https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

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

Leon Kuhn[1, 2], Steffen Beirle[2], Sergey Osipov[2, 3], Andrea Pozzer[2], and Thomas Wagner[1, 2]

[1]Institute for Environmental Physics, University of Heidelberg, Germany
[2]Satellite Remote Sensing Group, Max-Planck Institute for Chemistry, Mainz, Germany
[3]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Correspondence: Leon Kuhn (l.kuhn@mpic.de)

Abstract.

We introduce "NitroNet", a deep learning model for the prediction of tropospheric NO_2 profiles from satellite column measurements. NitroNet is a neural network, which was trained on synthetic NO_2 profiles from the regional chemistry and transport model WRF-Chem, operated on a European domain for the month of May 2019. This WRF-Chem simulation was constrained by in-situ and satellite measurements, which were used to optimize important simulation parameters (e.g. the boundary layer scheme). The NitroNet model receives vertical NO_2 column densities (VCDs) from the TROPOMI satellite instrument and ancillary variables (meteorology, emissions, etc.) as input, from which it reproduces NO_2 concentration profiles. Training of the neural network is conducted on a filtered dataset, meaning that NO_2 profiles with strong disagreement (> 20 %) to colocated TROPOMI column measurements are discarded.

We present a first evaluation of NitroNet on a variety of geographical domains (Europe, US west coast, India, and China) and different seasons. For this purpose, we validate the NO_2 profiles predicted by NitroNet against satellite, in-situ, and MAX-DOAS measurements. The training data were previously validated against the same datasets. During summertime, NitroNet shows small biases and strong correlations to all three datasets ($\text{bias} = +6.7$ % and $R = 0.95$ for TROPOMI NO_2 VCDs, $\text{bias} = -10.5$ % and $R = 0.75$ for AirBase surface concentrations). In the comparison to TROPOMI satellite data, NitroNet even shows significantly lower errors and stronger correlation than a direct comparison with WRF-Chem numerical results. During wintertime considerable low biases arise, because the summertime training data is not fully representative of all atmospheric wintertime characteristics (e.g. longer NO_2 lifetimes). Nonetheless, the wintertime performance of NitroNet is surprisingly good, and comparable to that of classic RCT models. NitroNet can demonstrably be used outside the geographic domain of the training data with only slight performance reductions. What makes NitroNet unique compared to similar existing deep learning models is the inclusion of synthetic model data, which has important benefits: Due to the lack of NO_2 profile measurements, models trained on empirical datasets are limited to the prediction of surface concentrations learned from in-situ measurements. NitroNet, however, can predict full tropospheric NO_2 profiles. Furthermore, in-situ measurements of NO_2 are known to suffer from biases, often larger than +20 %, due to cross sensitivities to photooxidants, which other models trained on empirical data inevitably reproduce.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

1 Introduction

Nitrogen oxides (NO_x = NO + NO_2) are an important marker of air pollution. The negative impact of NO_2 on human health has been widely recognized (see e.g. Faustini et al. (2014); Mills et al. (2015); Chowdhury et al. (2021)). In many European countries, the recommended annual-average exposure limit of 10 μg m^{-3} is exceeded continuously. Active monitoring of tropospheric NO_2 is a crucial step in identifying pollution hotspots, localizing emissions, and designing long-term solutions to the pollution problem. Different NO_2 measuring methods exist. Many countries across the world deploy in-situ measurements at the surface (see e.g. the AirBase network, European Environment Agency). The TROPOMI satellite instrument (see Veefkind et al. (2012)) yields measurements of the tropospheric NO_2 vertical column density (VCD) with daily near global coverage and a ground pixel size of up to $3.5\ \mathrm{km} \times 5.5\ \mathrm{km}$. Lastly, ground-based MAX-DOAS measurements ("multi-axis differential optical absorption spectroscopy", see Platt and Stutz (2008); Hönninger et al. (2004)) are used to obtain tropospheric NO_2 profiles in a few selected places, by means of scanning the troposphere at different elevation angles. Although further measuring platforms (e.g. sondes, aircraft) and methods (e.g. LIDAR instruments) exist, these are not routinely deployed (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)). Altogether, these measurements are valuable for the quantification of tropospheric vertical column densities, surface concentrations, and to some extent the tropospheric profile shapes. Nonetheless, the described methods also have drawbacks:

- The TROPOMI instrument can measure the tropospheric column density, but it can not resolve along the lightpath or the vertical axis, meaning it can principally not return vertical NO_2 profiles. Furthermore, the TROPOMI NO_2 VCD retrieval depends on a priori profiles. In the operational TROPOMI processor, these are taken from the TM5-MP model (see Krol et al. (2005)), whose low horizontal resolution of $1° \times 1°$ is known to be the main cause of significant negative biases of typically -10 to -20 % (see Ialongo et al. (2020); Tack et al. (2021); Liu et al. (2021); Douros et al. (2023)), but in some cases even up to -50% (Lange et al. (2023)). Alternative data products with higher resolved a priori profiles exist, but are not available globally.

- In-situ measurements often utilize the molybdenum-based chemiluminescence method, which is known for its severe cross sensitivities to other atmospheric oxidants, causing large biases in the reported NO_2 concentrations (see Dunlea et al. (2007); Steinbacher et al. (2007); Lamsal et al. (2008); Boersma et al. (2009); Villena et al. (2012)). These biases typically range from $+20$ % to $+100$ %, but Villena et al. (2012) even report biases of up to $+300$ % in extreme cases. As described in detail further into the manuscript, these biases can be strongly reduced down to a few percent within our model framework.

- MAX-DOAS measurements are quite sparsely located and can not provide dense spatial coverage. Additionally, the commonly used retrieval algorithms are only sensitive to the atmospheric layers close to the ground (up to $\sim 1\ \mathrm{km}$), but not to the upper troposphere.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Measurements are therefore often complemented by regional chemistry and transport (RCT) simulations. Examples of state-of-the-art RCT models are WRF-Chem (Grell et al. (2005)), COSMO/MESSy (Kerkweg and Jöckel (2012)), Lotos-Euros (Manders et al. (2017)), CAM-chem (Emmons et al. (2020)), and CHIMERE (Menut et al. (2021)). Such models can simulate realistic distributions of NO_2 and other atmospheric trace gases with horizontal resolutions on the scale of $3\ \mathrm{km} \times 3\ \mathrm{km}$ and vertical resolutions of $\sim 1\ \mathrm{m}$ at the surface, to $\sim 1\ \mathrm{km}$ in the upper troposphere. High-resolution RCT simulations can be used to estimate air pollution in the absence of in-situ measurement, and to obtain better resolved a priori profiles for the TROPOMI retrieval. Unfortunately, the continuous deployment of RCT simulations is no easy endeavour, due to their computational expense, dependence on input data which may not always be available in an up-to-date form at high resolution (in particular emission data), and the uncertainty in choice of simulation parametrizations. Another point of concern is the general accuracy of these models: RCT simulations reported in recent literature have shown significant deviations from observational reference data (see Visser et al. (2019); Kuik et al. (2016); Kuik et al. (2018); Poraicu et al. (2023)), e.g. an underestimation of the summertime surface-level NO_2 concentration of up to $-50\ \%$. A study by Douros et al. (2023) reveals overestimations of the winter-time NO_2 VCD by $+50\ \%$, and demonstrates that such biases even occur in ensemble models, such as CAMS (consisting of 11 different RCT models with $0.1° \times 0.1°$ horizontal resolution). In previous work, we showed that a recalibration of the vertical mixing parametrization can mostly resolve such biases in the WRF-Chem model in summer over Europe (see Kuhn et al. (2024)). However, the process of model recalibration is tedious, computationally expensive, and domain-dependent. Altogether, it can be concluded that high-resolution RCT simulations are of undisputed benefit, but their practical realization remains challenging.

In this article we introduce "NitroNet", a new machine learning model intended to complement existing RCT models and measurements of NO_2. NitroNet is a feed-forward neural network, which was designed to predict full tropospheric NO_2 profiles using TROPOMI VCDs alongside other ancillary data (meteorology, emissions, surface types, etc.) as input. Because neural networks are universal function approximators, they are the ideal tool to capture such complex data relationships. NitroNet is trained on numerically simulated data from the WRF-Chem model, operated on a European domain for the month of May 2019 as described in Kuhn et al. (2024). A data filtering scheme is used to ensure that only well-validated results from the WRF-Chem simulation are used for training the neural network, e.g. training examples with significant disagreement to co-located satellite observations are dismissed. Afterwards, NitroNet is used as a standalone model, without the necessity to run the RCT simulation again. NitroNet expands on previous deep learning models trained on empirical data (see e.g. Gardner and Dorling (1999); Kang et al. (2021); Chan et al. (2021); Ghahremanloo et al. (2021); Zhang et al. (2022); Jesemann et al. (2022); Cao (2023); all presented models were trained on in-situ surface observations) by inclusion of synthetic model data. This approach provides intrinsic advantages: Firstly, NitroNet can predict full NO_2 profiles, while models trained on empirical data can only be used for surface predictions. Secondly, the chemical mechanisms of RCT models allow for the explicit treatment of in-situ measurement biases (typically larger than $+20\ \%$) by computation of suitable correction factors, while empirically trained models can not compute such correction factors and inevitably reproduce the biases immanent to the training data. Thirdly, synthetic datasets of NO_2 profiles are typically much larger than the few empirical data, and also cover the spatial

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

domain continuously. This allows for the use of highly selective training data filtering, which demonstrably improves the neural network's performance.

The article is structured as follows: Section 2 gives an overview of the datasets used in our study. Section 3 gives a detailed explanation of the NitroNet model. Section 4 shows an evaluation of NitroNet against satellite, in-situ, and MAX-DOAS data on a European domain for May 2022 (i.e. on input data, which the neural network has never seen before). This study is then extended to different seasons and geographical domains (UK, Spain + Portugal, US west coast, India, and China). Section 5 concludes.

2 Datasets

The following datasets are used in our study:

2.1 Vertical NO_2 profiles from WRF-Chem

An RCT simulation using the WRF-Chem model (v. 4.2.2, see Grell et al. (2005) provides the NO_2 profiles on which NitroNet is trained. The simulation was run for the month of May 2019 on a domain over Europe with a spatial resolution of $3\ \mathrm{km} \times 3\ \mathrm{km}$, 43 terrain-following pressure levels, and hourly output. A detailed description, discussion, and validation study of this dataset was published in Kuhn et al. (2024). The simulation setup additionally deploys the vertical emission profiles from Bieser et al. (2011). We will refer to this dataset as "WRF-2019" from hereon. WRF-2019 contains approximately two million NO_2 profiles, which are split into three partitions: A training set (80 %), a validation set (15 %), and a test set (5 %). The training set is used for training NitroNet (described in sect. 3.3), the validation set for hyperparameter optimization (described in sect. 3.2 and Appendix A), and the test set for evaluation of the neural network on previously unseen data.

2.2 Input data for NitroNet

NitroNet uses tropospheric NO_2 vertical column densities (VCDs) from the TROPOMI satellite instrument as the main input. Additionally, although much less influential, total O_3 VCDs are used. The TROPOMI instrument on board of the S5P satellite observes spectra of backscattered light from space with near global coverage, a daily overpass at around 13:30 local time, and a pixel size of up to $3.5 \times 5.5\ \mathrm{km}$ (see Veefkind et al. (2012); van Geffen et al. (2022)). The retrieval of tropospheric NO_2 VCDs is comprised of three steps: First, the NO_2 total slant column density is obtained from the observed light spectra using differential optical absorption spectroscopy (DOAS, see Platt and Stutz (2008)). Then, the obtained total SCD is separated into a stratospheric and a tropospheric component ($\mathrm{SCD_{trop}}$). Finally, the tropospheric VCD is obtained by computing

$$\mathrm{VCD_{trop}} = \frac{\mathrm{SCD_{trop}}}{\mathrm{AMF_{trop}}} \tag{1}$$

where $\mathrm{AMF_{trop}}$ denotes the tropospheric air mass factor. Air mass factors are computed using an altitude-dependent look-up table together with simulated NO_2 a priori profiles from the RCT model TM5-MP (see Krol et al. (2005)) with a horizontal resolution of $1° \times 1°$. The process is described by van Geffen et al. (2022). Throughout our study, we only use data with a

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

high "quality assurance" value ($f_{\mathrm{QA}} > 0.75$), which is the general recommendation (see Eskes et al. (2019)). This also acts as a cloud filter, as it removes observations with cloud fractions of above 50 %. Throughout the rest of the paper, "$\mathrm{NO_2}$ VCD" refers to the *tropospheric* $\mathrm{NO_2}$ VCD, and "$\mathrm{O_3}$ VCD" refers to the *total* $\mathrm{O_3}$ VCD.

