., Eilerman, S. J., Neuman, J. A., Peischl, J., Ryerson, T. B., and Brown, S. S. (2017). On-road measurements of vehicle NO_2/NO_x emission ratios in Denver, Colorado, USA. *Atmospheric Environment*, 148:182–189.

Williams, J. E., Boersma, K. F., Le Sager, P., and Verstraeten, W. W. (2017). The high-resolution version of TM5-MP for optimized satellite retrievals: description and validation. *Geoscientific Model Development*, 10(2):721–750.

World Health Organization (2000). Air Quality Guidelines for Europe. `https://iris.who.int/bitstream/handle/10665/107335/9789289013581-eng.pdf?sequence=1`, visited: 14 March 2024.

World Health Organization (2021). WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. `https://iris.who.int/handle/10665/345329`, visited: 15 March 2024.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., and Deng, S.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, 17(1):26–40.

Yeo, I.-K. and Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Zhang, C., Liu, C., Li, B., Zhao, F., and Zhao, C. (2022). Spatiotemporal neural network for estimating surface NO_2 concentrations over north China and their human health impact. *Environmental Pollution*, 307:119510.

Appendix A

Fundamentals

A.1 Vertical temperature gradients in the troposphere

A.1.1 Dry-adiabatic case

The vertical temperature profile of the troposphere can be explained by thermodynamic considerations. When an air parcel at the surface is heated, its density will decrease, and it will subsequently rise. While rising, the air parcel expands, and performs mechanical work ($\mathrm{d}W = -p\,\mathrm{d}V$) against the ambient air pressure. The process is further assumed adiabatic (i.e. there occurs no exchange of heat, $\mathrm{d}Q = 0$). First, the case of dry air is treated. It follows from the the ideal gas assumptions ($\star$) and first law of thermodynamics ($\star\star$), that

$$\mathrm{d}U \overset{\star}{=} C_V\,\mathrm{d}T \overset{\star\star}{=} \mathrm{d}Q + \mathrm{d}W = \mathrm{d}Q - p\,\mathrm{d}V \tag{A.1}$$

where:

U : inner energy of the gas

C_V : specific molar heat capacity of the gas (at constant volume)

The specific molar heat at constant pressure is defined as

$$C_p = R + C_V \tag{A.2}$$

and the differential form of the ideal gas equation states

$$p\,\mathrm{d}V + V\,\mathrm{d}p = R\,\mathrm{d}T \tag{A.3}$$

From these prerequisites it can be derived that

$$\mathrm{d}Q \overset{(A.1)}{=} p\,\mathrm{d}V + C_V\,\mathrm{d}T \tag{A.4}$$

$$\overset{(A.3)}{=} -V\,\mathrm{d}p + R\,\mathrm{d}T + C_V\,\mathrm{d}T \tag{A.5}$$

$$\overset{(A.2)}{=} -V\,\mathrm{d}p + C_p\,\mathrm{d}T \tag{A.6}$$

$$= -\frac{RT}{p}\,\mathrm{d}p + C_p\,\mathrm{d}T \tag{A.7}$$

$$\overset{(2.33)}{=} Mg\,\mathrm{d}z + C_p\,\mathrm{d}T \tag{A.8}$$

and thus, because $\mathrm{d}Q = 0$,

$$\frac{\mathrm{d}T}{\mathrm{d}z} = -\frac{Mg}{C_p} \tag{A.9}$$

With $C_p \approx 28,97\ \mathrm{J\ K^{-1}}$, $M \approx 28,97\ \mathrm{g\ mol^{-1}}$, and $g \approx 9,81\ \mathrm{m\ s^{-2}}$, the dry adiabatic temperature gradient (also called the *dry adiabatic lapse rate* Γ) evaluates to

$$\frac{\mathrm{d}T}{\mathrm{d}z} \approx -0,00981\ \mathrm{K\ m^{-1}} \tag{A.10}$$

A.1.2 Moist-adiabatic case (no condensation)

As long as no condensation of the water vapor contained in moist air occurs, not much changes and eq. (A.9) can be adapted to account for the specific heat of water vapor. The specific heat of a gas at fixed pressure is defined as

$$c_p = C_p/M \tag{A.11}$$

The moist-adiabatic lapse rate then reads

$$\frac{\mathrm{d}T}{\mathrm{d}z} = -\frac{g}{s \cdot c_{p,\mathrm{water}} + (1-s) \cdot c_{p,\mathrm{air}}} \tag{A.12}$$

where:

$c_{p,\mathrm{water}}$: specific heat of water vapor ($\approx 1,00\ \mathrm{J\ K^{-1}\ g^{-1}}$)

$c_{p,\mathrm{air}}$: specific heat of dry air ($\approx 1,86\ \mathrm{J\ K^{-1}\ g^{-1}}$)

s : mass ratio of water vapor to total air

For a typical value of $s = 0.01$, the lapse rate changes by less than 1 %. Therefore, if no condensation occurs, the presence of water vapor can be practically ignored.

A.1.3 Moist-adiabatic case (with condensation)

If, however, the water content of the air exceeds the saturation level and begins to condensate, the situation changes drastically. The condensation of water releases energy (*latent heat* L), which compensates a fraction of the energy lost by the mechanical work $\mathrm{d}W$, performed during the expansion of the air cell. The latent heat released by condensation is given by

$$\mathrm{d}Q_V := \frac{n}{V}\,\mathrm{d}Q = -L\,\mathrm{d}\rho_{\mathrm{w,sat}} \tag{A.13}$$

where:

Q_V : heat "per volume" as opposed to "per mol"

$\frac{n}{V}$: mols per unit volume

$\rho_{\mathrm{w,sat}}$: absolute moisture at the condensation limit

In analogy to the energy balance obtained in the derivation of the dry lapse rate, it holds that

$$\mathrm{d}Q_V = \frac{n}{V}\,\mathrm{d}Q \overset{\star}{=} \frac{p}{RT}\,\mathrm{d}Q \overset{(A.8)}{=} \frac{p}{RT}\left(Mg\,\mathrm{d}z + C_p\,\mathrm{d}T\right) \tag{A.14}$$

Equations (A.13) and (A.14) combined yield

$$\frac{p}{RT}\left(Mg\,\mathrm{d}z + C_p\,\mathrm{d}T\right) = -L\,\mathrm{d}\rho_{\mathrm{w,sat}} = -L\frac{\mathrm{d}\rho_{\mathrm{w,sat}}}{\mathrm{d}T}\,\mathrm{d}T \tag{A.15}$$

$$\Rightarrow \quad \frac{p}{RT}Mg\,\mathrm{d}z = -\left(L\frac{\mathrm{d}\rho_{\mathrm{w,sat}}}{\mathrm{d}T} + \frac{p}{RT}C_p\right)\mathrm{d}T \tag{A.16}$$

$$\Rightarrow \quad \frac{\mathrm{d}T}{\mathrm{d}z} = -\frac{pMg}{RTL\dfrac{\rho_{\mathrm{w,sat}}}{\mathrm{d}T} + pC_p} \tag{A.17}$$

The larger $\frac{\mathrm{d}\rho_{\mathrm{w,sat}}}{\mathrm{d}T}$, the more dampening of the lapse rate occurs due to condensation. Because $\frac{\mathrm{d}\rho_{\mathrm{w,sat}}}{\mathrm{d}T}$ grows with temperature, the difference to the dry lapse rate is largest at high temperatures. For example, the *international standard atmosphere*, as defined by the *International Civil Aviation Organization*, assumes a lapse rate of $\frac{\mathrm{d}T}{\mathrm{d}z} = -0,0065\ \mathrm{K\ m^{-1}}$ (International Civil Aviation Organization, 1993).

A.2 International barometric height formula

Equation (2.34) can now be revised, incorporating the atmospheric lapse rate. $T(z)$ can be expressed as

$$T(z) = T_0 + \Gamma z \tag{A.18}$$

Here, the temperature gradient is denoted as Γ for brevity, but need not necessarily refer to the dry lapse rate. Insertion into eq. (2.33) yields:

$$\frac{\mathrm{d}p}{p} = -\frac{Mg}{R \cdot (T_0 + \Gamma z)}\,\mathrm{d}z \tag{A.19}$$

and subsequently, by integration on both sides

$$\ln\left(\frac{p(z')}{p_0}\right) = -\frac{Mg}{R}\int_0^{z'} \frac{1}{T_0 + \Gamma z}\,\mathrm{d}z \tag{A.20}$$

Using the integral rule

$$\int \frac{1}{b + ax} = \frac{1}{a}\ln\left(b + ax\right) \tag{A.21}$$

with $a = \Gamma$ and $b = T_0$ follows

$$\ln\left(\frac{p(z')}{p_0}\right) = -\frac{Mg}{R} \cdot \left[\frac{1}{\Gamma}\ln\left(T_0 + \Gamma z\right)\right]_{z=0}^{z=z'} \tag{A.22}$$

$$= -\frac{Mg}{R\Gamma} \cdot \left(\ln\left(T_0 + \Gamma z'\right) - \ln\left(T_0\right)\right) \tag{A.23}$$

$$= -\frac{Mg}{R\Gamma} \cdot \ln\left(1 + \frac{\Gamma}{T_0} z'\right) \tag{A.24}$$

$$\tag{A.25}$$

From hereon, we may use z instead of z' to denote altitude again. Solving for $p(z)$ yields:

$$p(z) = p_0 \cdot \exp\left(-\frac{Mg}{R\Gamma} \cdot \ln\left(1 + \frac{\Gamma}{T_0} z\right)\right) \tag{A.26}$$

$$= p_0 \cdot \left(1 + \frac{\Gamma}{T_0} z\right)^{-\frac{Mg}{R\Gamma}} \tag{A.27}$$

In order to obtain the *international barometric height formula*, the values of the *international standard atmosphere* (see International Civil Aviation Organization, 1993), being $T_0 = 288.15$ K, $p_0 = 1013.25$ hPa, $\Gamma = -0.0065$ K m^{-1}, are inserted into eq. (A.27). This yields

$$p(z) = 1013.25 \text{ hPa} \cdot \left(1 - \frac{0,0065 \text{ K m}^{-1}}{288.15 \text{ K}} \cdot z\right)^{5.255} \tag{A.28}$$

and

$$z(p) = \frac{288.15 \text{ K}}{0,0065 \text{ K m}^{-1}} \cdot \left(1 - \left(\frac{p}{1013.25 \text{ hPa}}\right)^{\frac{1}{5.255}}\right) \tag{A.29}$$

A.3 Derivation of the characteristic atmospheric layers

A.3.1 The free troposphere

The upper part of the troposphere (the "free troposphere") is characterized by *geostrophic flow*, where wind is controlled by the force equilibrium between the *pressure gradient force*

$$\boldsymbol{F}_p = -\boldsymbol{\nabla}_h p \tag{A.30}$$

where:

$\boldsymbol{\nabla}_h = (\partial_x, \partial_y, 0)$: the horizontal gradient operator

p : air pressure

i.e. the force exerted onto a parcel of air, pushing it towards regions of lower pressure, and the *Coriolis force*:

$$\boldsymbol{F}_c = 2\rho \cdot (\boldsymbol{v} \times \boldsymbol{\Omega}) \tag{A.31}$$

where:

ρ : density

$\boldsymbol{v}$: horizontal wind speed vector

$\boldsymbol{\Omega}$: Earth's angular velocity

These forces should be understood in relation to volume, i.e. they have the dimension M L^{-2} T^{-2} (corresponding units of e.g. N m^{-3}). This is helpful for writing out the equation of motion later on. Of course, there is also gravity at play, but for the horizontal dynamics studied here, gravity can be neglected. In the context of horizontal flow, only the fraction of $\boldsymbol{F}_c$ orthogonal to $\boldsymbol{\Omega}$, denoted as $\boldsymbol{F}_{c\perp}$ is of interest. By definition of the vector cross product, the magnitude of this vector is

$$|F_{c\perp}| = 2\rho \cdot (v\Omega \sin\varphi) \tag{A.32}$$

where φ denotes the latitude (which is also the angle between $\boldsymbol{v}$ and $\boldsymbol{\Omega}$ in the plain they span). $\boldsymbol{F}_{c\perp}$ must be orthogonal to $\boldsymbol{v}$, pointing to the right in the northern hemisphere, and to the left in southern hemisphere, i.e. in the direction of the (normalized) vector $\frac{1}{v}(v_y, -v_x, 0)^{\mathrm{T}}$. Combining direction and magnitude of the Coriolis force vector yields:

$$\boldsymbol{F}_{c\perp} = 2\rho \cdot (v\Omega \sin\varphi) \cdot \frac{1}{v}(v_y, -v_x, 0)^{\mathrm{T}} := \rho f \cdot (v_y, -v_x, 0)^{\mathrm{T}} \tag{A.33}$$

where $f = 2\,\Omega \sin\varphi$ is called the *Coriolis parameter*. The force equilibrium $\boldsymbol{F}_p + \boldsymbol{F}_{c\perp} = 0$ then implies

$$v_x = -\frac{1}{\rho f}\partial_y p \qquad \text{and} \qquad v_y = \frac{1}{\rho f}\partial_x p \tag{A.34}$$

The third line of this vector equation reads $0 = 0$, and is obviously irrelevant to the consideration of horizontal flow made here.