Additionally, NitroNet uses meteorological variables from the ERA5 reanalysis ($0.125° \times 0.125°$, see Hersbach and Dee (2017)) and emission data from the EDGARv5 global emission inventory ($0.1° \times 0.1°$, see Crippa et al. (2020)) as input data.

2.3 Validation data for NitroNet

The following three datasets are used to evaluate the NitroNet model:

1. The aforementioned tropospheric $\mathrm{NO_2}$ VCDs from the TROPOMI satellite instrument.

2. In-situ surface measurements of $\mathrm{NO_2}$ from the European AirBase instrument network (see European Environment Agency). This dataset is assembled from the submissions of individual countries of the European Union. The measurements are available as hourly mean values and are classified into three groups: background, traffic, and industrial. Traffic and industrial stations are typically located directly next to strong sources (e.g. near large streets or power plants), where strong horizontal $\mathrm{NO_2}$ gradients occur on the scale of a few meters (see e.g. Beckwith et al. (2019)). Such gradients can neither be resolved by TROPOMI, whose observations are used as input data, nor by WRF-Chem, whose simulation results were used for training NitroNet. Therefore, only background stations are included in our validation study.

3. $\mathrm{NO_2}$ concentration profiles from MAX-DOAS instruments, operated within the FRM$_4$DOAS project in Europe (see Fayt et al. (2021)). FRM$_4$DOAS uses the optimal-estimation based Mexican MAX-DOAS fit (MMF, see Friedrich et al. (2019)), and the Mainz Profile Algorithm (MAPA, see Beirle et al. (2019)) for profile inversion. The resulting $\mathrm{NO_2}$ profiles are defined on a vertical grid with ~ 200 m spacing, reaching to altitudes of up to 4 km. Each instrument produces approximately five $\mathrm{NO_2}$ profiles per hour.

3 NitroNet model description

The NitroNet model consists of an artificial neural network at its core and deploys additional non-machine learning code for efficient data pre-processing and Monte Carlo uncertainty estimation on high performance computing (HPC) architectures. NitroNet's neural network uses the feed-forward topology and is trained with the standard backpropagation method (see Rumelhart et al. (1986)). It has one output neuron, which is used to predict a single $\mathrm{NO_2}$ concentration value per query. Full $\mathrm{NO_2}$ profiles are obtained by concatenating multiple queries on a vertical grid of the user's choice. Throughout this article, a vertical grid with 186 levels is used, resulting in vertical resolutions of ~ 1 m near the surface, ~ 50 m up until 4 km altitude, and up to 400 m in the regions between 4 and 8 km altitude.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

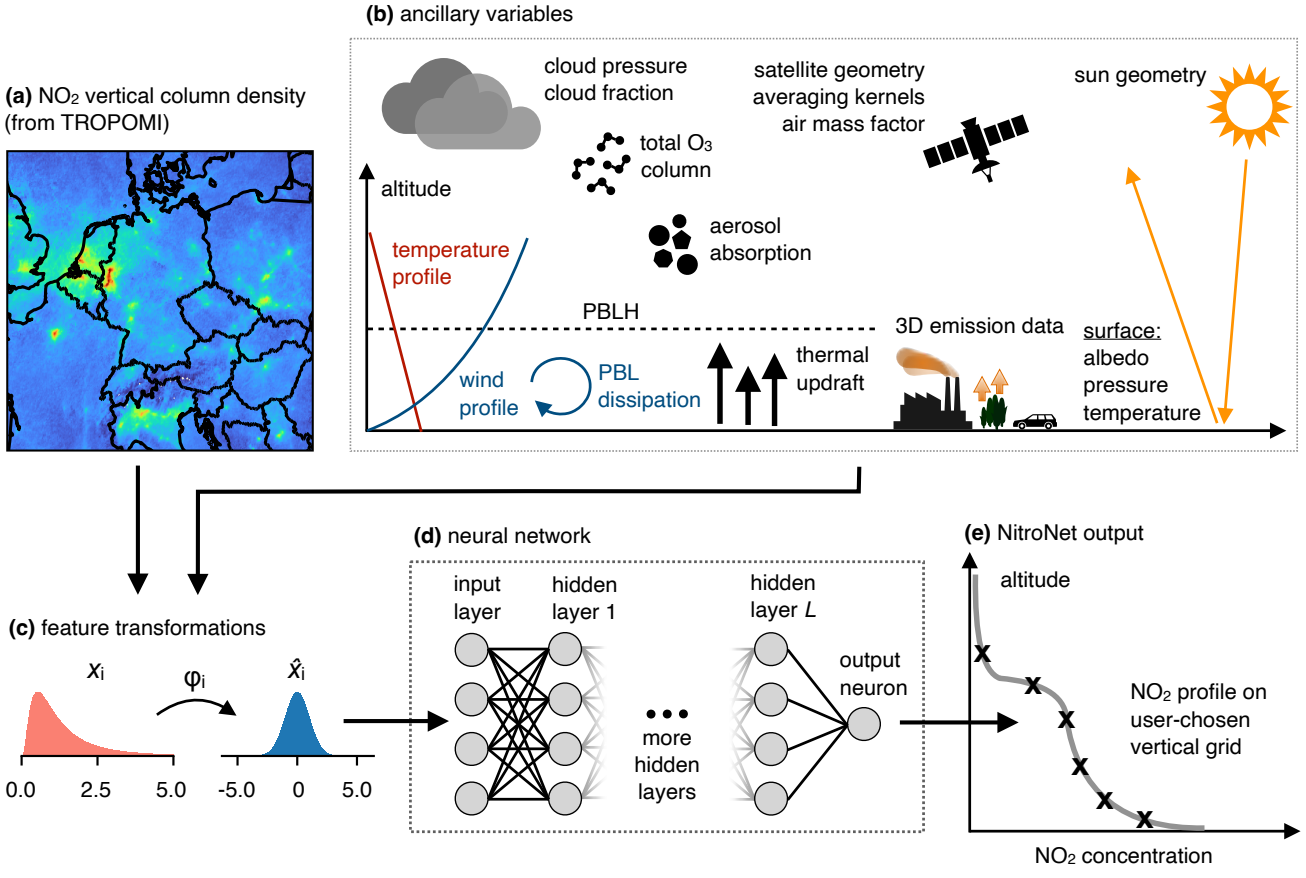

Figure 1. Overview of the NitroNet model. (a) and (b) depict the various input variables, which undergo feature transformation (c) before entering NitroNet's neural network (d). The output of the neural network is an NO_2 profile on a vertical grid of the user's choice.

3.1 Description of the model input

The purpose of our model is to provide realistic NO_2 profiles without the necessity to run computationally expensive RCT simulations. For this reason it is imperative, that NitroNet is only trained on variables from sources accessible both at training and runtime. This may include simulation data from other operational models (e.g. the planetary boundary layer height (PBLH) from ERA5), but excludes many potentially informative variables exclusive to WRF-2019 (e.g. various trace gas concentrations). The training targets (i.e. the NO_2 profiles) are exempt from this rule, because they can only be obtained from WRF-Chem.

Table 1 gives an overview of all input variables ("features") to the neural network. For the NO_2 and O_3 VCDs, the most recent TROPOMI product version (2.04) is used. Tropospheric averaging kernels (AKs) are computed according to Eskes et al. (2019) and defined on the vertical grid of the TM5 model. NitroNet uses the tropospheric AKs at the 9 lowest TM5 layers

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

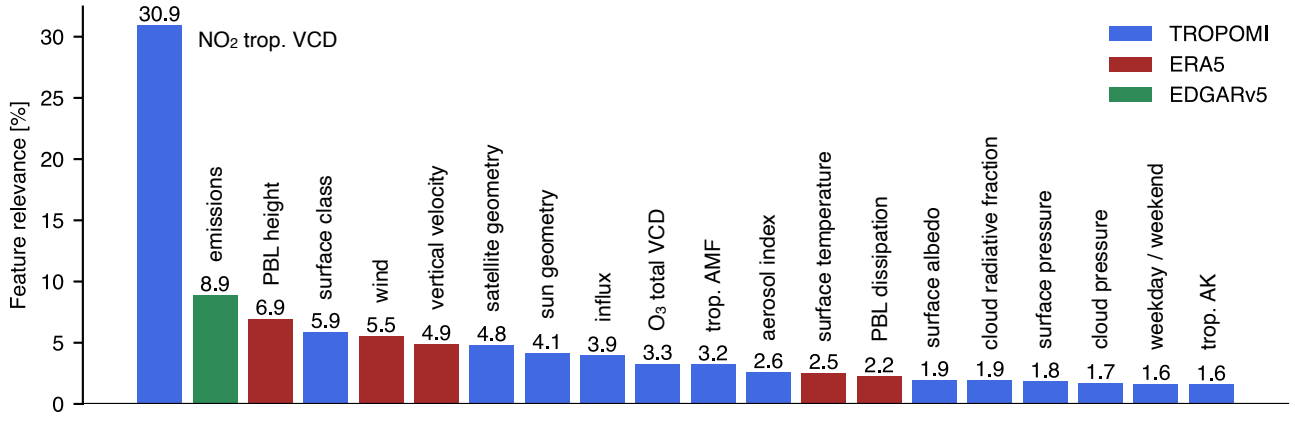

Figure 2. Feature relevance analysis of the NitroNet model. The legend in the top right indicates the data source of each input group.

(reaching up to $\sim 2300\ \mathrm{m}$ altitude), although in hindsight, it was discovered that the AKs contribute only very little to the overall prediction quality, most likely due to the redundancy with other input variables (cloud data, surface albedo, sun zenith angle, etc.). The ERA5 variables "wind speed" and "vertical velocity" are vertically resolved at 1000 hPa, 950 hPa, 900 hPa, 850 hPa, 750 hPa, and 700 hPa. "Wind speed" refers to the absolute wind speed profile, i.e. $\sqrt{u^2+v^2}$, where u, and v are the northward and eastward wind speeds, respectively. "Boundary layer dissipation" is an ERA5 variable, which measures the conversion of kinetic energy into heat due to small-scale eddies in the planetary boundary layer (PBL). NitroNet receives NO_x emissions from the EDGARv5 emission inventory, along with the corresponding relative contribution of four emission bins based on the "Standard Nomenclature for Air Pollution" (SNAP, see European Environment Agency (2023)). The intent is to inform the neural network 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", by which we refer to e.g. road traffic or agricultural emissions. NitroNet uses a ternary surface classification (urban, cropland, forest), which is available within the TROPOMI NO_2 product. The "VCD influx" variable represents the amount of NO_2 that an observed TROPOMI pixel receives from its eight immediate neighbouring pixels due to advection. The corresponding wind speeds are taken from the ERA5 reanalysis.

An in-depth analysis of the "feature importance" of each input variable was conducted, see Fig. 2. The intention is to compute the relevance of each input variable for the model's prediction quality in a rigorous manner, here using the so-called *Shapley scores*. As expected, the NO_2 VCD is by far the most important input feature ($F = 30.9\ \%$), followed by the emission variables ($F = 8.9\ \%$) and the PBLH ($F = 6.9\ \%$). A detailed explanation, and further interpretation are found in Appendix B.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Table 1. NitroNet's input variables

Input variable name	Data source	Note
NO_2 VCD (tropospheric)	TROPOMI	v. 2.04
O_3 VCD (total)	TROPOMI	v. 2.04
tropospheric air mass factor	TROPOMI	
tropospheric averaging kernels	TROPOMI	9 lowest TM5 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	
satellite viewing geometry (zenith and azimuth angle)	TROPOMI	
planetary boundary layer height (PBLH)	ERA5	
planetary boundary layer dissipation	ERA5	
surface temperature	ERA5	
vertical velocity	ERA5	see https://codes.ecmwf.int/grib/param-db/?id=135
wind speed	ERA5	total absolute wind speed, i.e. $\sqrt{u^2+v^2}$
NO_2 emissions (total)	EDGARv5	
NO_2 emissions (rel. contribution from SNAP 1)	EDGARv5	
NO_2 emissions (rel. contribution from SNAP 3)	EDGARv5	
NO_2 emissions (rel. contribution from SNAP 4)	EDGARv5	
NO_2 emissions (rel. contribution from surface sources)	EDGARv5	
surface classification (urban / cropland / forest)	TROPOMI	ternary mask
day	—	binary mask ($0 =$ weekday, $1 =$ weekend)
VCD influx	TROPOMI + ERA5	
vertical grid	—	vertical grid, on which the resulting NO_2 profiles are defined

3.2 Neural network design

NitroNet's neural network design is based on an extensive hyperparameter study (see Bergstra and Bengio (2012)), in which 300 different variants of the neural network (with different number of hidden layers, neurons per layer, training algorithm, etc.) were tested. The performance of a neural network can strongly depend on these parameters, but their ideal values can not be determined on prior knowledge. The different variants were ranked based on their *mean absolute percentage error* (MAPE) on

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

the validation set of WRF-2019. The MAPE is defined as

$$\mathrm{MAPE}(y_{\mathrm{pred}}, y_{\mathrm{true}}) = \frac{1}{n} \sum_{i=1}^{N} \left| \frac{y_{\mathrm{pred}}}{y_{\mathrm{true}}} - 1 \right| \tag{2}$$

where N is the number of instances in the validation set, y_{pred} the neural network prediction, and y_{true} the ground truth. The best neural network with regard to this metric was chosen for NitroNet and is described in the following.

The neural network has 8 hidden layers with 326 neurons each. It uses the Parametric Rectified Linear Unit activation function (PReLU, see He et al. (2015)), the Nesterov Adam optimizer (NAdam, see Ruder (2016)), a learning rate of $3.4 \cdot 10^{-4}$, a batch size of 2048, and the L_1 loss function, defined as

$$L_1(y_{\mathrm{pred}}, y_{\mathrm{true}}) = |y_{\mathrm{pred}} - y_{\mathrm{true}}| \tag{3}$$

The learning rate was halved whenever training progress had stalled over several epochs (i.e. full iterations over the training set). Detailed information about the hyperparameter optimization procedure can be found in Appendix A. NitroNet further deploys feature transformations (e.g. the quantile transformation from the sklearn library, see Pedregosa et al. (2012)) to reduce scale differences and skewedness of the input variables. Prediction uncertainties are computed via the Monte Carlo method, for which a comprehensive summary is found in Anderson (1976). Figure 1 shows an overview of the NitroNet model.

3.3 Training NitroNet on filtered data

The overall performance of NitroNet can be significantly enhanced by the implementation of a training data filtering scheme. The idea is to rank the NO_2 profiles from WRF-2019 by their agreement to reference data, and only use the best few percent for training. More specifically, we define two thresholds Δ_{VCD} and Δ_{PBLH} and remove all training instances where

$$\left| \frac{\mathrm{VCD}_{\mathrm{WRF}} - \mathrm{VCD}_{\mathrm{TROPOMI}}}{\mathrm{VCD}_{\mathrm{TROPOMI}}} \right| > \Delta_{\mathrm{VCD}} \quad \text{or} \quad \left| \frac{\mathrm{PBLH}_{\mathrm{WRF}} - \mathrm{PBLH}_{\mathrm{ERA5}}}{\mathrm{PBLH}_{\mathrm{ERA5}}} \right| > \Delta_{\mathrm{PBLH}} \tag{4}$$

Here, $\mathrm{VCD}_{\mathrm{WRF}}$ denotes the simulated NO_2 VCD from WRF-2019, $\mathrm{VCD}_{\mathrm{TROPOMI}}$ the observed NO_2 VCD from TROPOMI (using the simulated NO_2 a priori profiles), $\mathrm{PBLH}_{\mathrm{WRF}}$ the simulated PBLH from WRF-2019, and $\mathrm{PBLH}_{\mathrm{ERA5}}$ the PBLH from ERA5. This way, profiles with poor agreement to the TROPOMI NO_2 VCD (representing the total amount of NO_2) or the ERA5 PBLH (representing atmospheric mixing depth and profile shape) are identified and dismissed from training. The lower Δ_{VCD} and Δ_{PBLH} are chosen, the fewer instances remain in the training set. Therefore, we face a trade-off between training data quality and quantity, which we resolve by including Δ_{VCD} and Δ_{PBLH} in the hyperparameter optimization mentioned in sect. 3.2. By this means, ideal values of $\Delta_{\mathrm{VCD}} = 0.2$ and $\Delta_{\mathrm{PBLH}} = 0.1$ were determined. With these thresholds, only the best 7 % of all profiles (approximately 100.000) remain for training.

However, caution is warranted: If the training dataset is manipulated in such a way, it may become unrepresentative of the "real world" (e.g. by extinction of feature modes). Evaluation on the validation set shows, that the use of filtered training data introduces a low bias of approximately -10% to the NitroNet predictions in the lower layers of the atmosphere. This

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

bias can be determined immediately after training, stored in an altitude-dependent look-up table, and automatically subtracted from NitroNet's predictions. From a machine learning perspective, this look-up table is simply another hyperparameter, whose optimization is justified via validation on the independent test set.

3.4 Treatment of out-of-distribution instances

Neural networks are known to struggle when presented with out-of-distribution (OOD) instances, i.e. input data which lies outside the joint distribution of the training set. In the case of NitroNet (trained on one month of summertime RCT data in Europe), OOD instances are likely to occur in previously unseen geographical regions or seasons. The impact of OOD input variables on the neural network's performance can be detrimental, even if the neural network's sensitivity to the variable was low in the in-distribution case. In order to minimize the influence of OOD input variables, we implement a variant of the *winsorization* method (see e.g. Ruppert (2014)): First, the marginal probability density distributions $p_{x_i}(x)$ of the features x_i are estimated using kernel density estimation (KDE) on the training set. Instance entries are considered OOD, if they lay in regions of relatively low probability density, e.g. if $p_{x_i}(x) < 0.15$. In that case, they are replaced with a sample from p_{x_i}. The NO_2 VCD and categorical input features (i.e. surface classifications) are exempt from this treatment.

3.5 Correction of NO_z biases of in-situ measurements

An important part of the validation study presented in sect. 4 will be the comparison of NitroNet predictions to in-situ measurements at the surface. Over 90 % of the European in-situ measurements rely on the molybdenum-based chemiluminescence method, which is demonstrably cross-sensitive to other atmospheric oxidants (summarized as "NO_z"), such as peroxyacetyl nitrate (PAN), nitric acid (HNO_3) and the alkyl nitrates (see Dunlea et al. (2007); Steinbacher et al. (2007); Lamsal et al. (2008); Boersma et al. (2009); Villena et al. (2012)). Consequently, the reported NO_2 values are often too large, because a fraction of the NO_z is falsely registered as NO_2. Lamsal et al. (2008) give an empiric formula for the overestimation of the NO_2 concentration in the presence of NO_z:

$$F := \frac{[NO_2^*]}{[NO_2]} = 1 + \frac{0.95[\mathrm{PAN}] + 0.35[HNO_3] + \sum \text{alkyl nitrates}}{[NO_2]} \tag{5}$$

where [PAN], [HNO_3], and [NO_2] denote the true surface mixing ratios of PAN, HNO_3, and NO_2, while [NO_2^*] denotes the biased measurement result. The same formula was used in Kuhn et al. (2024), and was found to be crucial for the agreement between simulation data and in-situ measurements. NitroNet was trained to predict F (as learned from WRF-2019) as an additional output, so that when comparing NitroNet predictions to in-situ measurements, the measurement bias can be compensated. Because alkyl nitrates are not included in the MOZART chemical mechanism used in WRF-2019, we must assume $\sum \text{alkyl nitrates} = 0$. According to Elshorbany et al. (2012), the contribution of the alkyl nitrates to F can be estimated in the range of 2 % - 6 %. Based on the evaluation on the test set, NitroNet can reproduce the F-values from WRF-2019 with a relative precision of ± 5 %.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

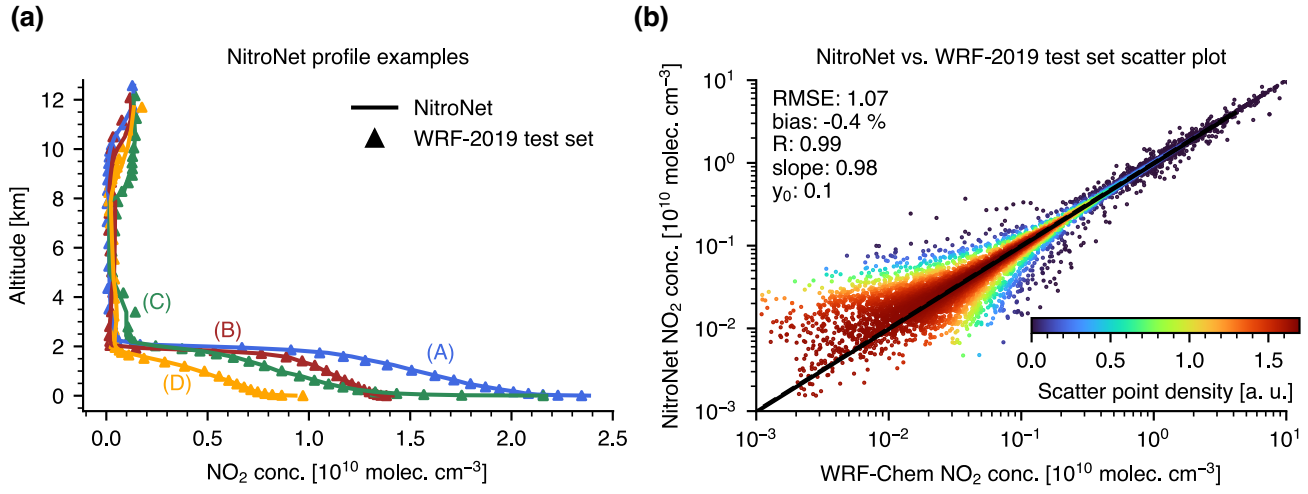

Figure 3. Evaluation of NitroNet on the WRF-2019 test set. **(a)** Four exemplary NO_2 profiles from the test set (triangular markers) with corresponding NitroNet predictions (solid lines). **(b)** Scatter plot of all NO_2 concentrations in the test set vs. their corresponding NitroNet predictions. RMSE and intercept are expressed in units of 10^9 molec. cm^{-3}.

4 Results

4.1 Evaluation of NitroNet in May 2019

From hereon, we deal with the validation of the trained NitroNet model. The easiest way to confirm successful training of the model is to validate it against new examples from the test set. Figure 3a shows four exemplary NO_2 profiles from the test set and the corresponding predictions from NitroNet. Our model reproduces the shape and magnitude of the profiles well, although there are small deviations, e.g. in profile (C) at $\sim$ 3 km altitude. Within the boundary layer, almost no discrepancies are observed. A noteworthy feature of the NO_2 profiles is their upper-tropospheric portion starting at 8 km altitude. Here, a sudden enhancement of the NO_2 concentration is found, which could be linked e.g. to aircraft emissions. Figure 3b shows a scatter plot of all NO_2 concentrations in the (filtered) test set against their corresponding NitroNet predictions. Here we refer to the same filter (by VCD and PBLH agreement), as described in sect. 3.3. The linear regression reveals excellent agreement, a strong correlation of $R = 0.99$, and a negligible bias of -0.4 %. Next, we verify that the training on filtered data as described in sect. 3.3 does indeed have the desired effect. For this purpose, we inter-compare observed and simulated NO_2 VCDs and surface concentrations from WRF-2019, NitroNet, TROPOMI, and AirBase. Figure 4a shows the comparison of monthly-mean NO_2 VCDs from TROPOMI and the corresponding simulation results from WRF-2019. The simulated VCDs are computed as

$$\mathrm{VCD}_{\mathrm{sim}} = \sum_{l<l_{\mathrm{tp}}} c_l \cdot \Delta h_l \tag{6}$$

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

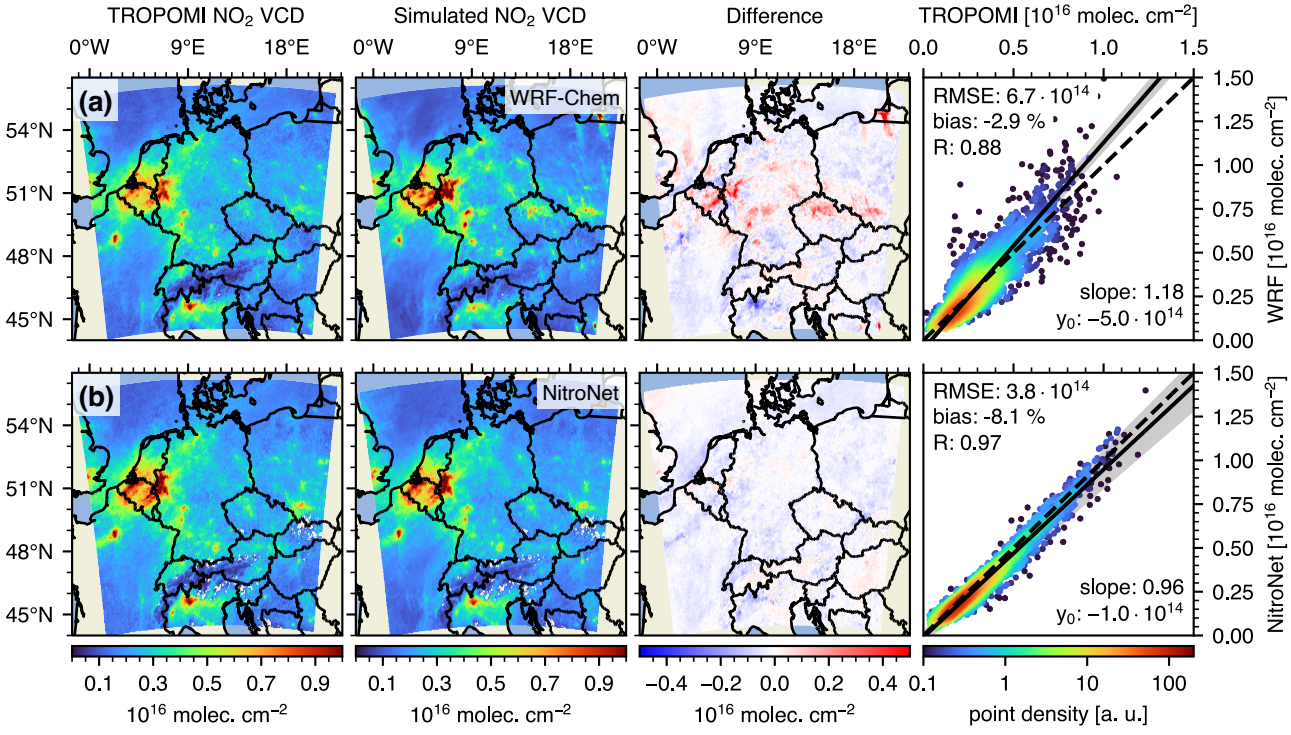

Figure 4. Comparison of monthly-mean TROPOMI NO_2 VCDs against simulated NO_2 VCDs from WRF-Chem **(a)** and NitroNet **(b)** (May 2019). The NO_2 a priori profiles used in the air mass factor computation of the TROPOMI VCDs were replaced with those from WRF-Chem and NitroNet, respectively. RMSE and intercept are given in units of 10^{14} molec. cm^{-2}.