A.3.2 Ekman layer and the influence of friction

Friction in fluid motion is a consequence of *momentum exchange* between the fluid layers, moving at different speeds. Consider two layers of a fluid, as depicted in Fig. A.1. Here, flow is in the horizontal x-direction, and the velocity gradient points along the vertical z-axis. Individual fluid particles (by diffusion) and fluid parcels (by turbulence) may move from the slower layer with velocity $v_1(z_1)$ at altitude z_1 to the faster moving layer with velocity $v_2(z_2)$ and vice versa. This equates to an exchange of *momentum density* ρv_x, expressed in the form of the *shear stress tensor*, whose xz-component (describing the transport of x-directed momentum along the z-axis) reads:

$$\tau_{xz} = -(K + \nu)\rho\frac{\partial v_x}{\partial z} \tag{A.35}$$

where:

K : the turbulent diffusion coefficient

ν : the kinematic viscosity

Note, that τ_{xz} is naturally expressed relative to some area A, through which the momentum transfer occurs. The force exerted onto the volume element $V := A\,\mathrm{d}z$ reads

$$F_{fx} \cdot V = \overbrace{-\tau_{xz}(z + \mathrm{d}z) \cdot A}^{\text{from above}} + \overbrace{\tau_{xz}(z) \cdot A}^{\text{from below}} \tag{A.36}$$

$$= -\left(\tau_{xz}(z) + \frac{\mathrm{d}\tau_{xz}}{\mathrm{d}z}\,\mathrm{d}z\right) \cdot A + \tau_{xz}(z) \cdot A \tag{A.37}$$

$$= -\frac{\mathrm{d}\tau_{xz}}{\mathrm{d}z} \cdot A\,\mathrm{d}z \tag{A.38}$$

Dividing by the volume $A\,\mathrm{d}z$ and inserting the expression from eq. (A.35) yields

$$F_{fx} = \frac{\mathrm{d}}{\mathrm{d}z}\left[(K + \nu)\rho \cdot \frac{\partial v_x}{\partial z}\right] \tag{A.39}$$

The expression for the friction force in y-direction is analogous. It is further assumed that $F_{fz} = 0$, i.e. it is sufficient to consider only the shear stress vector $\boldsymbol{\tau} = (\tau_{xz},\ \tau_{yz}, 0)^{\mathrm{T}}$. The effects of surface friction become more relevant closer to Earth's surface. The friction force $\boldsymbol{F}_f$ adds to the force equilibrium in the stationary case:

$$\boldsymbol{F}_c + \boldsymbol{F}_p + \boldsymbol{F}_f = 0 \tag{A.40}$$

This results in an *inclination* of the wind trajectory: In a strictly geostrophic scenario, wind moves parallel to the isobars (lines of constant pressure), but with friction involved, the trajectory is pushed towards the region of lower pressure.

Consider a free tropospheric regime with geostrophic flow (where friction is negligible), and a regime closer to the surface, where surface friction becomes more important. A layer of air at the interface between the two regimes will be dragged along by the geostrophic flow above it, but decelerated by the layers below. As a result, velocity decreases, and the inclination increases, with growing distance to the interface. This vertical wind spiral is called the *Ekman spiral*, and is a dominant feature between the free troposphere and ~ 100 m above the ground. Figure A.2 gives a graphical representation of geostrophic flow, the influence of friction, and the Ekman spiral.

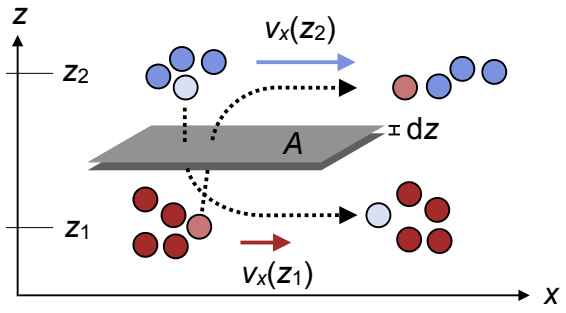

Figure A.1: Momentum exchange between two layers of a fluid. The velocity gradient points in z-direction. Fluid particles may diffuse from the slower layer (red, z_1) to the faster layer (blue, z_2) and vice versa, transporting momentum between the two layers. Turbulent friction, which involves the exchange of entire fluid parcels, can be understood in analogy.

A.3.3 The Prandtl layer, the molecular-viscous layer and overview of the planetary boundary layer

Based on the characteristics of friction, the atmosphere can be divided into two further sublayers: the Prandtl layer, and the molecular-viscous layer. The largest part of the troposphere, the free troposphere, has already been described. Here, it can be assumed, that no interaction with the surface takes place, i.e. $\boldsymbol{\tau} \approx 0$. Below the free troposphere, friction becomes relevant.

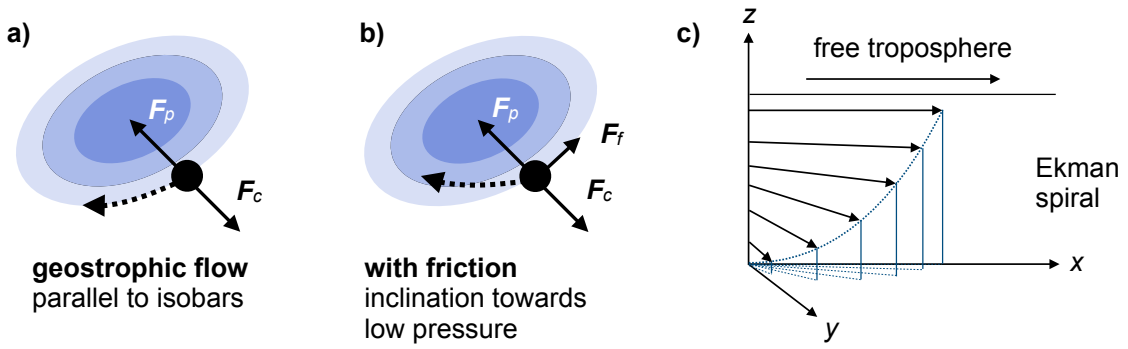

Figure A.2: Effect of surface friction on horizontal flow. a) Purely geostrophic wind, following the isobars. Low pressure regions are depicted in deeper shades of blue, and higher pressure regions in lighter shades. **b)** Wind under the additional influence of surface friction with inclined trajectory. **c)** The resulting Ekman spiral, rotating the wind direction with distance to the interface of the free troposphere.

Recall eq. (A.39), which may be written as

$$F_{fx} = \underbrace{\frac{\mathrm{d}}{\mathrm{d}z}\left[K\rho\frac{\partial v_x}{\partial z}\right]}_{\text{governs Prandtl- and Ekman layer}} + \underbrace{\frac{\mathrm{d}}{\mathrm{d}z}\left[\nu\rho\frac{\partial v_x}{\partial z}\right]}_{\text{governs molecular-viscous layer}} \tag{A.41}$$

(here, for the x-component). In other words, friction can stem from either *turbulent diffusion* (K) or thermal molecular diffusion (ν). The layer in which thermal diffusion dominates only spans a few millimeters above the ground and is called the *molecular-viscous* layer. The remaining part of the PBL is dominated by turbulent diffusion.

For this purpose, we may write the horizontal equation of motion (i.e. Newton's 2nd law of motion applied to the continuum, ($\star$)), assuming the stationary case ($\frac{\mathrm{d}v}{\mathrm{d}t} = 0$, ($\star\star$)), and dominance of turbulence ($K \gg \nu$, ($\star\star\star$))

$$\rho\frac{\mathrm{d}v}{\mathrm{d}t} \overset{\star}{=} \boldsymbol{F}_p + \boldsymbol{F}_c + \boldsymbol{F_f} \overset{\star\star}{=} 0 \tag{A.42}$$

$$\Rightarrow \quad -(\partial_x p, \partial_y p)^{\mathrm{T}} + \rho f \cdot (v_y, -v_x)^{\mathrm{T}} - \frac{\mathrm{d}}{\mathrm{d}z}(\tau_{xz}, \tau_{yz})^{\mathrm{T}} = 0 \tag{A.43}$$

$$\overset{\star\star\star}{\Rightarrow} \quad -(\partial_x p, \partial_y p)^{\mathrm{T}} + \rho f \cdot (v_y, -v_x)^{\mathrm{T}} + \frac{\mathrm{d}}{\mathrm{d}z}\left(K\rho \cdot \left(\frac{\partial v_x}{\partial z}, \frac{\partial v_y}{\partial z}\right)\right)^{\mathrm{T}} = 0 \tag{A.44}$$

in the free troposphere, without friction, the equation reads

$$-(\partial_x p, \partial_y p)^{\mathrm{T}} + \rho f \cdot (v_{gy}, -v_{gx})^{\mathrm{T}} = 0 \tag{A.45}$$

Here, subscript g denotes "geostrophic". Subtraction yields an expression for the horizontal

wind speeds below the free troposphere, relative to the geostrophic wind:

$$v_x - v_{gx} = \frac{1}{f}\frac{\mathrm{d}}{\mathrm{d}z}\left(K\frac{\partial v_y}{\partial z}\right) \tag{A.46}$$

$$v_y - v_{gy} = -\frac{1}{f}\frac{\mathrm{d}}{\mathrm{d}z}\left(K\frac{\partial v_x}{\partial z}\right) \tag{A.47}$$

It is intuitive to assume, that the influence of friction becomes more important closer to the surface, and that a layer of depth Δz exists, in which friction even dominates the pressure gradient and Coriolis force. According to eq. (A.43), this is the case if $\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}z} \approx 0$, or in other words, $\boldsymbol{\tau} = \text{const}$. Although this is not the case at the surface (notice, that eq. (A.47) even predicts a maximum of $\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}z}$ at the surface, because the left-hand side becomes maximal), a statement of similar meaning, $\Delta z \cdot \frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}z} \ll \boldsymbol{\tau}$, holds true in the lowest ~ 50 m of the troposphere. This layer is called the *Prandtl layer*.

To summarize: The atmosphere can coarsely be divided into the stratosphere ($\sim 10 - 50$ km), the tropopause ($\sim 10 - 15$ km), the troposphere ($\lesssim 10 - 15$ km), and further layers (mesosphere, etc.). The troposphere can be divided into the free troposphere ($\gtrsim 1000$ m), and three layers below, in which friction due to the Earth's rigid surface is involved: The Ekman layer ($\sim 100 - 1000$ m), where wind is inclined (following the Ekman spiral), the Prandtl layer ($\sim 0 - 100$ m), where friction by turbulent diffusion is the dominant force, and lastly the molecular-viscous layer ($\lesssim 10$ mm), where friction by thermal diffusion dominates.

A.4 Backpropagation

Backpropagation is the application of the chain rule from calculus to the operations within the layers of a neural network. Here we adopt the following notation:

- C : The loss value, computed from a single network prediction, i.e. $C = \mathcal{L}(y_i, \hat{y}_i)$
- L : The number of network layers
- n_l : The number of neurons in the l-th layer
- a_j^l : The output ("activation") of the j-th neuron of the l-th layer
- w_{jk}^l : weight between the j-th neuron of the l-th layer, and the k-th neuron of the $(l-1)$-th layer
- b_j^l : the bias of the k-th neuron of the l-th layer

First, an analytic expression for the gradient of C wrt. the parameters of the last layer L is derived. The reader may refer to Fig. A.3 along the following calculations. The output, or

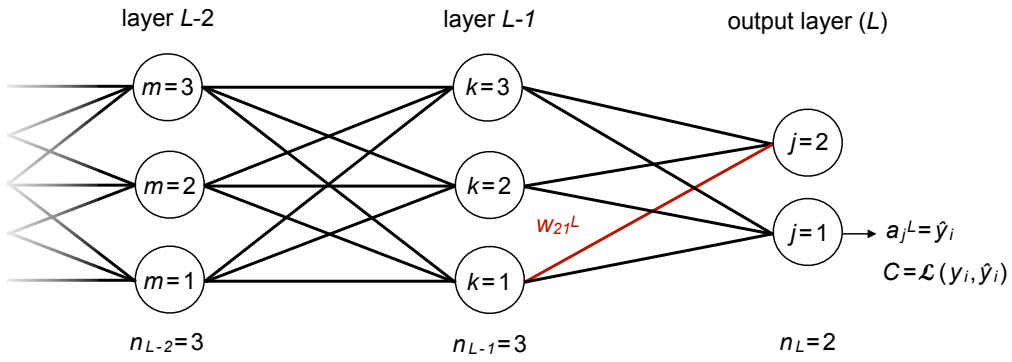

Figure A.3: Helper figure for the explanation of the backpropagation calculus. The neural network's output is computed in forward direction (from left to right). The loss gradient is obtained in backwards direction (from right to left), by application of the chain rule to each neuron's output with respect to its parameters (weights and biases).

activation, of a neuron is written as

$$a_j^L = \phi\left(z_j^L\right), \qquad z_j^L = \sum_{k=1}^{n_{L-1}} w_{jk}^L a_k^{L-1} + b_j^L \tag{A.48}$$

This equates to eq. (2.91). The derivative of C wrt. a weight of a neuron then reads

$$\frac{\partial C}{\partial w_{jk}^L} = \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L} \frac{\partial z_j^L}{\partial w_{jk}^L} \tag{A.49}$$

$\frac{\partial C}{\partial a_j^L}$ depends on the choice of loss function. For example, if C is the sum of squared errors, then

$$\frac{\partial C}{\partial a_j^L} = 2\left(a_j^L - y_j\right) \tag{A.50}$$

$\frac{\partial a_j^L}{\partial z_j^L}$ depends on the choice of activation function. For example, e.g. If $\phi = \mathrm{ReLU}$, then

$$\frac{\partial a_j^L}{\partial z_j^L} = \begin{cases} 0 & z_j^L \leq 0 \\ 1 & z_j^L > 0 \end{cases}, \tag{A.51}$$

whereby the undifferentiability of ReLU at $x = 0$ is resolved by using its subgradient (any value between -1 and 1) in place of the gradient. $\frac{\partial z_j^L}{\partial w_{jk}^L}$ simply equates to a_k^{L-1}. In analogy, the

gradient of C wrt. the bias of a neuron in the last layer reads

$$\frac{\partial C}{\partial b_j^L} = \frac{\partial C}{\partial a_j^L}\,\frac{\partial a_j^L}{\partial z_j^L}\,\frac{\partial z_j^L}{\partial b_j^L} = \frac{\partial C}{\partial a_j^L}\,\frac{\partial a_j^L}{\partial z_j^L} \tag{A.52}$$

because $\frac{\partial z_j^L}{\partial b_j^L} = 1$. Next, we extend the calculus to the second to last layer $L-1$ as follows:

$$\frac{\partial C}{\partial w_{km}^{L-1}} = \frac{\partial C}{\partial a_k^{L-1}}\,\frac{\partial a_k^{L-1}}{\partial z_k^{L-1}}\,\frac{\partial z_k^{L-1}}{\partial w_{km}^{L-1}} \tag{A.53}$$

The latter two terms $\frac{\partial a_k^{L-1}}{\partial z_k^{L-1}}$ and $\frac{\partial z_k^{L-1}}{\partial w_{km}^{L-1}}$ can be computed just as before. Only the first term $\frac{\partial C}{\partial a_k^{L-1}}$ changes, namely to

$$\frac{\partial C}{\partial a_k^{L-1}} = \sum_{j=1}^{n_L-1} \frac{\partial C}{\partial a_j^L}\,\frac{\partial a_j^L}{\partial z_j^L}\,\frac{\partial z_j^L}{\partial a_k^{L-1}} \tag{A.54}$$

of which the first two terms are known, and $\frac{\partial z_j^L}{\partial a_k^{L-1}} = w_{jk}^L$. In analogy

$$\frac{\partial C}{\partial b_k^{L-1}} = \frac{\partial C}{\partial a_k^{L-1}}\,\frac{\partial a_k^{L-1}}{\partial z_k^{L-1}}\,\frac{\partial z_k^{L-1}}{\partial b_k^{L-1}} = \frac{\partial C}{\partial a_k^{L-1}}\,\frac{\partial a_k^{L-1}}{\partial z_k^{L-1}} \tag{A.55}$$

because $\frac{\partial z_k^{L-1}}{\partial b_k^{L-1}} = 1$. This recursive calculus can be continued up until the first layer of the network. It may suffer from a complicated notation, but the individual steps are easy to compute, because the gradient terms of layer l mostly contain recycled terms of the layer $l+1$. The full loss gradient $\nabla_{\boldsymbol{\theta}} C$ is computed by traversing through the network's layers in backwards direction, hence the term "backpropagation".

Note, that the backpropagation algorithm allows to optimize not only the weights and biases of a network's neurons, but *all* trainable parameters. This may, for example, include the slope parameter a of the PReLU activation function, see eq. (2.90).

Appendix B

Regional chemistry and transport modelling with WRF-Chem

B.1 Statements on the Mo-CL bias from the UBA

During the early research phase of this thesis, it was initially unclear whether the Mo-CL biases explained in sect. 2.3.6 apply to the German in situ measurements of NO_2 and NO_x. For this reason, it was decided to contact the UBA personally in order to request an official statement on the topic. The exchange below is shared with explicit consent on behalf of the UBA.

E-Mail to the UBA, 30 March 2022

Dear (retracted),

recently, a number of publications regarding the measurement principle of NO_x surface instruments have come to my attention. For example in Visser et al. (2019), *European NO_x emissions in WRF-Chem derived from OMI: impacts on summertime surface ozone* write:

"The comparison against in situ NO_2 observations from the AirBase network may be hindered by interference of reactive N species for measurements with molybdenum converters. The type of converter is not reported in the database. Literature-reported estimates of measurement overestimations due to this interference are 22 % (Dunlea et al., 2007) and 5 % − 18 % (Boersma et al., 2009) at urban sites and 20 % − 42 % at a rural site (Steinbacher et al., 2007). A correction factor can be applied to obtain corrected NO_2 measurements from observations using a molybdenum converter, which is on average $0.4 - 0.6$ in summer, but with a large spread $(0.2 - 0.8)$ (Lamsal et al., 2008, 2010). The strongest corrections of molybdenum-based in situ NO_2 measurements are needed in remote environments, where NO_x is a relatively smaller component of the total reactive nitrogen budget compared to areas closer to NO_x

sources (Lamsal et al., 2008). We hypothesize that this can partially explain the remaining modelobservation mismatch for NO_2 after the use of topdown emissions."

Here, the authors compare modelled NO_2 concentrations at the surface to "AirBase", a network of surface instruments, which to my knowledge includes UBA instruments. Can you comment on the authors' thesis? After all, they mention correction factors of 0.4 – 0.6, meaning that the surface instruments are biased by up to +100 %. However, this apparently only applies to instruments, that use a molybdenum converter. Are you aware, of whether these are used in the instruments you deploy? Have any measurement biases of this kind been discussed before?

All information on the matter would be greatly appreciated. Many thanks!

With kind regards

Leon Kuhn

Reply from the UBA, 7 April 2022

Dear Mr. Kuhn,

Comparing model results and measurements is always tricky. I cannot retrace the supposed error factors of 0.4 – 0.6. The instruments are tested for cross-sensitivities during quality assurance testing. Instruments with Mo converters show a surplus of approx. 7 % in the presence of other nitrogen compounds. The direct NO_2 measurements have undergone similar quality assurance and equivalence testing. The errors were found to be much lower than described. More information can be found on the website `www.qal1.de`.

The European Union demands a unified measurement protocol, which is explained in detail in EN 14211. Only instruments, which were tested according to this norm may be used. In Germany, instruments are listed in the "Bundesanzeiger", once they have passed quality testing. Additionally, they are listed on the website `www.qal1.de`, along with the corresponding test summaries and possibly made changes. If other systems are used, they must first be tested for equivalence to the reference system. These tests follow the European guidelines to prove equivalence (GDE). All of this is determined by the air quality guidelines.

With kind regards

(retracted)

B.2 Supplementary material

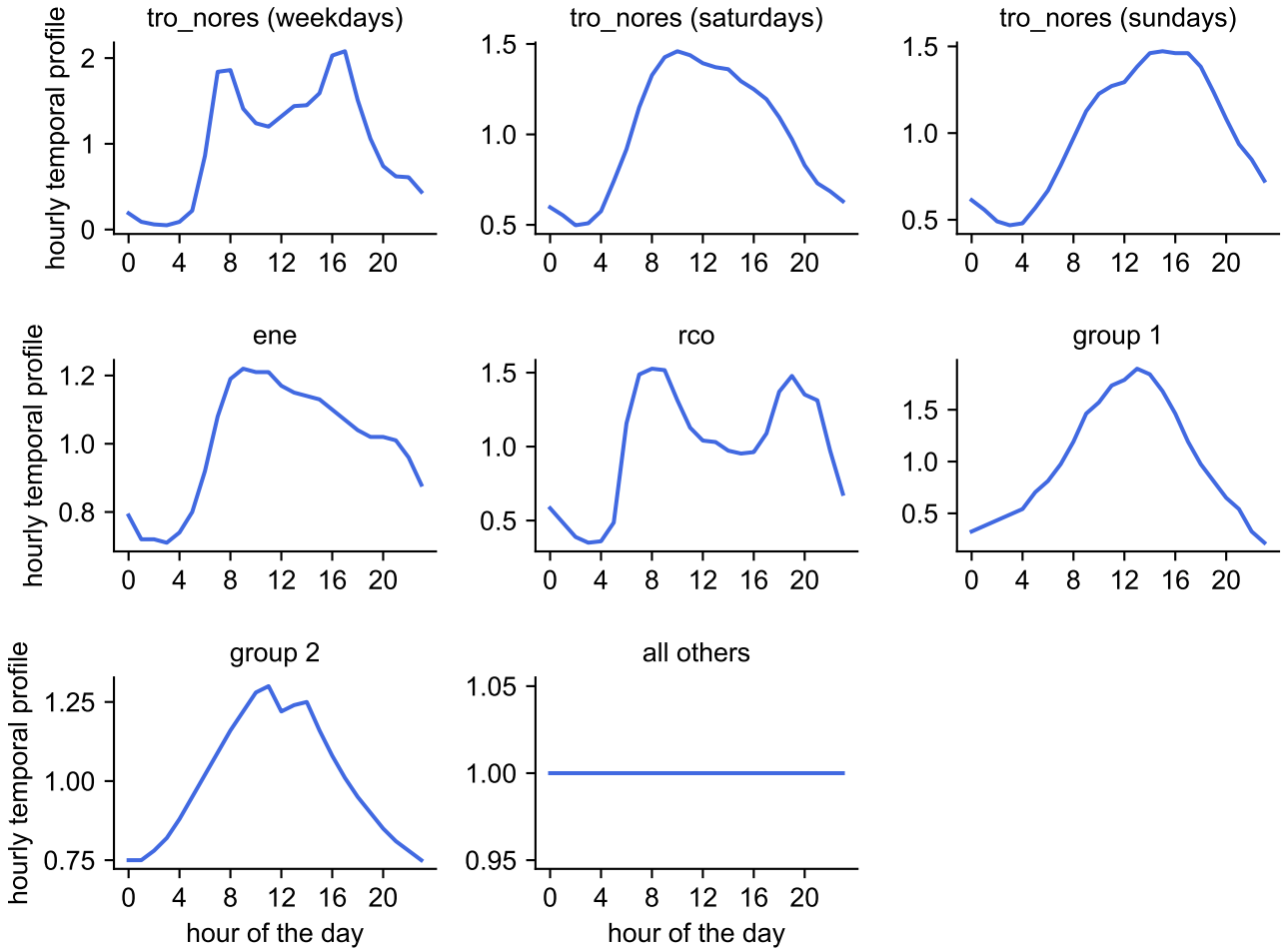

Figure B.1: Overview of the hourly temporal emission profiles used in the WRF-Chem simulation. Group 1 refers to the sectors ags, mnm, awb. Group 2 refers to the sectors tnr_ship, ref_trf, tnr_aviation_lto, tnr_other, pro, swd_inc, neu. For these abbreviations, refer to Table 2.1.