where l denotes the layer index, l_{tp} the tropopause layer index, c_l the NO_2 concentration in layer l, and Δh_l the vertical extent of layer l. The NO_2 a priori profiles used in the air mass factor computation of the TROPOMI VCDs were replaced with those from WRF-Chem, following Eskes et al. (2019):

$$\mathrm{VCD}_{\mathrm{obs,\ corr}} = \mathrm{VCD}_{\mathrm{obs}} \cdot \frac{\mathrm{AMF}_{\mathrm{trop}}}{\mathrm{AMF}} \cdot \frac{\sum_{l<l_{\mathrm{tp}}} c_l \cdot \Delta h_l}{\sum_{l<l_{\mathrm{tp}}} c_l \cdot \Delta h_l \cdot A_l} \tag{7}$$

where $\mathrm{VCD}_{\mathrm{obs,\ corr}}$ denotes the VCD with exchanged a priori profile, $\mathrm{VCD}_{\mathrm{obs}}$ the original VCD, AMF the total air mass factor, $\mathrm{AMF}_{\mathrm{trop}}$ the tropospheric air mass factor, and A_l the tropospheric averaging kernel of layer l. Figure 4a reveals significant biases in the WRF-Chem simulation of up to 10^{16} molec. cm^{-2} (e.g. in western Germany, northern Austria, and the Kaliningrad Oblast). The simulated and observed NO_2 VCDs agree with a mean bias of -2.9 %, an RMSE of $6.7 \cdot 10^{14}$ molec. cm^{-2}, and a correlation coefficient of $R = 0.88$. Here, and throughout the rest of the article, "correlation coefficient" refers to the Pearson correlation coefficient. A more detailed discussion of the WRF-Chem simulation results can be found in Kuhn et al. (2024).

Fig. 4b shows the same comparison, but using the NO_2 profiles from NitroNet instead of WRF-Chem. Overall, much better agreement is observed. In particular, the major overestimations observed with WRF-Chem have disappeared, while some

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

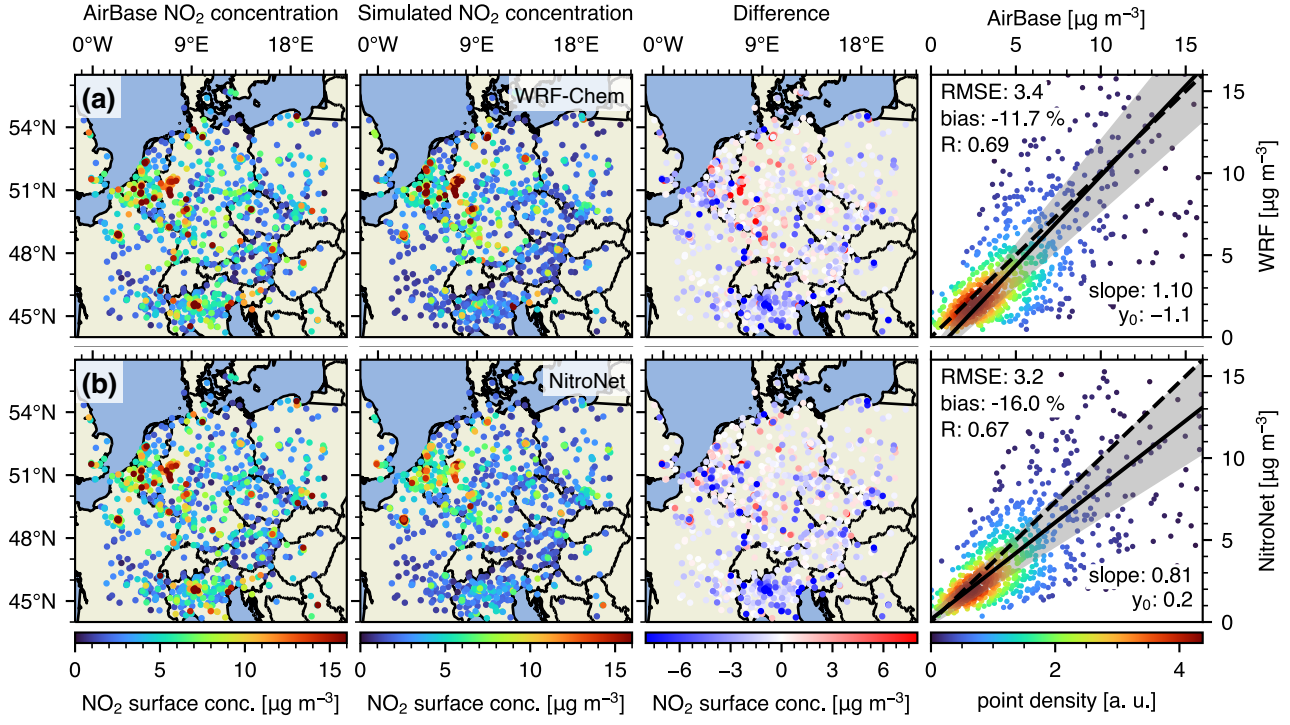

Figure 5. Comparison of monthly-mean AirBase NO_2 surface observations against simulated surface concentrations from WRF-Chem **(a)** and NitroNet **(b)** at TROPOMI overpass time (May 2019). The AirBase observations were corrected for NO_z biases, using WRF-Chem model results for **(a)** and NitroNet predictions for **(b)**, respectively. RMSE and intercept are given in units of $\mu g\ m^{-3}$.

weak underestimations remain. Although the absolute mean bias is slightly larger (-8.1 %), the correlation is much stronger ($R = 0.97$) and the RMSE was almost halved ($3.8 \cdot 10^{14}$ molec. cm^{-2}). In some regions of the domain (e.g. near the cities of Frankfurt and Mannheim, Germany), these improvements are easily explained by the considerable reduction of the simulated column. In other regions (e.g. at the border between Belgium, Netherlands, and Germany), the improvements must be partially attributed to larger TROPOMI reference VCDs, resulting from the use of presumably more realistic a priori NO_2 profiles.

Figure 5 shows the comparison of monthly-mean NO_2 surface concentrations from AirBase to the corresponding model results at TROPOMI overpass time. The NO_z bias correction described in sect. 3.5 was applied to the AirBase data, using WRF-2019 and NitroNet model results, respectively. The "Difference" subplots of Fig. 4 and Fig. 5 show a clear correlation, e.g. in western Germany and northern Italy. Nonetheless, different spatial patterns can be identified between NitroNet and WRF-2019: In some model regions (e.g. in western Germany) NitroNet produced smaller errors than WRF-Chem with respect to the VCDs and the surface concentrations. However, the opposite is observed in other regions. For example, NitroNet produced smaller VCD errors, but larger surface concentration errors in northern Italy. This demonstrates that filtering of the training data based on VCD and PBLH criteria alone may not always lead to better neural network predictions at the surface. Scatter

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

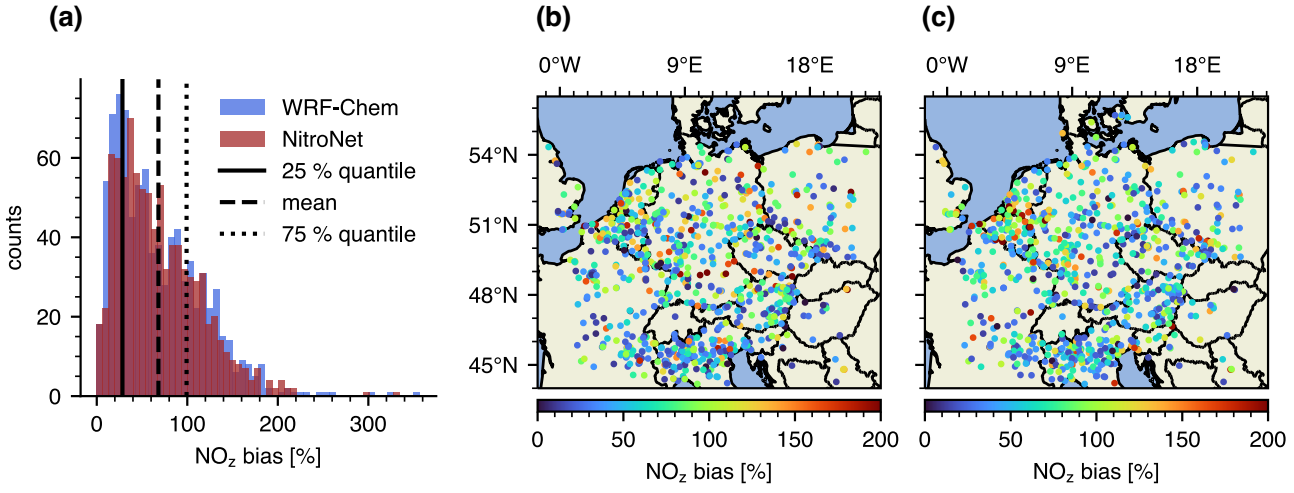

Figure 6. Histogram **(a)** and geographic distribution of the monthly-mean NO_z biases from **(b)** WRF-Chem and **(c)** NitroNet corresponding to the AirBase observations shown in Fig. 5.

plots for individual countries (Germany, Netherlands, and Italy) with differing response to the data filtering (improvement, neutral, worsening) can be found in Fig. C1. This finding is important for the interpretation of the presented results: WRF-Chem produces positive and negative errors in moderate balance, while NitroNet produces similar negative, but much smaller positive errors. Subsequently, NitroNet shows a smaller RMSE (3.2 μg m^{-3} vs 3.4 μg m^{-3}), but larger absolute mean bias (-16.0 % vs. -11.7 %). In such a case, the increase in absolute mean bias is obviously not a suitable measure for overall model skill. The slight reduction in correlation coefficient ($R = 0.67$ vs. $R = 0.69$) escapes this argument, but can be considered insignificant.

Figure 6 shows a histogram of the NO_z biases of the in-situ measurements, computed from modelled PAN and HNO_3 mixing ratios according to Lamsal et al. (2008), see sect. 3.5. The results obtained from WRF-Chem and NitroNet show values of up to $+200$ %. We show this figure with the intent of emphasizing that caution is required when using in-situ measurements for training and validation of RCT and machine learning models without a proper correction strategy.

The results of this section demonstrate that our training method has the intended effect: Using filtered data, NitroNet produces NO_2 profiles of overall more realistic magnitude and/or shape than WRF-Chem. Although the improvement to the simulated surface concentrations is rather small, a much stronger improvement to the VCDs is obtained. Even better results are expected by further filtering the training data by their agreement to the in-situ observations. However, this is impossible here, as the surface observations are so sparse that too few data would remain for the training of the neural network.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

4.2 Evaluation of NitroNet on unseen data (May 2022)

We now address the validation of NitroNet on completely new input data from the month of May 2022. From hereon we use NitroNet without comparison to RCT simulation data on a domain ranging from 42° to 56° latitude, and from -5° to 23° longitude.