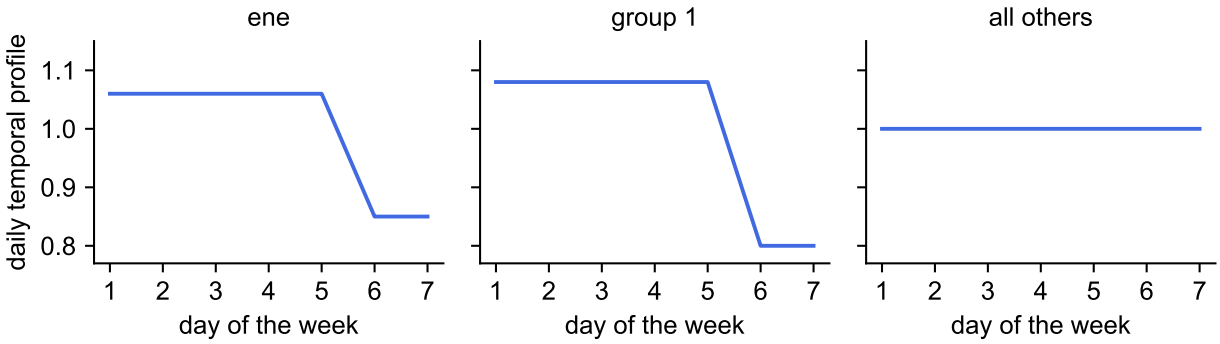

Figure B.2: Overview of the daily temporal emission profiles used in the WRF-Chem simulation. Group 1 refers to the sectors tro_nores, rco, ind, nmm, che, iro, foo_pap, nfe. For these abbreviations, refer to Table 2.1.

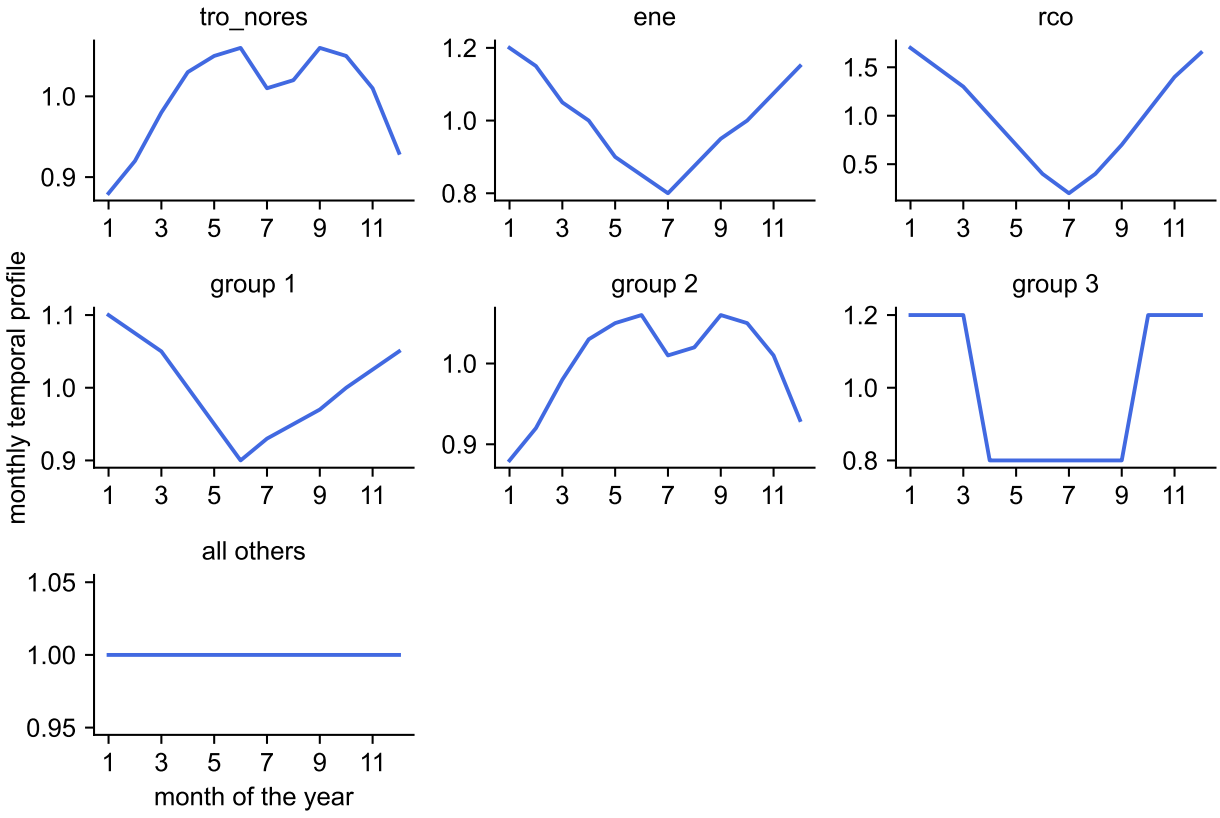

Figure B.3: Overview of the monthly temporal emission profiles used in the WRF-Chem simulation. Group 1 refers to the sectors ind, nmm, che, iro, foo_pap, nfe. Group 2 refers to the sectors tnr_ship, tnr_aviation_lto, tnr_other. Group 3 refers to the sectors ref_rtf, pro. For these abbreviations, refer to Table 2.1.

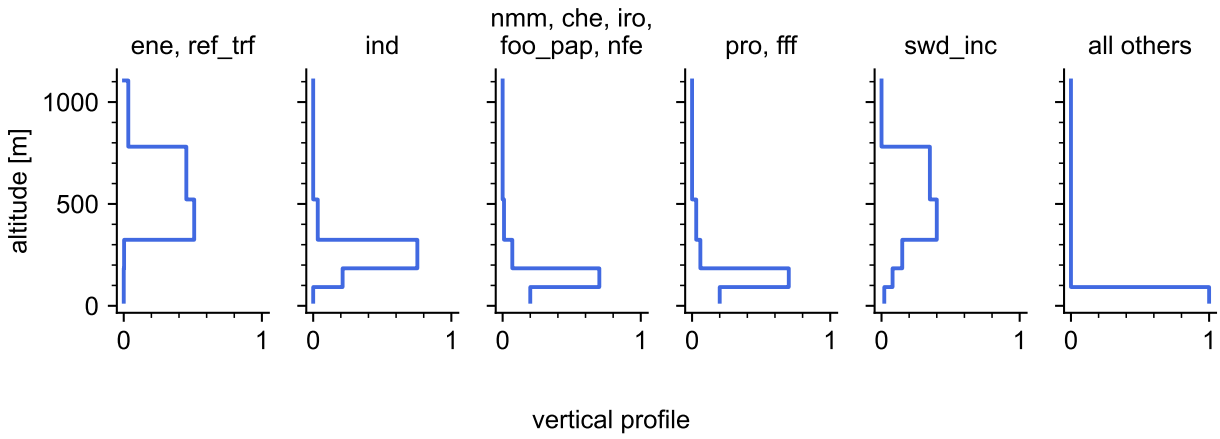

Figure B.4: Overview of the vertical emission profiles used in the WRF-Chem simulation. For the abbreviations in the subplot titles, refer to Table 2.1.

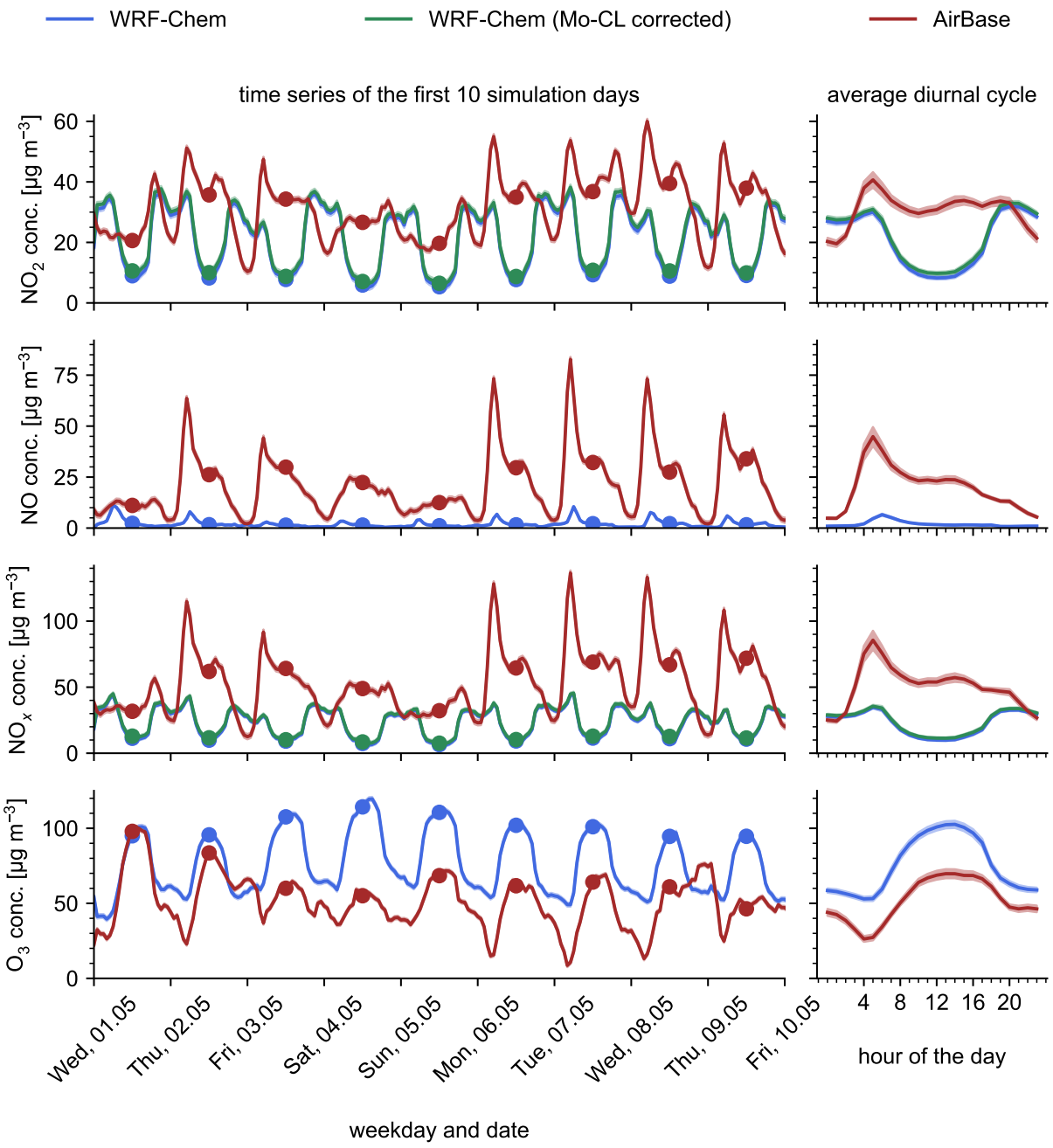

Figure B.5: Like Fig. 3.2, but for traffic measurements instead of background measurements.

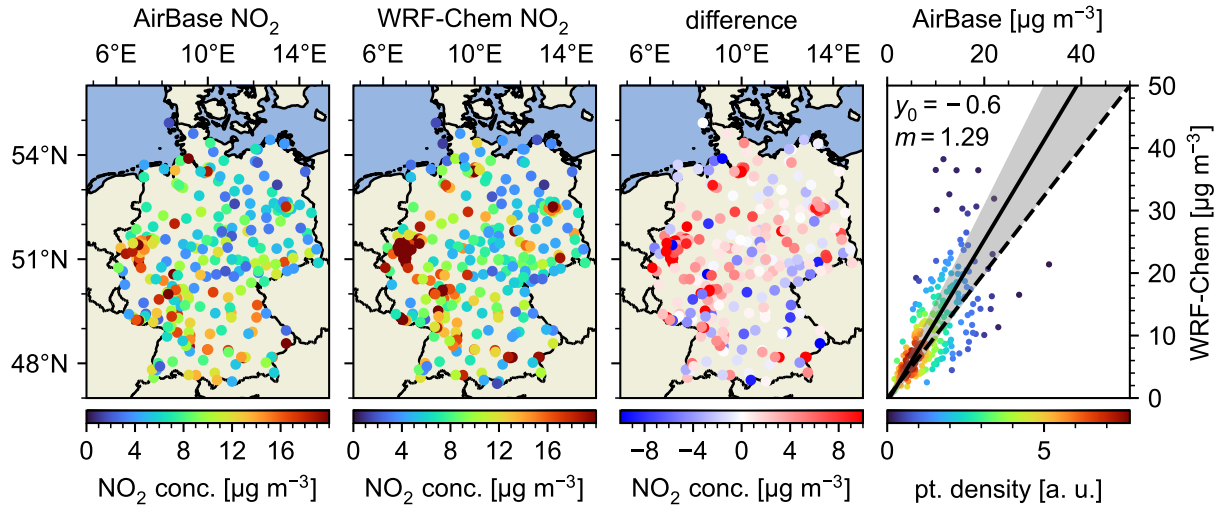

Figure B.6: Like Fig. 3.4a, but for 4 PM instead of noontime.

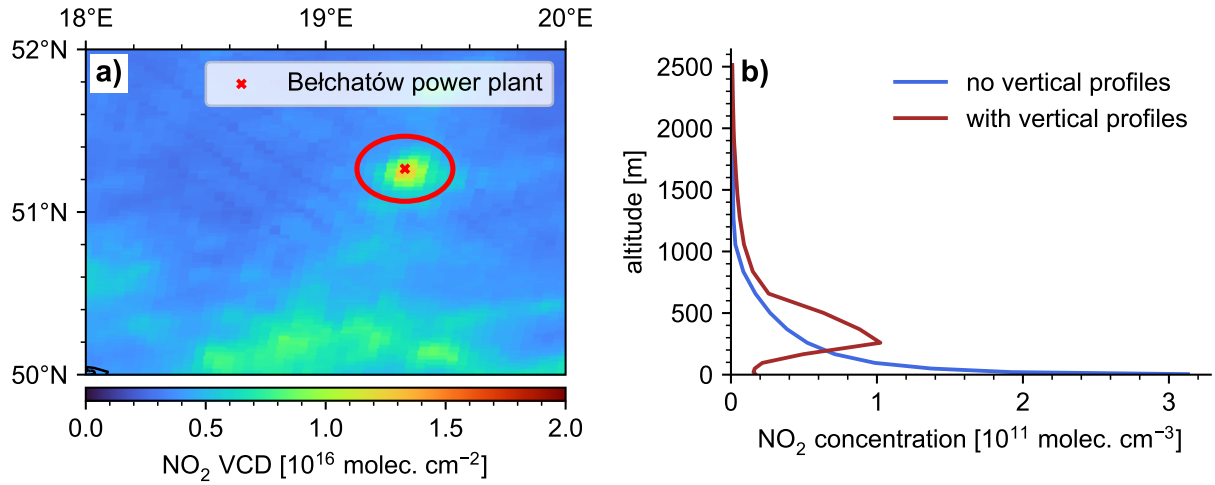

Figure B.7: Impact of vertical emission profiles on WRF-Chem simulation results near a strong emitter (Bełchatów power plant). a) Map of the average NO_2 column density from TROPOMI near the Bełchatów power plant (marked by a red ellipse). The air mass factors were recomputed with the NO_2 profiles from the WRF-Chem simulation run S-YSU-5-5. **b)** NO_2 profiles from the WRF-Chem simulations S-YSU-5-5 (without vertical emission profiles) and S-YSU-5-5-B (with vertical emission profiles), sampled at the center pixel of the Bełchatów power plant (lat = 51.266 °N, lon = 19.330 °E).

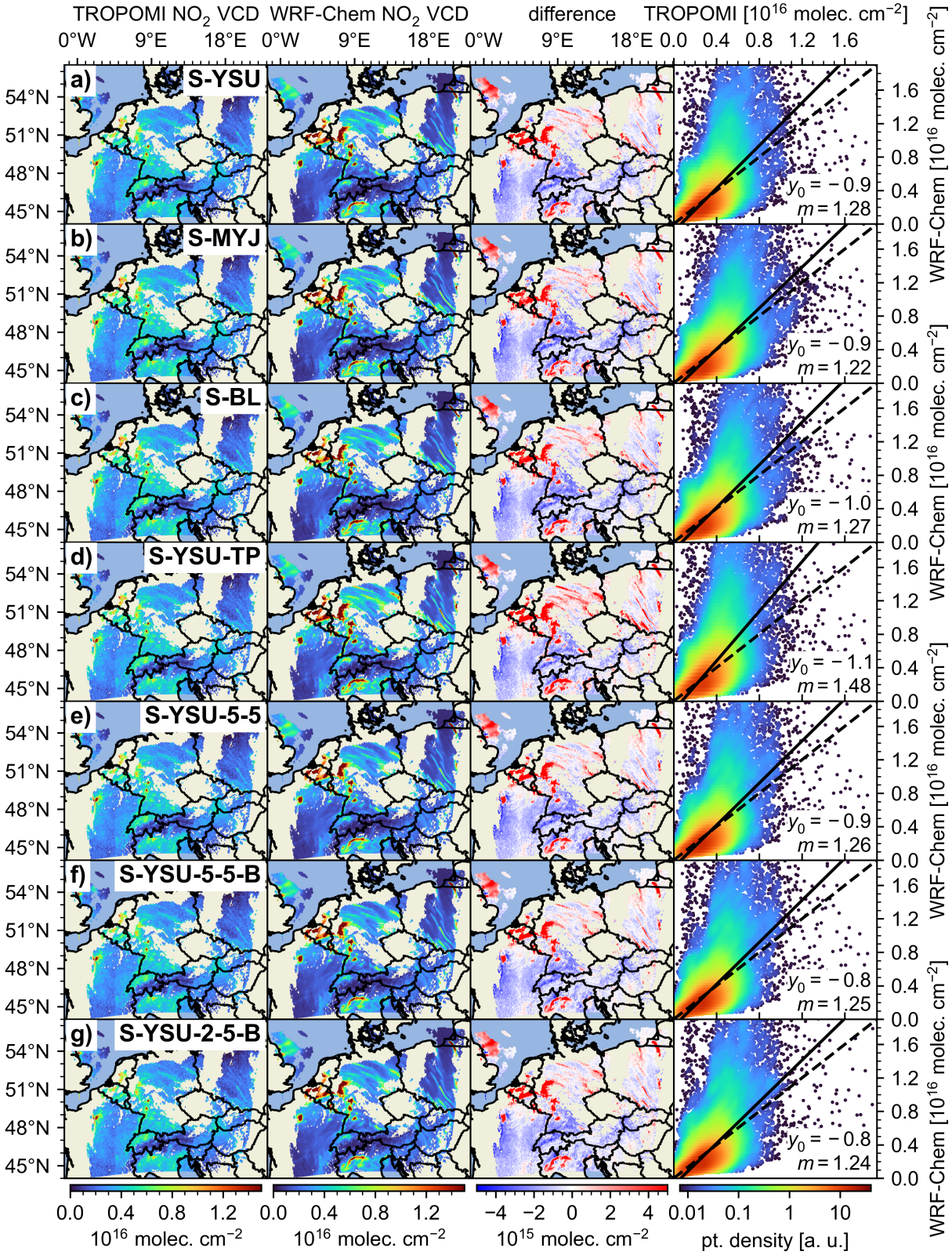

Figure B.8: Like Fig. 3.16, but for a single orbit. Shown here: orbit 08017, recorded on 1 May 2019.

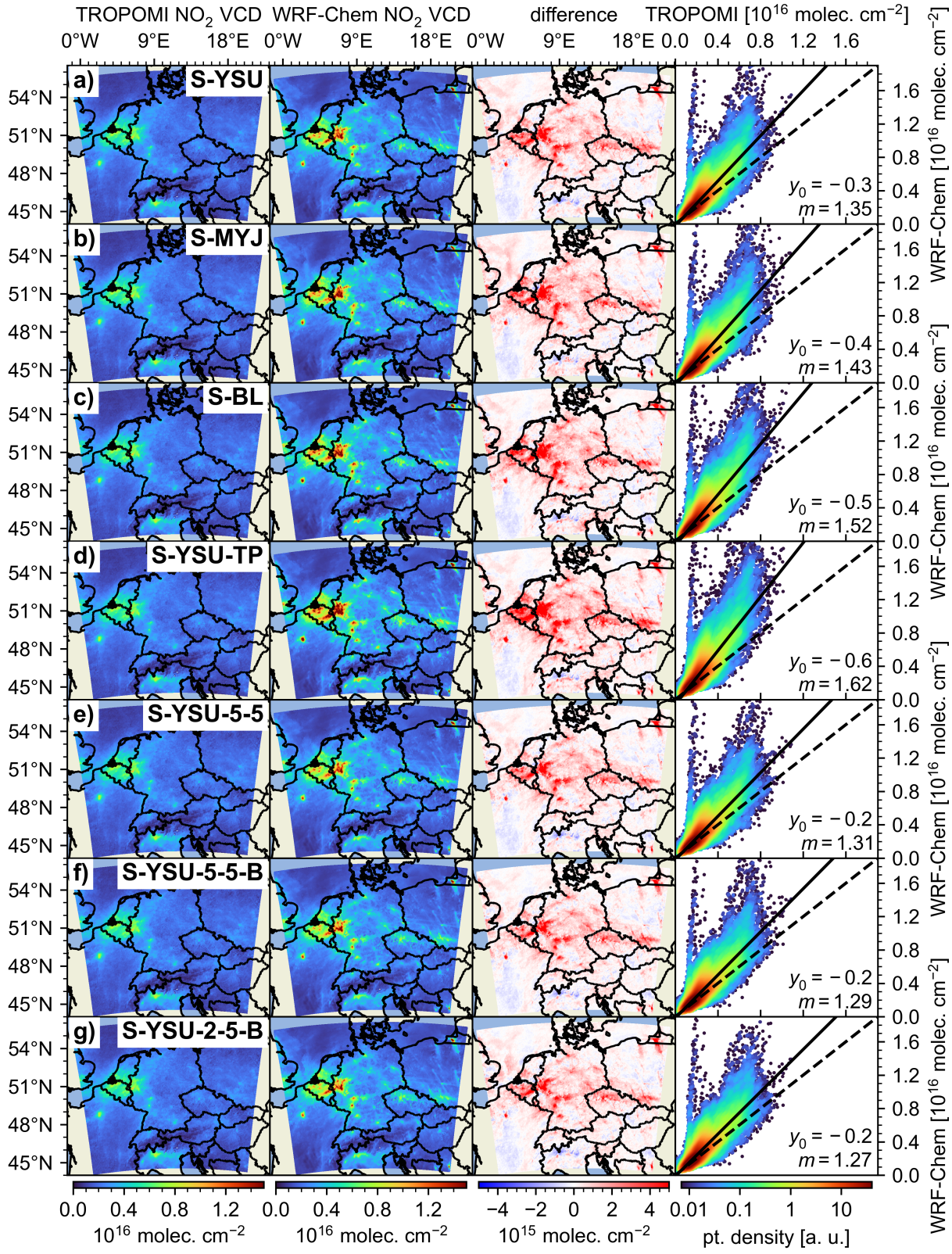

Figure B.9: Like Fig. 3.16, but with the original air mass factors.

Appendix C

The NitroNet model

C.1 Data transformations

In this section, an overview of the data transformations deployed by NitroNet is given. Let X denote any of NitroNet's input or output variables (here to be understood as a random variable). NitroNet supports the following data transformations:

- Manual normalization. The corresponding transformation is defined as

$$\Phi_{\text{man}}(X, \mu, \sigma) = (X - \mu)/\sigma \tag{C.1}$$

  It is conceptually similar to the popular *min-max scaling* (see e.g. Pedregosa et al., 2011) which maps the variable X into the interval $[0, 1]$. However, min-max scaling is not used here, because it can yield subpar transformations in the presence of outliers. In our described manual normalization procedure, μ and σ are chosen manually and disregard outliers, at the cost of softer interval boundaries (i.e. the transformed variable is not strictly confined to $[0, 1]$).

- Automatic normalization, which uses the same transformation formula as the manual normalization, but estimates μ and σ from quantiles of the variable's distribution:

$$\mu = Q(X, 0.01) \tag{C.2}$$
$$\sigma = Q(X - \mu, 0.99) \tag{C.3}$$
$$\Phi_{\text{auto}}(X) = \Phi_{\text{man}}(X, \mu, \sigma) \tag{C.4}$$

  where $Q(X, q)$ denotes the quantile function, that returns the q-quantile of X.