4.2.1 Validation against TROPOMI satellite data and AirBase in-situ measurements

Figure 7 shows the comparison of monthly-mean NO_2 VCDs from TROPOMI against NitroNet predictions. The computations were conducted as explained in sect. 4.1. The NitroNet NO_2 VCDs show similar magnitudes, geographical distribution, and errors as in May 2019. However, the results for May 2022 show lower RMSE ($2.8{\cdot}10^{14}$ molec. cm^{-2} vs. $3.8{\cdot}10^{14}$ molec. cm^{-2}), and increased mean bias (+6.7 % vs. -8.1 %). This apparent improvement could be purely coincidental: Figure 4b indicates a slight underestimation of the NO_2 VCDs on behalf of NitroNet. On the other hand, the NO_2 VCDs in May 2019 were on average 18 % higher than in May 2022; Subsequently, NitroNet can overestimate the true VCDs because it attempts to reproduce the approximate magnitudes learned from 2019. If the two effects cancel each other out, this could reasonably explain the smaller VCD errors observed in 2022.

Figure 7b shows the comparison of monthly-mean NO_2 surface concentrations from AirBase against NitroNet predictions. NitroNet correctly identifies surface pollution hotspots (e.g. in Paris (France), Essen (Germany), and Hamburg (Germany)), but somewhat underestimates surface NO_2 concentrations in various regions of the domain. Compared to May 2019, the results show a smaller mean bias (-10.5 % vs. -16.0 %), a higher correlation coefficient ($R = 0.72$ vs. $R = 0.67$), and significantly reduced RMSE (1.2 μg m^{-3} vs 3.2 μg m^{-3}). A key contribution to these differences is found in the Lombardy region of northern Italy. Here, significant underestimations were observed in 2019, but the corresponding data points are missing entirely in 2022. Inspection of the AirBase metadata reveals, that in May 2019 over 92 % of the Italian measurements were flagged as "valid", 5 % as "invalid", and 2 % as "below detection limit". In May 2022, however, only 48 % of the measurements were flagged as "valid", 13 % as "invalid", and 39 % as "below detection limit". Additionally, the total number of Italian instruments was reduced from 320 in 2019 to just 69 in 2022. It remains unclear, why these measurements were removed from AirBase.

Another interesting observation is the dependence of NitroNet's low bias on the measuring stations' type. Here we refer to the entire domain shown in Fig. 7. As explained in sect. 2.2, we exclusively use background stations throughout our study, based on the argument that accurate modelling of traffic and industrial scenarios is known to require simulations of much higher resolution ("local scale"). So far we have assumed no errors in the classification of the AirBase instruments. However, based on the resolutions of modern emission inventories, the variability of trace gas transport, and the scarce documentation on classification criteria, it can be argued that the category "urban background" is a grey zone within this classification. After all, emission inventories clearly show that urban regions are always affected by traffic emissions. We therefore investigated, whether the comparison of NitroNet's results to in situ observations would improve by removing the urban background stations, as shown in Fig. 7c. Significant improvements were revealed, manifesting in increased slopes (from 0.84 to 1.00), lower absolute mean bias (-10.5 % to +2.2 %), and lower RMSE (1.7 μg m^{-3} to 1.2 μg m^{-3}). These improvements can be explained

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

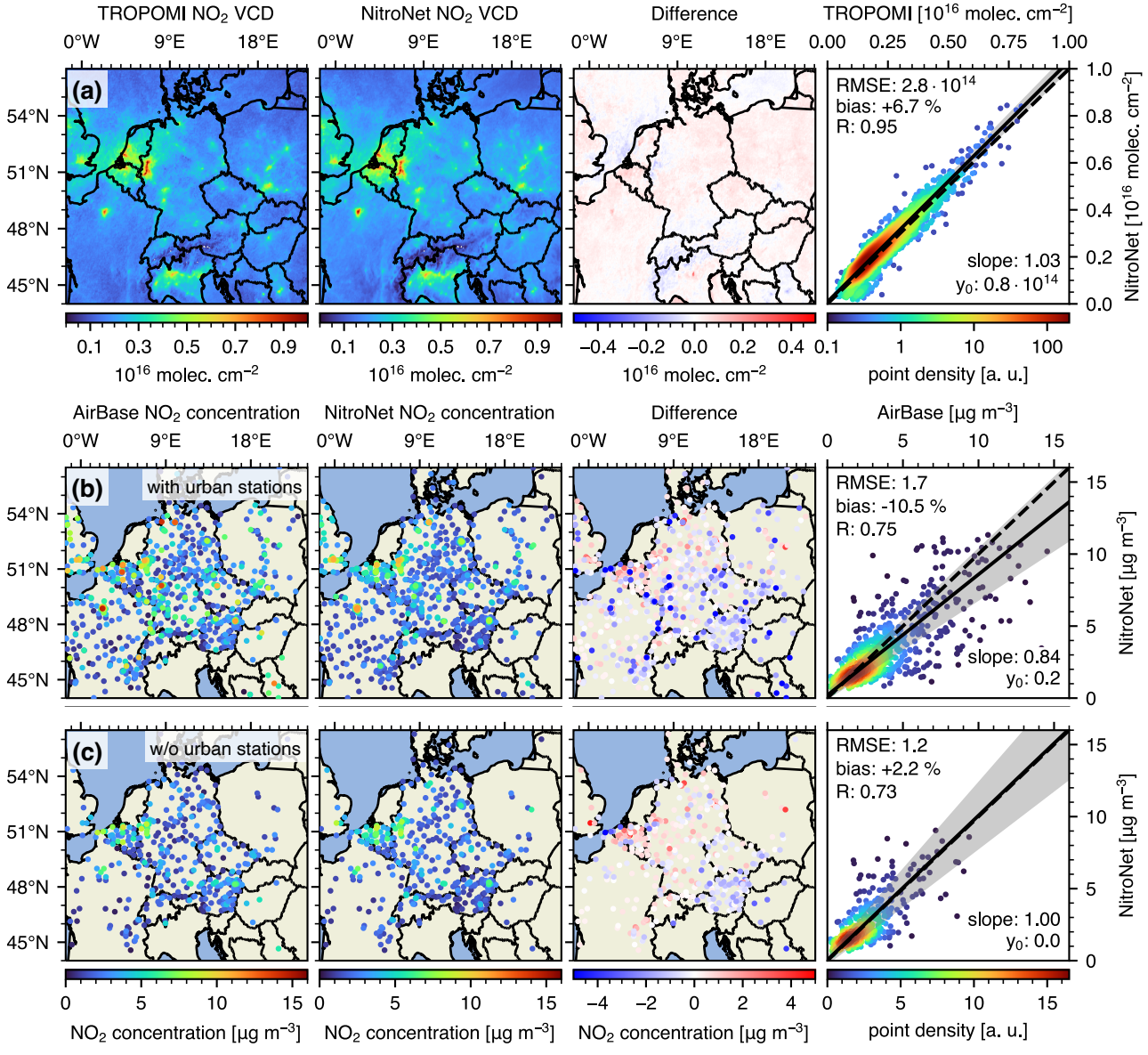

Figure 7. Comparison of monthly-mean TROPOMI NO_2 VCDs **(a)** and AirBase surface observations **(b)** against NitroNet predictions (May 2022). Subfigure **(c)** is identical to **(b)**, except that AirBase instruments of the type "urban background" were removed. RMSE and intercept are displayed in molec. cm^{-2} for the VCDs, and $\mu g\ m^{-3}$ for the surface concentrations.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

either by a tendency of NitroNet to underestimate NO_2 concentrations in urban areas, or by an erroneous categorization of the measurements. Due to the lack of information about the classification process, we will omit the urban background stations in our evaluations from hereon.

4.2.2 Validation against FRM$_4$DOAS MAX-DOAS measurements

We now validate the NO_2 profiles from NitroNet against MAX-DOAS measurements from the FRM$_4$DOAS dataset at six European locations. Averaging kernels are available from the MMF retrieval algorithm and given as an $n \times n$ matrix $\mathbf{A}$, where n denotes the number of vertical layers in the retrieval. The i-th row of $\mathbf{A}$ describes the retrieval sensitivity of the concentration value of layer i to the other n layers. An ideal retrieval would be characterized by $\mathbf{A} = \mathbb{1}$, where $\mathbb{1}$ denotes the unity matrix. In practice, the AK matrix diagonal is usually close to unity at the surface, but quickly drops below 50 % within the first 1-2 km above ground (see e.g. Fig. C3, showing the AK matrix of the instrument in Heidelberg, Germany). The AKs are applied to the NitroNet profiles following Rodgers (2000) by computing

$$\boldsymbol{c}_{\mathrm{sim,\ corr}} = \mathbf{A}\boldsymbol{c}_{\mathrm{sim}} + (\mathbb{1} - \mathbf{A})\boldsymbol{c}_{\mathrm{ap}} \tag{8}$$

where $\boldsymbol{c}_{\mathrm{sim}}$ denotes the original NitroNet profile and $\boldsymbol{c}_{\mathrm{ap}}$ the assumed a priori profile. The AKs are applied as described when comparing NitroNet to MMF profiles. MAPA, on the other hand, does not supply AKs.

Figure 8 shows the results obtained with this procedure. The thin scatter points (legend handles "MAPA" and "MMF") represent a one-to-one comparison of NO_2 concentration values from NitroNet and MAPA or MMF. The thick scatter points (legend handles "MAPA (monthly)" and "MMF (monthly)") show the monthly-mean NO_2 concentrations of each retrieval layer. The level of agreement between FRM$_4$DOAS and NitroNet varies, depending on the instrument location. NitroNet and MAPA show significant differences in some locations, with biases ranging from -3.6 % (San Pietro Capofiume) to $+99.6$ % (Heidelberg), RMSE values on the scale of $6 \cdot 10^9$ molec. cm^{-3} and correlation coefficients ranging from $R = 0.86$ (San Pietro Capofiume) to $R = 0.95$ (De Bilt). NitroNet and MMF show overall better agreement, with biases ranging from -34.3 % (San Pietro Capofiume) to $+8.7$ % (Bremen), RMSE values on the scale of $4 \cdot 10^9$ molec. cm^{-3} and correlation coefficients larger than 0.90. The linear regressions show significantly steeper slopes for MMF than for MAPA, but similar intercepts. MAPA tends to produce higher NO_2 concentrations than MMF in the lowest few hundred meters above ground, but smaller concentrations above. The NitroNet predictions are somewhere in between, manifesting in an "S"-shaped distribution of the scatter markers (see e.g. the comparison to MMF in Heidelberg). The corresponding plots of monthly-mean NO_2 profiles can be found in Fig. 9. NitroNet shows good agreement with the co-located surface observations (except for the station "BETR012" in Uccle). This is made possible by NitroNet's high vertical resolution at the surface (~ 1 m), which is adequate for the steep prevailing concentration gradients. This is not the case for MAPA and MMF, because the vertical sampling of FRM$_4$DOAS (~ 200 m) is too coarse. Our observations in this regard align well with the findings of Bösch (2018), who presents a a detailed comparison of MAX-DOAS measurements and co-located surface observations. The differences between MAPA, MMF, and NitroNet can partly be linked to the models' implementations and limitations: MMF uses a single, fixed NO_2 a priori profile for all retrievals, which was obtained from a WRF-Chem simulation in Mexico (Friedrich et al. (2019)). However, datasets like our