- Quantile transformation, as defined in the scikit-learn documentation (Pedregosa et al., 2011), see the entry of the `sklearn.preprocessing.QuantileTransformer` class.

variable name	units	transformation type	transformation parameters
NO_2 VCD (tropospheric)	10^{16} molec. cm^{-2}	man. norm.[(1)]	$\mu = 0.05$, $\sigma = 0.70$
O_3 VCD (total)	10^{16} molec. cm^{-2}	auto. norm.[(2)]	—
tropospheric air mass factor	—	man. norm.	$\mu = 0.4$, $\sigma = 1$
tropospheric averaging kernels	—	man. norm.	$\mu = 0$, $\sigma = 10$
cloud radiance fraction	—	—	—
cloud pressure	Pa	man. norm.	$\mu = 60000$, $\sigma = 35000$
aerosol absorbing index	—	man. norm.	$\mu = -1.5$, $\sigma = 1.5$
surface albedo	—	man. norm.	$\mu = 0.025$, $\sigma = 0.1$
surface pressure	Pa	auto. norm.	—
sun geometry (zenith and azimuth angle)	° (degrees)	auto. norm.	—
viewing geometry (zenith and azimuth angle)	° (degrees)	auto. norm.	—
surface classification	—	—	—
NO_2 VCD influx	10^{16} molec. cm^{-1} s^{-1}	man. norm.	$\mu = 0$, $\sigma = 7$
planetary boundary layer height (PBLH)	m	man. norm.	$\mu = 200$, $\sigma = 1950$
planetary boundary layer dissipation	J m^{-2}	man. norm.	$\mu = 0$, $\sigma = 52140$
surface temperature	K	auto. norm.	—
vertical velocity (profile)	Pa s^{-1}	auto. norm.	—
wind speed (profile)	m s^{-1}	man. norm.	$\mu = 0$, $\sigma = 28$
NO_x emissions (total)	kg m^{-2} s^{-1}	quant.[(3)]	—
NO_x emissions (SNAP 1)	kg m^{-2} s^{-1}	quant.	—
NO_x emissions (SNAP 3)	kg m^{-2} s^{-1}	quant.	—
NO_x emissions (SNAP 4)	kg m^{-2} s^{-1}	quant.	—
NO_x emissions (surface sources)	kg m^{-2} s^{-1}	quant.	—
day	—	—	—
altitude	m	logarithmic	$a = 0.096$, $b = 2.840$
NO_2 concentration (targets)	molec. cm^{-3}	logarithmic	$a = 0.083$, $b = 1.664 \cdot 10^{-6}$
F (targets)[(4)]	—	—	—

Table C.1: Data transformations in NitroNet.
(1) manual normalization (2) automatic normalization (3) quantile transformation (4) see sect. 4.2.5

- Logarithmic transformation, defined via

$$\Phi(X, a, b) = a \cdot \ln(b \cdot (X + 1)) \tag{C.5}$$

The logarithmic transformation is not strictly positive; e.g. the NO_2 concentrations (targets) of NitroNet are transformed with $a = 0.083$, $b = 1.664 \cdot 10^{-6}$. This returns negative values for $X \lesssim 6 \cdot 10^5$ (in implicit units of molec. cm^{-3}). This poses no problem, because all NO_2 concentrations in NitroNet's training set are above this threshold.

C.2 Supplementary material

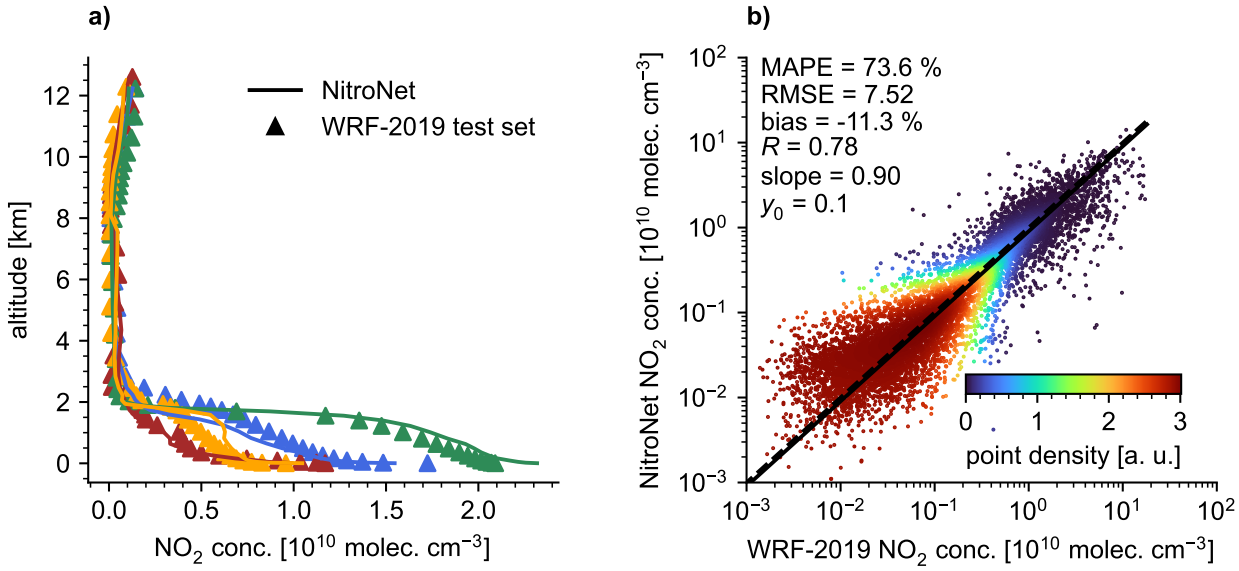

Figure C.1: Like Fig. 4.11, but without application of NitroNet's empirical bias correction.

WRF-2019 layer	altitude [m]	bias [%]	RMSE	MAPE [%]	slope	intercept	R
1	4	−8.0	18.5	36.0	0.7	3.2	0.71
2	20	−8.9	16.9	36.6	0.7	3.3	0.69
3	48	−8.8	15.7	36.8	0.8	2.5	0.67
4	93	−8.5	15.0	37.1	0.7	2.8	0.66
5	161	−7.5	14.3	37.7	0.7	2.6	0.66
6	251	−6.6	13.7	38.5	0.8	1.8	0.66
7	358	−6.3	13.0	39.1	0.8	1.9	0.66
8	487	−6.3	12.2	39.8	0.8	1.8	0.66
9	638	−6.9	11.4	40.6	0.8	1.6	0.66
10	812	−8.3	10.5	41.9	0.8	1.2	0.65
11	1025	−11.4	9.5	44.3	0.8	1.0	0.65
12	1236	−12.9	8.3	63.5	0.8	1.0	0.64
13	1415	−17.3	7.2	63.8	0.8	0.6	0.62
14	1575	−21.1	6.3	64.9	0.7	0.5	0.60
15	1718	−24.2	5.3	64.2	0.7	0.4	0.57
16	1835	−24.3	4.3	63.2	0.7	0.2	0.55
17	1930	−23.0	3.6	60.1	0.7	0.2	0.53
18	2050	−21.1	2.8	57.3	0.7	0.2	0.49
19	2216	−17.0	2.0	57.2	0.7	0.2	0.42
20	2450	−13.2	1.4	55.9	0.7	0.1	0.36
21	2826	−9.8	0.9	54.7	0.6	0.1	0.35
22	3436	−1.6	0.6	64.5	0.4	0.2	0.34
23	4231	+10.8	0.4	87.1	0.2	0.2	0.31
24	5087	+22.4	0.3	116.1	0.1	0.2	0.20
25	5957	+39.5	0.3	147.9	0.1	0.2	0.12
26	6815	+55.9	0.3	178.4	0.1	0.2	0.11
27	7519	+48.4	0.3	184.0	0.1	0.2	0.12
28	8035	+1.6	0.2	119.0	0.2	0.1	0.24
29	8455	−7.5	0.2	112.1	0.3	0.1	0.33
30	8806	−6.4	0.3	123.6	0.3	0.1	0.37
31	9109	−5.2	0.3	132.5	0.4	0.1	0.40
32	9369	−4.7	0.3	135.7	0.5	0.1	0.42
33	9603	−5.7	0.4	127.7	0.5	0.2	0.44
34	9858	−7.5	0.4	106.2	0.5	0.2	0.45
35	10196	−9.1	0.5	79.1	0.5	0.2	0.46
36	10685	−8.2	0.5	53.8	0.6	0.3	0.45
37	11359	−6.5	0.4	31.4	0.6	0.4	0.43
38	12271	−5.5	0.4	20.7	0.4	0.7	0.32

Table C.2: Like Table 4.6, but without application of NitroNet's empirical bias correction.

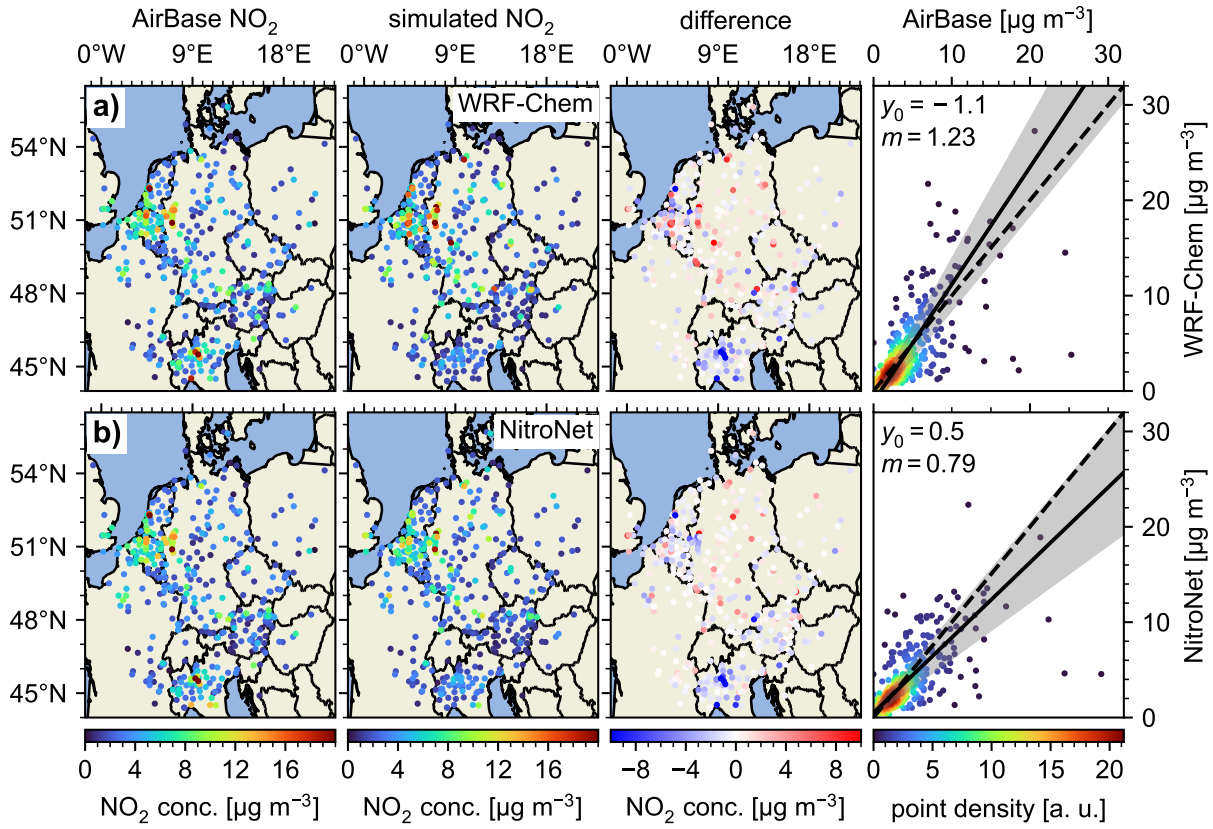

Figure C.2: Like Fig. 4.17, but without urban background instruments.

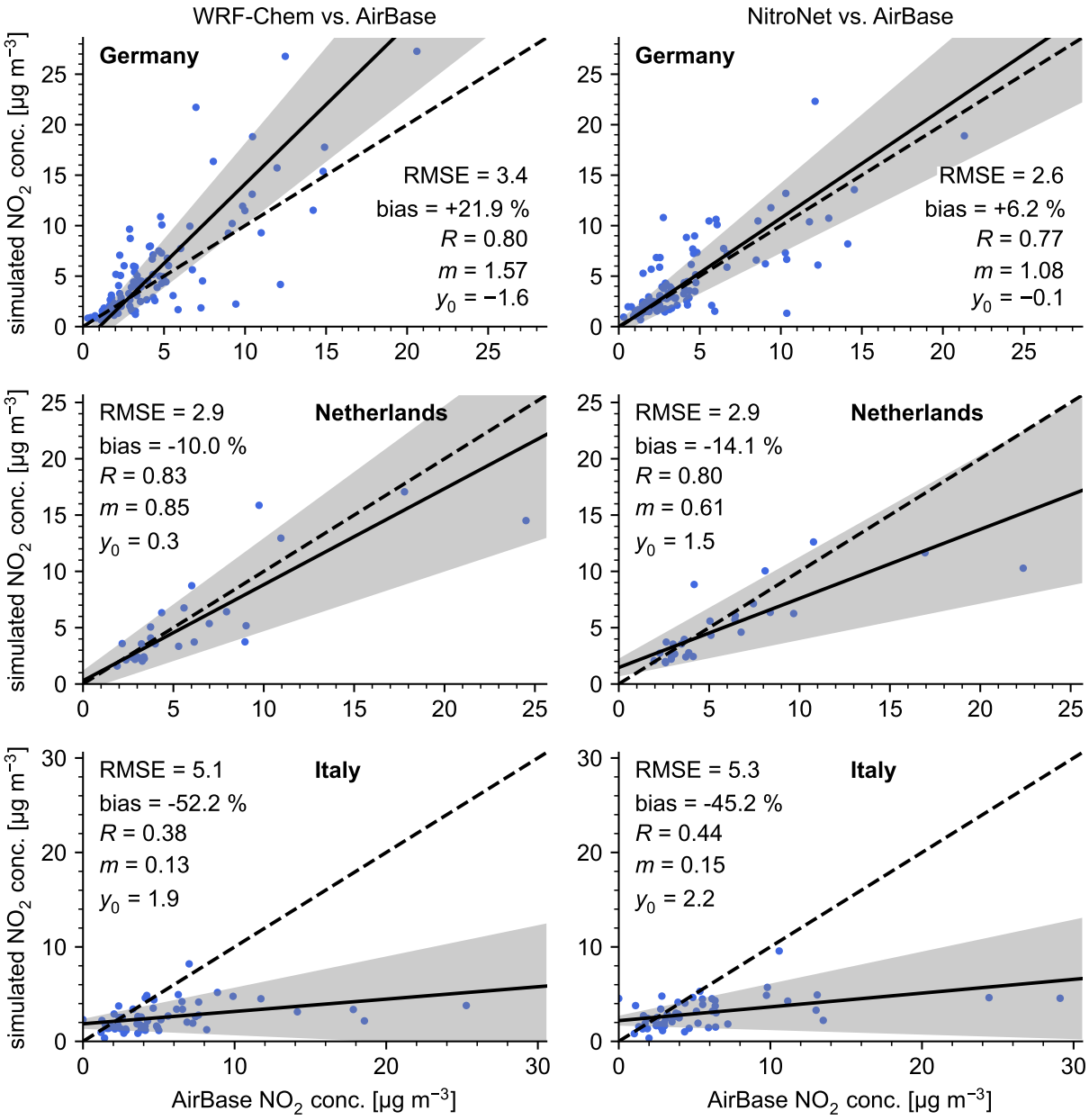

Figure C.3: Like Fig. 4.18, but without urban background instruments.

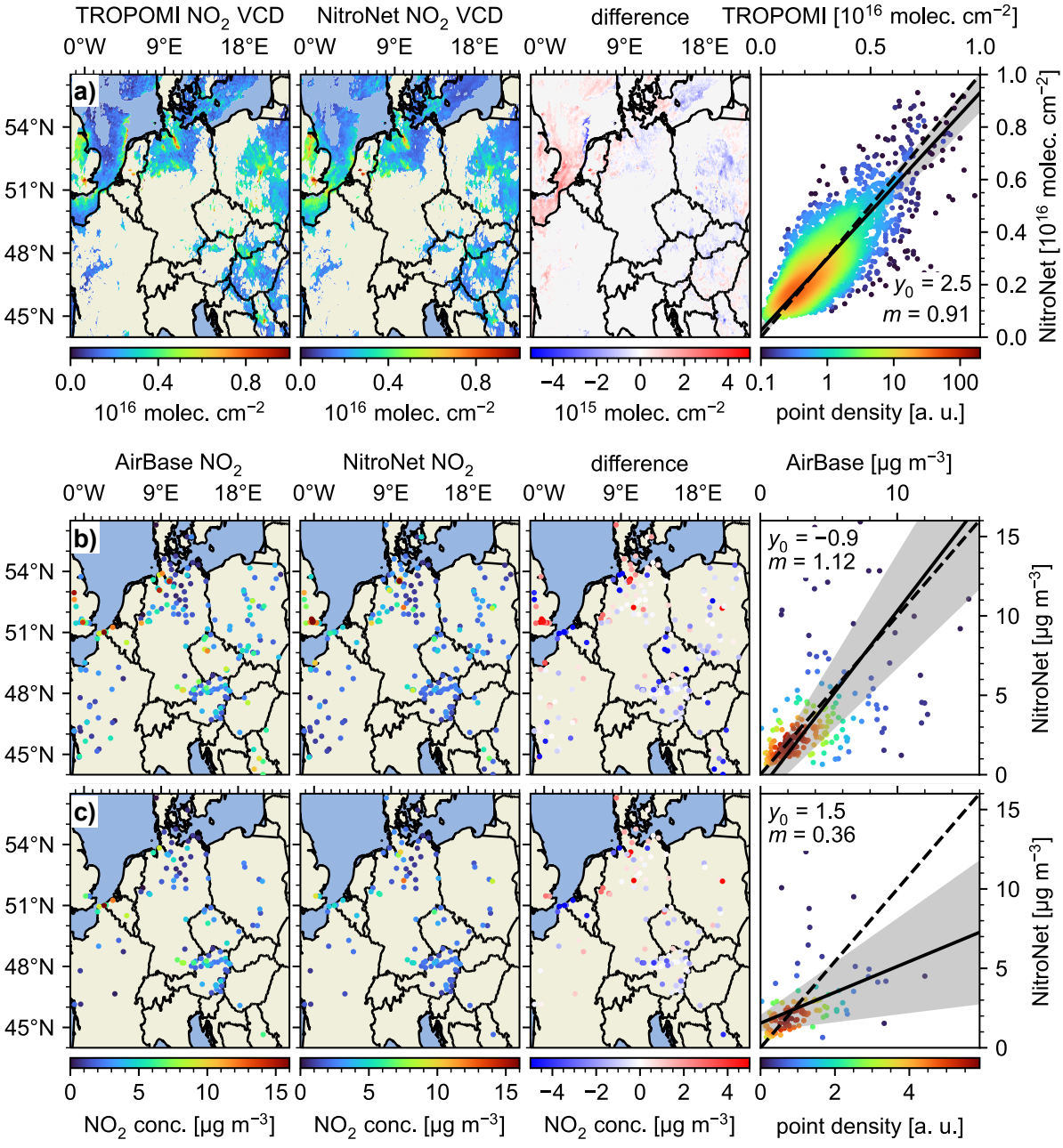

Figure C.4: Like Fig. 4.20, but for a single day (5 May 2022).

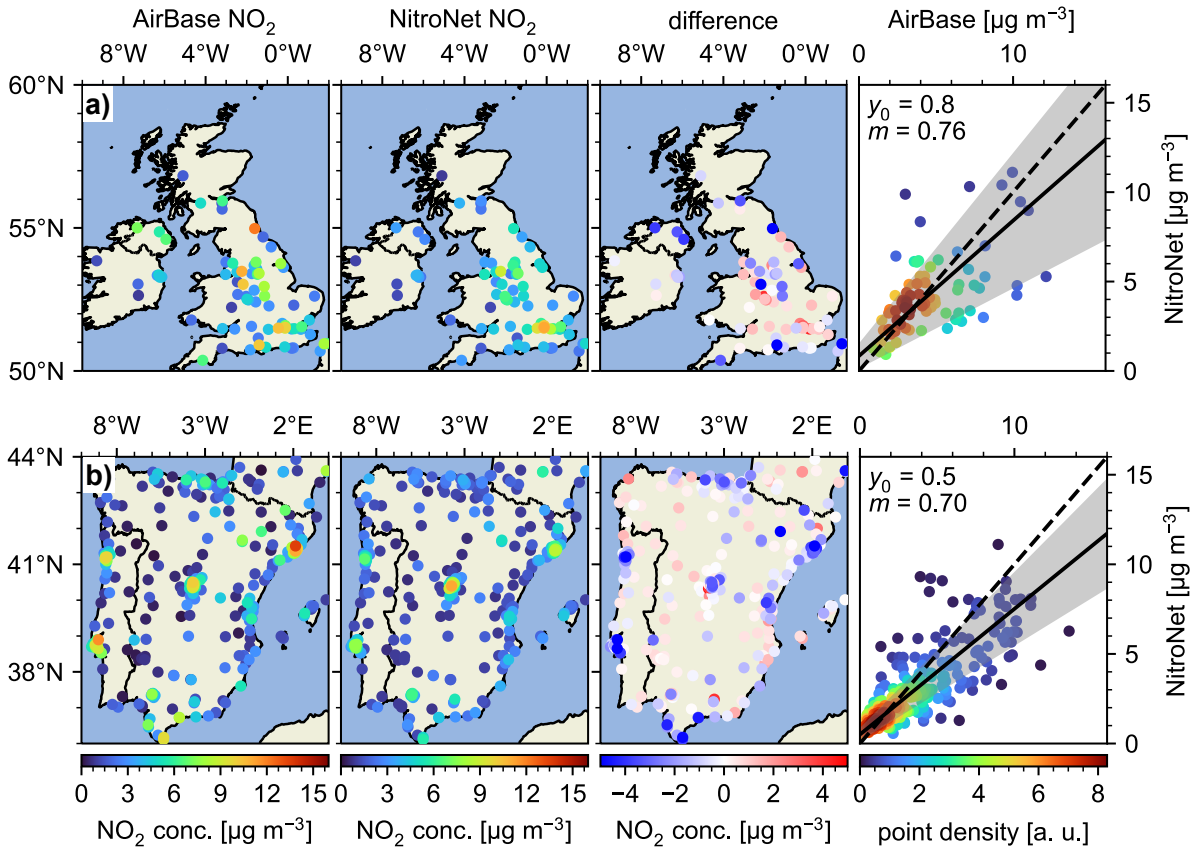

Figure C.5: Like Fig. 4.24, but with urban background stations.

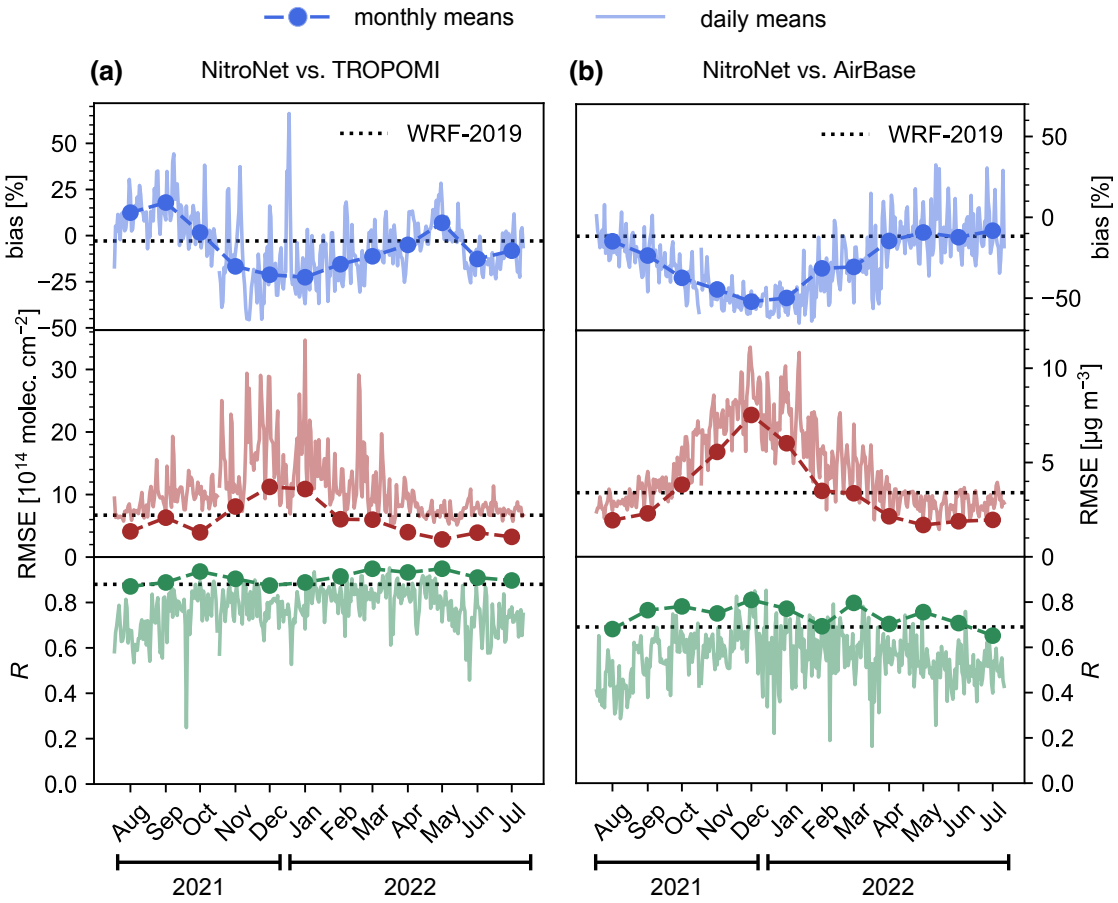

Figure C.6: Like Fig. 4.26, but with urban background stations.