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

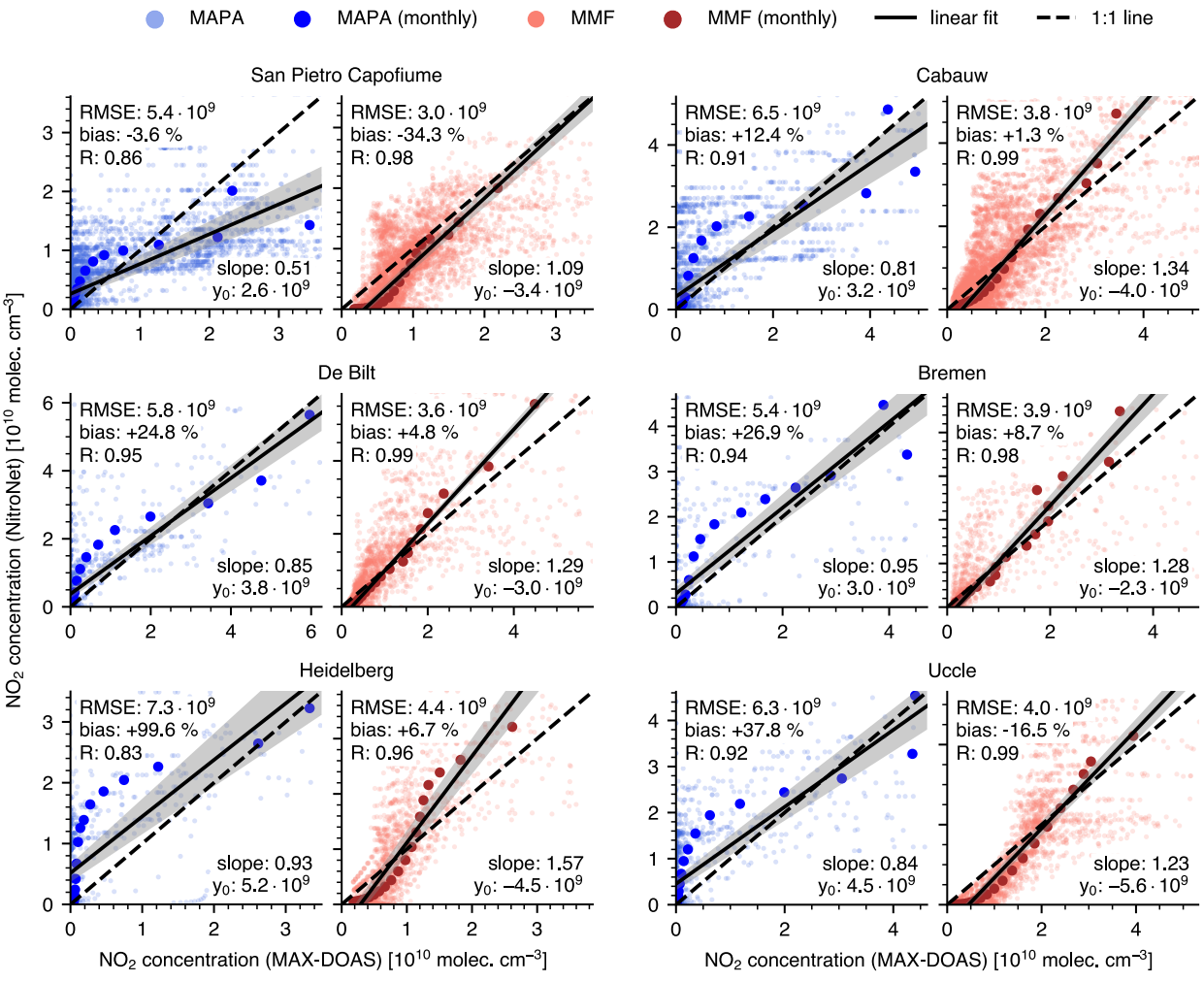

Figure 8. Comparison of FRM$_4$DOAS NO_2 concentrations against NitroNet predictions (May 2022). MAPA results are drawn in blue, and MMF results in red. The thin scatter points represent a one-to-one comparison of NO_2 concentration values (i.e. the concentrations of individual profiles). The thick scatter points show the monthly-mean NO_2 concentrations of each retrieval layer. RMSE and intercept are displayed in molec. cm^{-3}.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

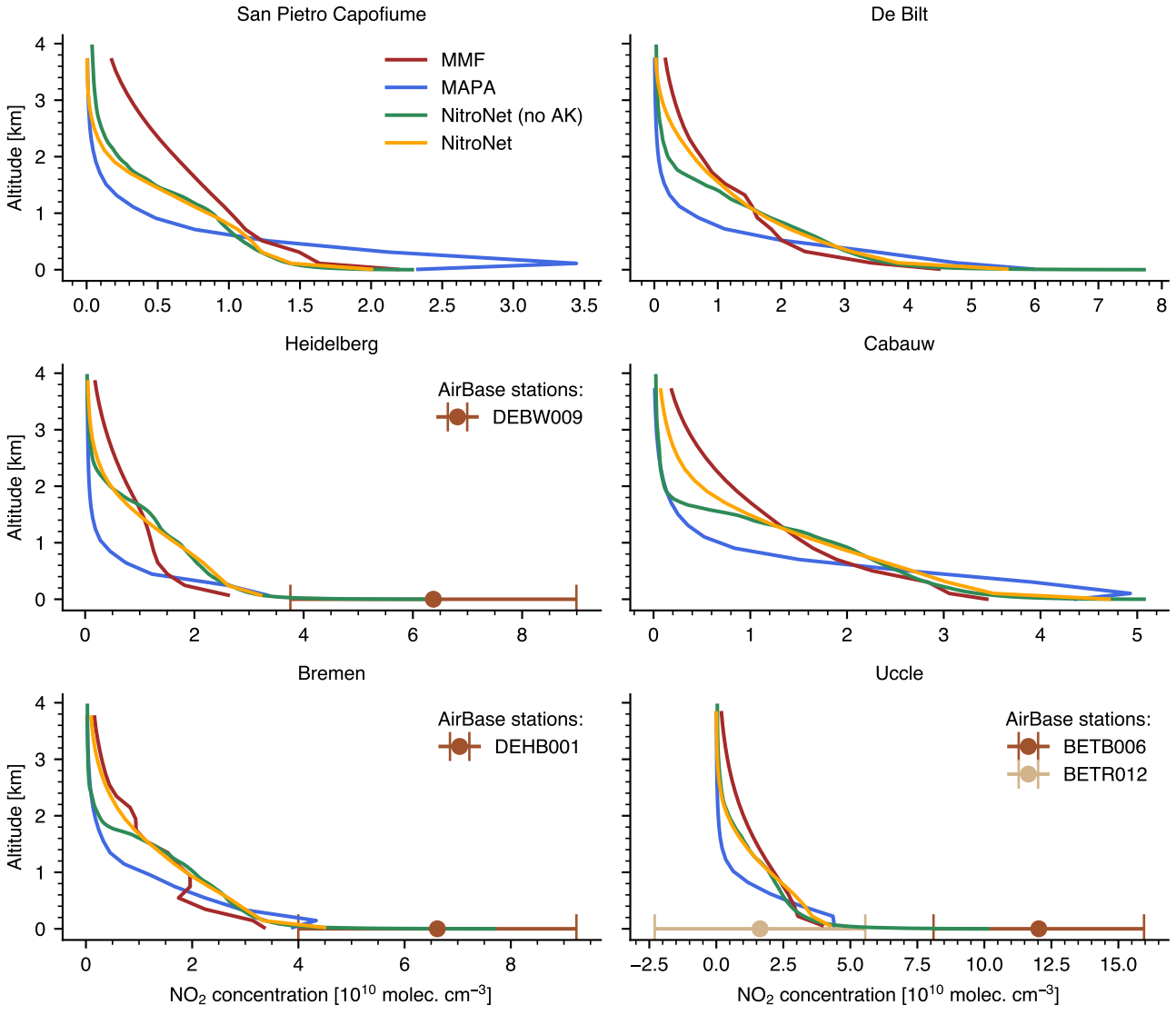

Figure 9. Comparison of monthly-mean FRM$_4$DOAS NO_2 profiles against NitroNet profiles (May 2022). Where available, co-located AirBase measurements of the surface NO_2 concentration within a radius of 5 km were drawn at 0 m altitude.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

own WRF-2019 show strong horizontal variability of NO_2 profiles on the scale of just a few kilometers. A single a priori profile is therefore not sufficient to fully represent the diversity of profile shapes and magnitudes. Moreover, horizontal gradients also systematically affect the MAX-DOAS profile retrievals. Subsequently, it is not surprising to see larger differences between MMF, MAPA and NitroNet (without AKs) in the regions of reduced sensitivity (small AKs) above 1 km altitude. Application of the AKs reduces the differences significantly in 3/6 locations (De Bilt, Cabauw, Bremen). MAPA, on the other hand, makes a priori assumptions in the form of a pre-defined profile parametrization. The profiles shown in Fig. 9 are qualitatively similar to those from MAPA's original publication paper (Beirle et al. (2019)), with a strong exponential shape and an optional peak in the 2nd or 3rd layer (San Pietro Capofiume and Cabauw). This could indicate the presence of an elevated NO_2 layer. NitroNet is unable to reproduce this profile type, most likely because the training dataset contains very few corresponding examples. On the other hand, the elevated layers are not reproduced by MMF either. In that regard, it is possible that they are falsefully produced by an incompatibility between the true NO_2 profile and MAPA's profile parametrization (technically a form of "model misspecification error"). Overall, the differences between MAPA and MMF demonstrate the large uncertainty from the choice of retrieval algorithm alone. Further sources of uncertainty (e.g. the influence of horizontal gradients), as well as the low statistical relevance of only six measurement locations must be considered withal, and are not easily quantifiable. Within these limitations, the comparison to MAX-DOAS data shows no glaring discrepancies, as much as it allows for no more than an approximate validation of profile shapes and magnitudes.

4.3 Evaluation of NitroNet in other seasons and regions of the world

Lastly, we present an analysis of NitroNet's ability to generalize to other seasons and regions of the world. The evaluations shown in sect. 4.2.1 were made for the same region (central Europe) and time of the year (May), on which the neural network was trained. Hence they represent the least challenging test case. Good generalization to other domains and seasons is not guaranteed, and associated with two challenges: Firstly, the neural network must respond reasonably to fundamentally different input data (e.g. much lower temperatures in winter than in summer). This is controlled by the network's regularization, which we enforce mainly via the winsorization technique described in sect. 3.4. Secondly, the training data is expected to be "epistemically incomplete", meaning that it does not contain all relevant training examples for other seasons and regions. This is a property of the training set, which we regard as a principal limitation that can not be resolved in the scope of this article. Nonetheless it is not implausible, that the fundamental relationships between the input and output data, as learned by NitroNet, hold at least partly for other seasons and regions, as well.

We first investigate the regional generalization capability of the model using reference data from May 2022. Figure 10 shows the comparison to TROPOMI NO_2 VCDs over the United Kingdom (UK, Fig. 10a) and the Mediterranean region of Portugal and Spain (Fig. 10b). The results are overall very similar to those from the central European domain investigated previously. However, Fig. 10b shows significant overestimations of approximately 10^{15} molec. cm^{-2} over the southern waterbodies (the Alboran Sea and the Gulf of Cadiz). It is not generally unexpected to see such systematic errors in the predictions of a neural network. The most likely explanation for this is that the training dataset possibly does not contain (enough) representative cases. For example, the water regions of the training set's spatial domain are pervaded by shipping routes, which may lead NitroNet

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

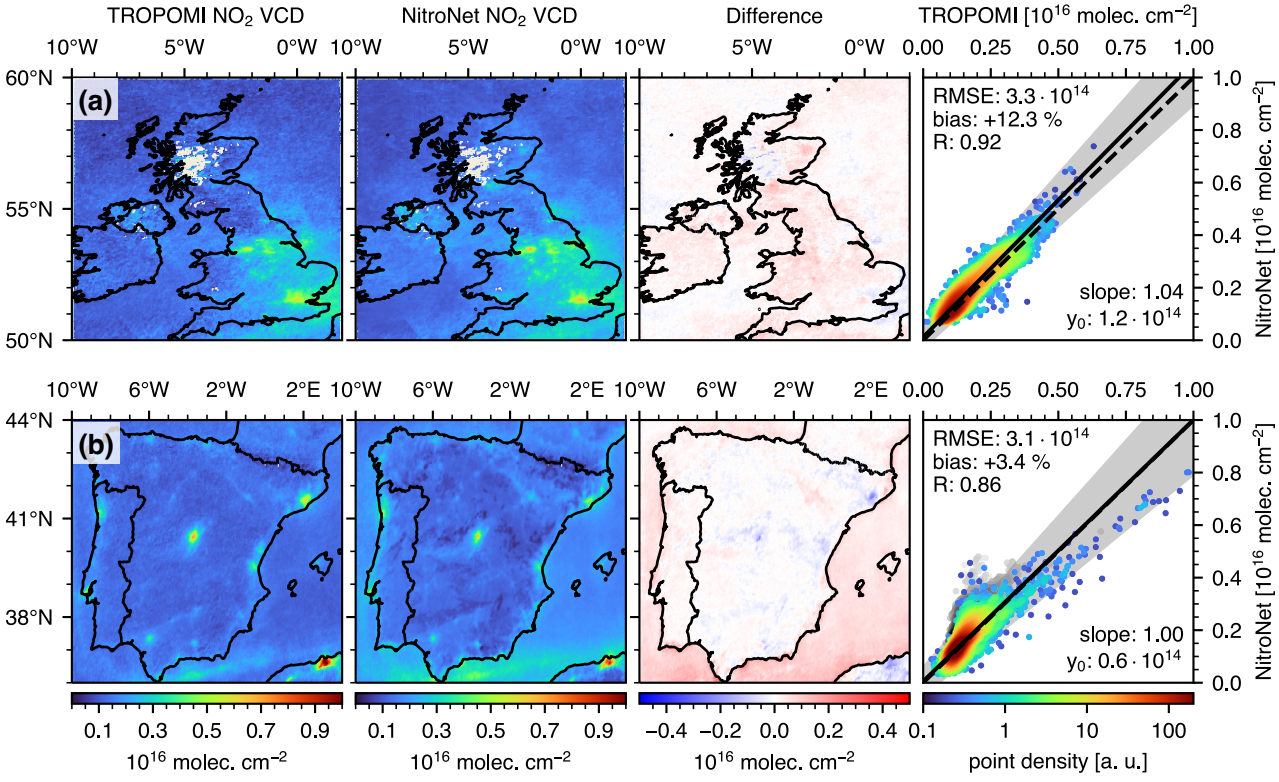

Figure 10. Like Fig. 7a, but for **(a)** the UK, and **(b)** the countries of Spain and Portugal. Water-pixels are drawn as gray dots in the right-side scatter plots and excluded from the statistical analysis. RMSE and intercept are displayed in molec. cm^{-2}.