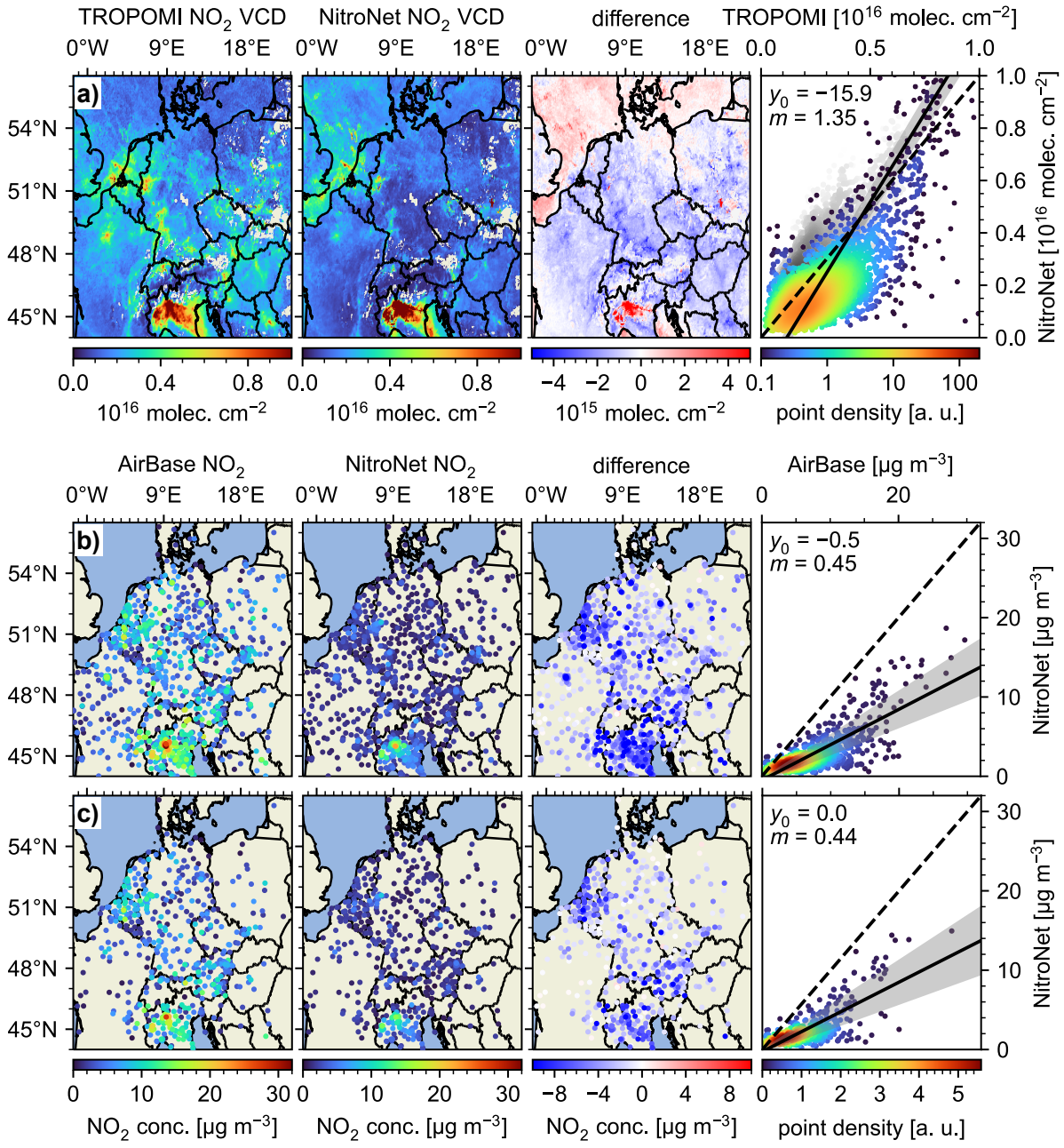

Figure C.7: Like Fig. 4.27, but with winsorization turned off. Note, that this affects the AirBase reference values as well, which depend on NitroNet's predictions of the Mo-CL correction factors.

NitroNet compared to	domain	bias	RMSE	R	reference
TROPOMI (monthly means)	UK	+12.8 %	$3.3 \cdot 10^{14}$ molec. cm^{-2}	0.92	Fig. 4.23a
TROPOMI (monthly means)	ES + PT	+18.4 %	$5.1 \cdot 10^{14}$ molec. cm^{-2}	0.74	Fig. 4.23b
TROPOMI (monthly means)	US	+11.4 %	$4.6 \cdot 10^{14}$ molec. cm^{-2}	0.61	Fig. 4.25a
TROPOMI (monthly means)	India	+65.5 %	$9.1 \cdot 10^{14}$ molec. cm^{-2}	0.90	Fig. 4.25b
TROPOMI (monthly means)	China	+21.2 %	$12.2 \cdot 10^{14}$ molec. cm^{-2}	0.69	Fig. 4.25c
TROPOMI (dec., monthly means)	EU	−21.2 %	$11.2 \cdot 10^{14}$ molec. cm^{-2}	0.87	Fig. 4.27a
TROPOMI (dec., no wins., monthly means)	EU	−21.4 %	$13.0 \cdot 10^{14}$ molec. cm^{-2}	0.68	Fig. C.7a
TROPOMI (individual orbits)	UK	+10.1 %	$6.5 \cdot 10^{14}$ molec. cm^{-2}	0.81	Fig. 4.23a
TROPOMI (individual orbits)	ES + PT	+19.2 %	$8.2 \cdot 10^{14}$ molec. cm^{-2}	0.58	Fig. 4.23b
TROPOMI (individual orbits)	US	+13.4 %	$8.1 \cdot 10^{14}$ molec. cm^{-2}	0.58	Fig. 4.25a
TROPOMI (individual orbits)	India	+58.8 %	$11.4 \cdot 10^{14}$ molec. cm^{-2}	0.82	Fig. 4.25b
TROPOMI (individual orbits)	China	−1.1 %	$18.5 \cdot 10^{14}$ molec. cm^{-2}	0.63	Fig. 4.25c
TROPOMI (dec., individual orbits)	EU	−21.6 %	$18.9 \cdot 10^{14}$ molec. cm^{-2}	0.78	Fig. 4.27a
TROPOMI (dec., no wins., individual orbits)	EU	−17.1 %	$27.9 \cdot 10^{14}$ molec. cm^{-2}	0.60	Fig. C.7a

Table C.3: Supplement to Tables 4.12 and 4.13. Shown here are the full statistical diagnostics with water pixels included. “dec.” refers to an evaluations in December 2021. “no wins.” refers to evaluations without winsorization. All other entries refer to regular evaluations for the month of May 2022.

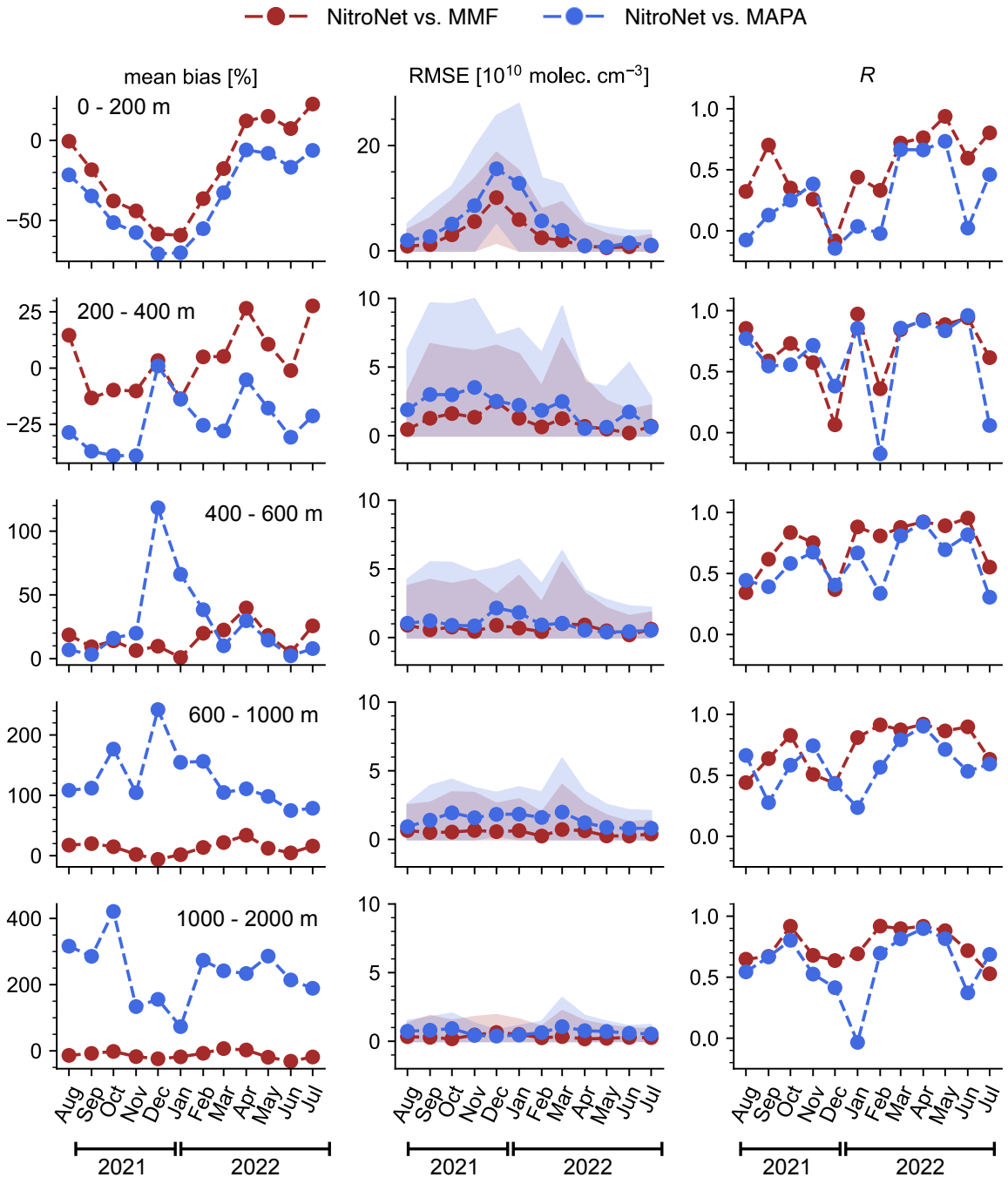

Figure C.8: Like Fig. 4.28, but without error bands for the mean bias.

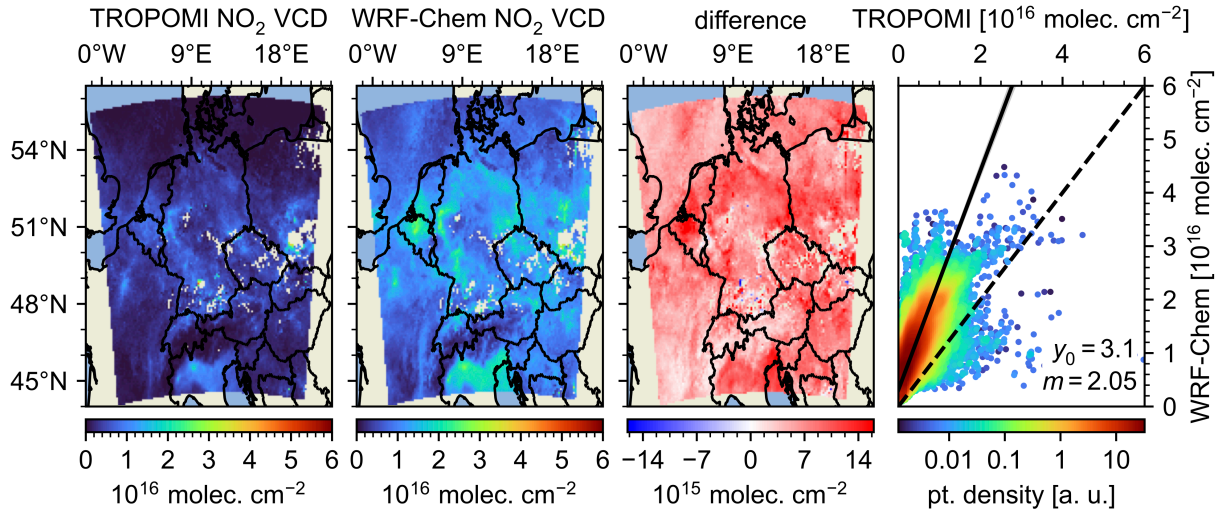

Figure C.9: Validation of simulated monthly-mean tropospheric NO_2 VCDs from WRF-Chem against TROPOMI observations, February 2019. The simulation setup S-YSU-2-5-B (see Table 3.4) was used. The intercept (y_0) is given in units of 10^{15} molec. cm^{-2}.