to overestimate NO_2 over more remote water bodies. We exclude these pixels from the statistical analysis, because they skew the results, while being much less relevant than land pixels. Compared to the central European domain, the RMSE values are increased from $2.8 \cdot 10^{14}$ molec. cm^{-2} to $3.3 \cdot 10^{14}$ molec. cm^{-2} (UK) and $3.1 \cdot 10^{14}$ molec. cm^{-2} (Spain and Portugal), while the correlation coefficients are reduced from $R = 0.95$ to $R = 0.92$ (UK) and $R = 0.86$ (Spain and Portugal). The mean biases are +12.3 % (UK) and +3.4 % (Spain and Portugal), respectively. For context, an RMSE of $5.0 \cdot 10^{14}$ molec. cm^{-2}, a bias of +18.0 %, and a correlation coefficient of $R = 0.74$ is obtained for the domain of Spain and Portugal if water pixels are included. The statistical analysis of the UK domain, however, is practically unaffected by water pixels. Figure 11 shows the corresponding comparison to AirBase surface observations in analogy to Fig. 7, including the omission of "urban background" stations. A version of Fig. 11 including urban stations is found in Fig. C2. The results are similar: On the UK domain, the RMSE is slightly increased from 1.2 μg m^{-3} to 1.8 μg m^{-3}, and the correlation coefficient is reduced significantly from $R = 0.73$ to $R = 0.45$. This is caused by the two outliers in the south-eastern corner of the domain and amplified by the low number of total observations. On the Mediterranean domain the number of observations is much larger, and the results are overall better, with an RMSE of 1.6 μg m^{-3} and a correlation coefficient of $R = 0.71$. This demonstrates, that NitroNet can generalize to

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

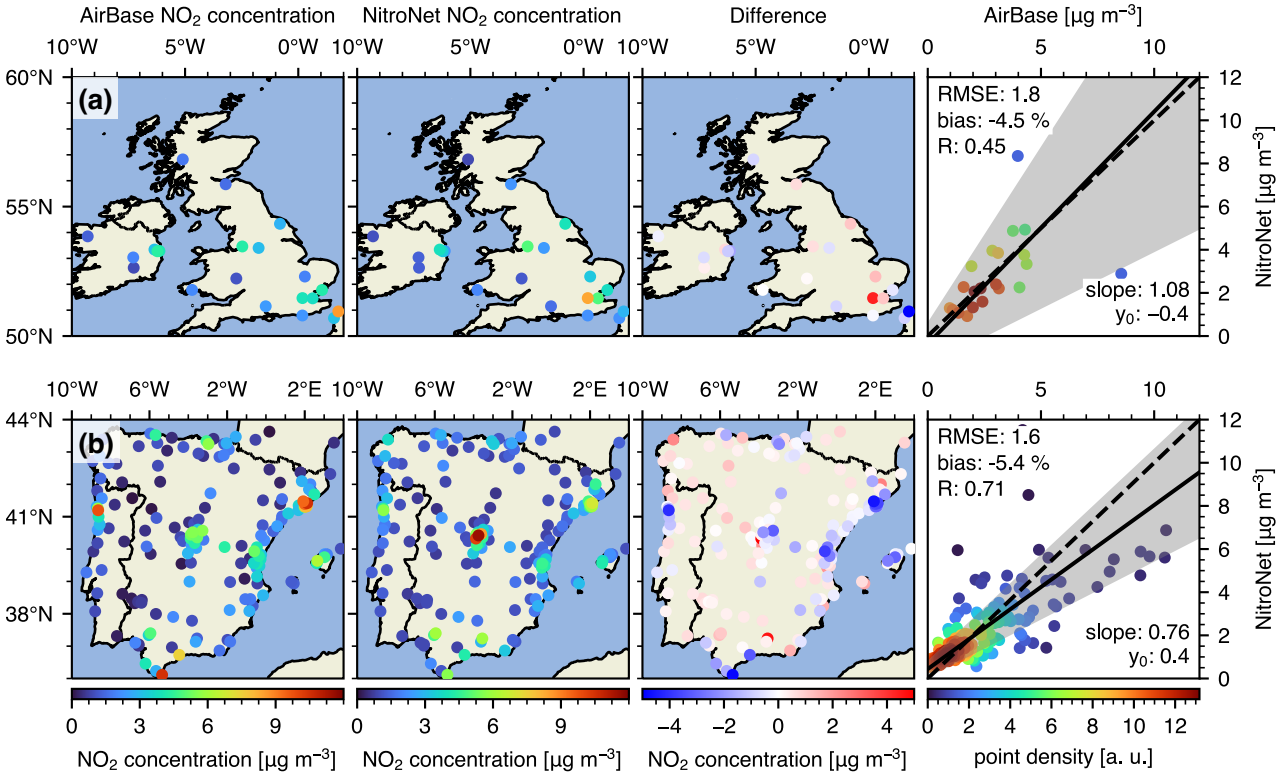

Figure 11. Like Fig. 7c, but for **(a)** the UK, and **(b)** the countries of Spain and Portugal. RMSE and intercept are displayed in μg m^{-3}.

new, but qualitatively similar domains with minor loss of prediction accuracy. NitroNet was also tested on three more "distant" domains covering the United States (US) west coast, India, and western China (see Fig. 12). We obtain good agreement for the US west coast (RMSE = $2.7 \cdot 10^{14}$ molec. cm^{-2}, bias = $+2.7$ %, R = 0.84). The Indian domain shows stronger correlation, but lower accuracy due to significant overestimations (RMSE = $8.0 \cdot 10^{14}$ molec. cm^{-2}, bias = $+41.5$ %, $R = 0.91$). The biggest deviations and weakest correlations are observed on the Chinese domain (RMSE = $12.6 \cdot 10^{14}$ molec. cm^{-2}, bias = $+12.5$ %, $R = 0.70$). Here, as shown in Fig. 12c, NitroNet ignores entire pollution hotspot areas in the northern Shanxi and Shaanxi provinces. These regions are known for their strong emissions from coal, steel, chemical and military industry (see e.g. Peng et al. (2023)). China's rapid economic development combined with fewer environmental state regulations make it plausible, that the EDGARv5 emission data of the year 2015 might already be outdated in such locations. Besides, NitroNet may struggle with the differences in atmospheric composition, e.g. the vastly higher aerosol pollution which prevails in China (see e.g. Meng et al. (2022)). The previously mentioned overestimation over waterbodies is observed in all three domains.

Finally, we investigate the seasonal performance of NitroNet. For this purpose, a whole year of data (August 2021 - July 2022) was processed on the central European domain. The NitroNet predictions are evaluated against TROPOMI and AirBase observations. The resulting time series are displayed in Fig. 13. The figure shows both monthly mean values in analogy to the

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

other evaluations presented up to this point, as well as daily mean values. In the following, we will focus on the monthly means. NitroNet's performance shows a clear seasonal cycle: The mean biases increase during wintertime and reach maximal values of -22.4 % (vs. TROPOMI, January) and -50.1 % (vs. AirBase, December). Likewise, the RMSE increases during wintertime and reaches maximal values of $10.8 \cdot 10^{14}$ molec. cm^{-2} (vs. TROPOMI, January) and 6.3 μg m^{-3} (vs. AirBase, December). The correlation coefficients are on the scale of $R \approx 0.90$ vs. TROPOMI and $R \approx 0.70$ vs. AirBase, with no conclusive annual cycle. The decrease in model performance in winter is expected due to the reasons discussed earlier. In particular, the oxidative capacity (via the hydroxyl and peroxy radicals) is reduced in winter and results in increased $\mathrm{NO_2}$ lifetimes of more than 20 hours, as opposed to 2 - 6 hours in summer (see e.g. Liu et al. (2016); Shah et al. (2020)). The results show that without specifically training on wintertime data, NitroNet's prediction for deep winter are only of limited value. Besides the obvious challenge of achieving good generalization from summertime training data to wintertime predictions, higher uncertainties in the input satellite data should also be taken into account in this context (see e.g. Douros et al. (2023)). Nonetheless, compared to the typical performance of RCT simulations, NitroNet performs well for the majority of the analyzed time series. Compared to WRF-2019, with equivalent filter criteria, the RMSE values of NitroNet's $\mathrm{NO_2}$ VCDs and surface concentrations are lower in 9 out of 12 months. Likewise, the correlation coefficients are larger in 10 out of 12 months (see the dotted gray lines in Fig. 13). It should be noted, that the performance of RCT simulations is expected to also drop significantly in wintertime. The scientific literature on the topic is sparse, but a study by Douros et al. (2023) shows that CAMS (an ensemble model consisting of 11 RCT models) produces summertime VCD biases of ~ 15 % and wintertime VCD biases of ~ 50 % in Europe. In light of such results, NitroNet's seasonal performance on the European domain can be considered competitive to most recent RCT simulations. Figure C4 shows examples of the comparison between NitroNet and TROPOMI for two individual days in summer and winter. In contrast to the monthly-mean comparisons shown previously, the data contains a significant amount of gaps (e.g. due to clouds), the correlation is reduced ($R \approx 0.80$), and the prediction errors are larger. This is expected, since averaging over an entire month of data reduces the statistical noise of the model. Nonetheless, as reflected in Fig. 13, NitroNet's daily performance is still competitive to that of WRF-Chem, indicating that it can reasonably be used for unaveraged predictions.

5 Conclusions, discussion, and outlook

In this article we have introduced NitroNet, a new deep-learning $\mathrm{NO_2}$ profile retrieval prototype for the TROPOMI satellite instrument. NitroNet is trained on one month of RCT simulation data from the WRF-Chem model in central Europe, May 2019. The use of synthetic data allows to overcome several obstacles associated with the empirical datasets used in other studies. The main benefits of our approach can be summarized as follows:

1. Because measurements of $\mathrm{NO_2}$ profiles are still sparse, empirical training data are effectively restricted to surface in-situ observations. A synthetic training dataset allows the neural network to learn the prediction of full $\mathrm{NO_2}$ profiles instead. These training profiles also cover the spatial domain continuously, and might cover scenarios which escape the in-situ observations altogether due to the strategic placement of the instruments.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

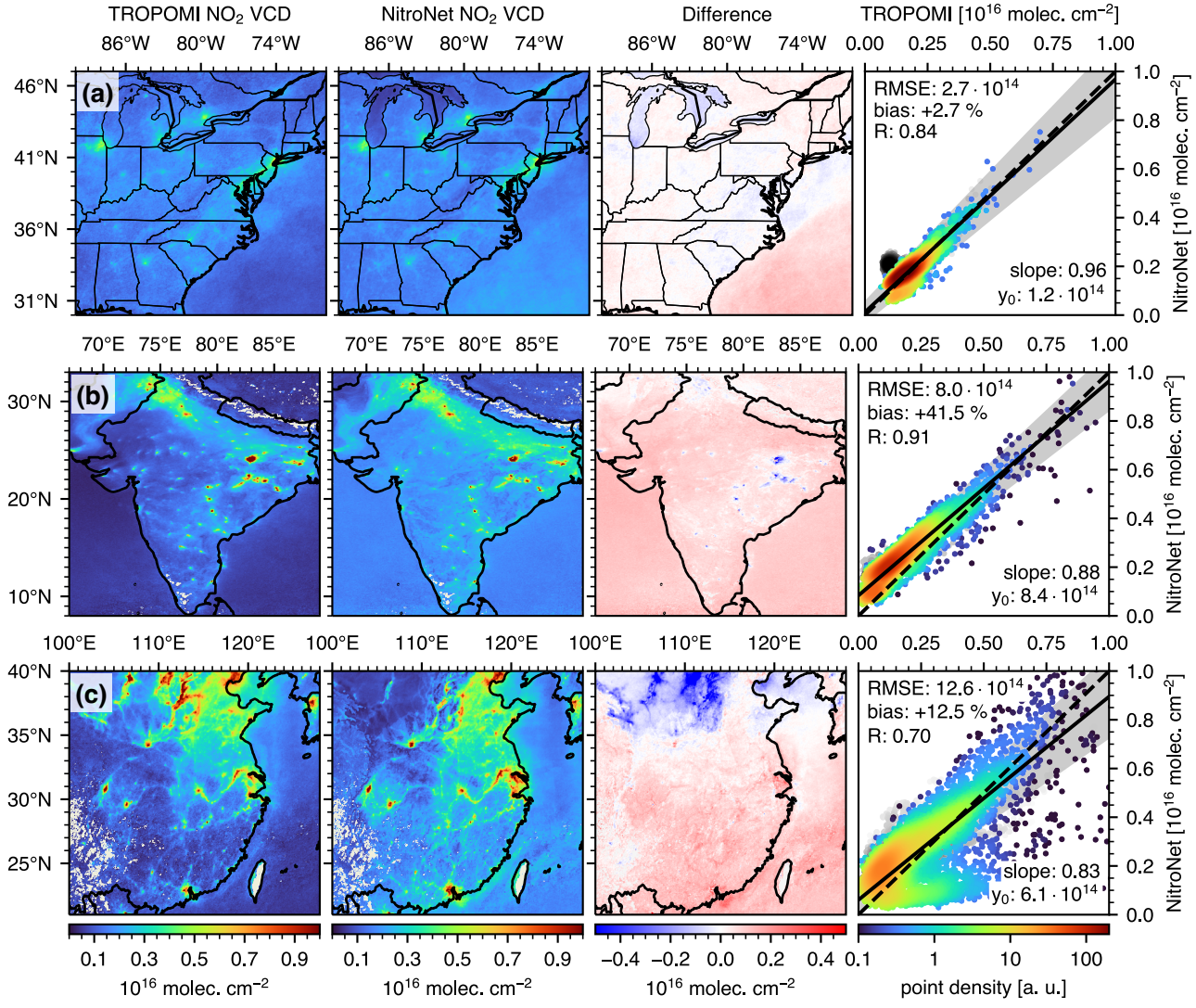

Figure 12. Like Fig. 10, extended to **(a)** the US west coast, **(b)** India, and **(c)** western China. The uncoloured scatter markers in panel **(a)** symbolize the entries over water, which were dismissed from the statistical analysis.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

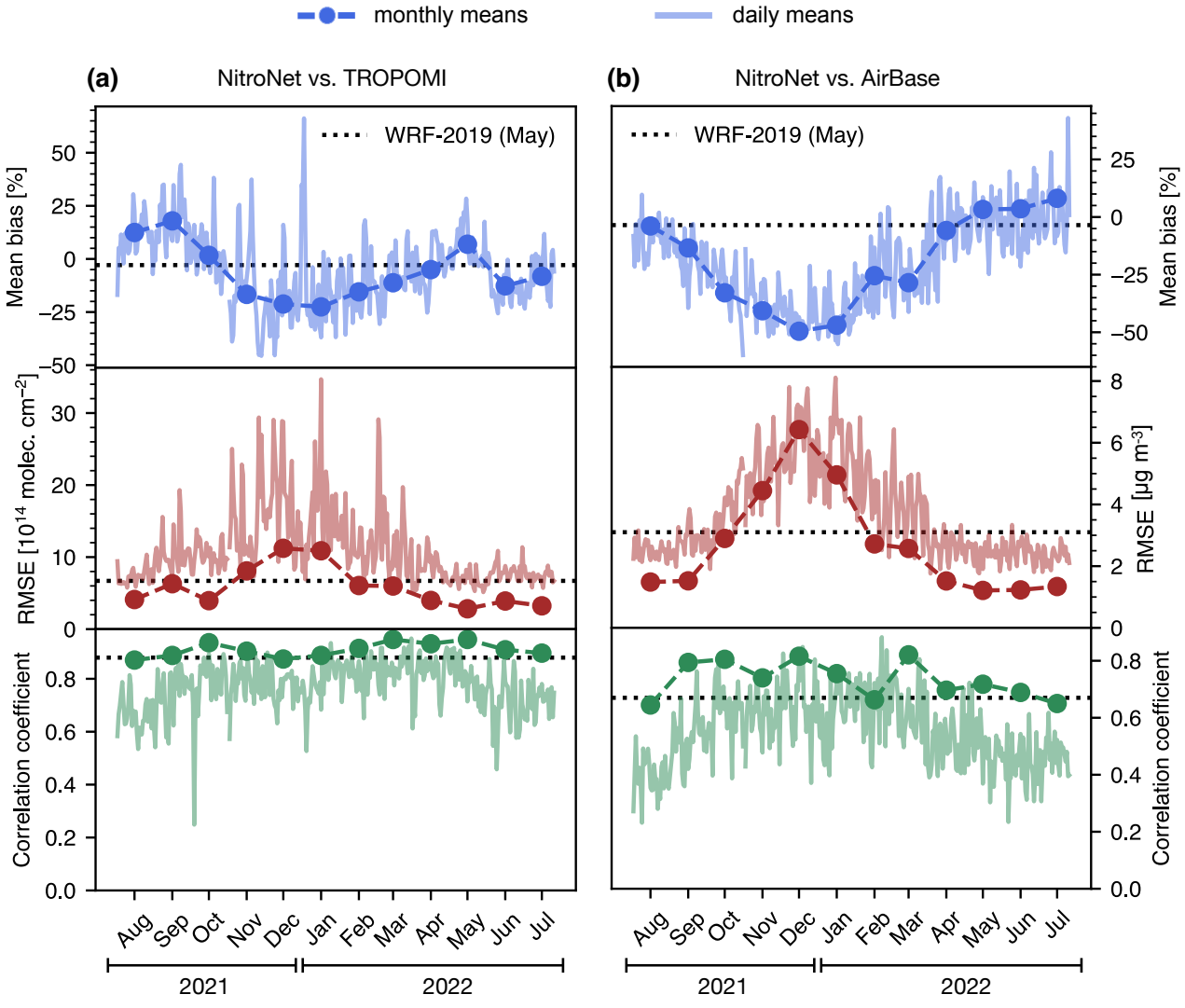

Figure 13. Seasonal evaluation of NitroNet on the central European domain against **(a)** NO_2 VCDs from TROPOMI and **(b)** surface observations from AirBase. The gray dotted line "WRF-2019 (May)" shows the value of the statistical diagnostics (mean bias, RMSE, and correlation coefficient) obtained from WRF-2019 for comparison.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

2. The NO_2 in-situ measurements used in empirical training sets contain a hidden NO_z bias of typically > 20 % due to cross sensitivities to atmospheric oxidants. Without access to model data, this bias can not be corrected, and is silently reproduced by other neural networks.

3. The abundance of training data from the RCT simulation allows for generous dismissal of untrustworthy training examples without running into data shortage. We can therefore train the neural network on filtered data, which was purged from erroneous example profiles. The neural network can then exceed the prediction quality of the original RCT simulation.

The latter concept of "learning from the good examples, but dismissing the errors" of a data generating model was explored in other publications (e.g. Sayeed et al. (2023); Li et al. (2023)), although in a somewhat different context. These publications describe the development of synergistic neural network + RCT combination models, while NitroNet is designed for standalone use as a surrogate model for the computationally expensive and slow RCT simulations. To put this into perspective: Using 800 CPUs, it took ~ 5 days to produce one month of WRF-Chem simulation data, while NitroNet can process the same amount of data in just ~ 20 minutes using 31 GPUs, with obvious operational advantages. Nevertheless, this functionality is limited to the prediction of NO_2 profiles, and NitroNet can not be considered a full replacement for RCT simulations, which can predict the concentrations of many other trace gases and aerosols, as well as meteorological variables.

Our main results were reported in sect. 4.2 in the form of an extensive evaluation of the NitroNet model. Three observational datasets (NO_2 VCDs from TROPOMI, background in-situ observations from AirBase, NO_2 profiles from FRM$_4$DOAS) were used as monthly-mean reference data. First, an inter-comparison between NitroNet, WRF-Chem, TROPOMI and AirBase was performed for May 2019. Hereby, the benefits of training the neural network on filtered data were demonstrated. NitroNet showed far better agreement to TROPOMI NO_2 VCDs than did WRF-Chem, while the comparison to AirBase surface observations returned similar results for both models. The NO_z cross sensitivities of the in-situ measurements were estimated based on modelled PAN and HNO_3 mixing ratios, resulting in significant bias correction factors of up to +200 %.

Next, NitroNet was evaluated on previously unseen data of May 2022. The comparison to TROPOMI NO_2 VCDs showed a strong correlation of $R = 0.95$, a bias of $+6.7$ % and an RMSE of $2.8 \cdot 10^{14}$ molec. cm^{-2}. The comparison to FRM$_4$DOAS NO_2 profiles showed good agreement when using the MMF retrieval algorithm (RMSE $\approx 4 \cdot 10^9$ molec. cm^{-3}), and slightly worse results when using the MAPA retrieval (RMSE $\approx 6 \cdot 10^9$ molec. cm^{-3}). The comparison to AirBase surface observations resulted in a correlation of $R = 0.75$, a bias of -10.5 % and an RMSE of 1.7 μg m^{-3}. By omitting the instruments categorized as "urban background", the bias and RMSE were reduced to $+2.2\%$, and 1.2 μg m^{-3}.

Lastly, the model evaluation was extended to different seasons (central European domain, August 2021 - July 2022) and regions of the earth (May 2022, UK, Spain and Portugal, US west coast, India, and China). Over the UK, Spain and Portugal, and the US west coast, NitroNet performed similarly well as in the original central European training domain. Over India and China, larger deviations and weaker correlations were found. The strongest differences occurred in the heavily industrialized regions of northern China, where the emission data used as model input might have been outdated. In all domains (except for the UK), NitroNet consistently overestimated the NO_2 load over waterbodies by approximately 10^{15} molec. cm^{-2}. The

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

seasonal analysis revealed stable model performance in spring, summer, and early fall (March - September), but significant low biases of up to -50 % in surface concentrations during late fall and winter (October - February). Part of these underestimations may be attributed to the higher uncertainties of the main model input, the NO_2 VCD, during wintertime.

In closing this article, we give an outlook on future improvements and use cases of NitroNet. We plan to produce a full year of synthetic training data, possibly in more diverse geographical regions. This will result in more consistent model accuracy across different seasons and regions of the world. Until then, NitroNet should be considered a prototype. Furthermore, the inclusion of more data from new instruments will strongly influence the training and validation of future model versions. Here the most promising outlook is the advent of geostationary satellites, such as GEMS (see Kim et al. (2020)), TEMPO (see Naeger et al. (2021)), and Sentinel-4 (see Stark et al. (2013)). These will provide hourly resolved NO_2 columns, allowing for the implementation of diurnal cycles into our model. The ongoing efforts in harmonizing observational datasets (see e.g. the GHOST dataset, see Bowdalo et al. (2024)) will allow for easier model validation at the surface in all regions of the Earth. The use of more intricate MAX-DOAS retrieval algorithms could allow for better sensitivity to higher layers of the troposphere (see e.g. Schofield et al. (2004), who achieve sensitivity to the stratosphere and upper troposphere with a zenith-sky viewing geometry). Lastly, more complex neural network designs, such as the invertible neural networks (INNs, see Ardizzone et al. (2018)), or physically informed neural networks (PINNs, see Raissi et al. (2019)), may be implemented once the remaining parts of the project are deemed mature enough. The NitroNet model can be used for scientific research, such as:

1. A revision of existing studies on near-surface air pollution and the associated effect on human health, with explicit treatment of the NO_z biases of in-situ measurements.

2. Reprocessing of the TROPOMI NO_2 columns by replacing the poorly resolved NO_2 a priori profiles from the TM5 model (horizontal resolution: $1° \times 1°$) with the much better resolved NO_2 profiles from NitroNet (horizontal resolution: $3.5\ \mathrm{km} \times 5.5\ \mathrm{km}$).

3. Possibly the prediction of other trace gas profiles, such as SO_2 or HCHO.

Altogether, the combined efforts of machine learning, RCT modelling, and instrumental development hold promising potential for the near future.

Data availability. All data is available from the authors upon reasonable request.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Author contributions. LK developed the question of research under the supervision of TW and SB. LK, SO, and AP produced the training data of the neural network. LK wrote the text and produced the remaining content of the article, with all authors contributing by revising it interactively.

Competing interests. Some authors are members of the editorial board of the Atmospheric Measurement Techniques journal. The authors declare that they have no other conflict of interest.

Acknowledgements. We acknowledge Vinod Kumar for his invaluable help on RCT modelling, which has preceded this article. We thank Andreas Richter, Udo Frieß, Ankie Piters, Michel van Roozendael, Alexis Merlaud, Elisa Castelli, and Paolo Pettinari for maintaining the MAX-DOAS instruments and sharing their data within the FRM_4DOAS network. Data analysis and visualization were performed using the Python programming language, including the libraries NumPy, SciPy, pandas, Xarray, matplotlib, and cartopy. The neural network of NitroNet was implemented using the PyTorch package.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

References

Anderson, G.: Error propagation by the Monte Carlo method in geochemical calculations, Geochimica et Cosmochimica Acta, 40, 1533–1538, https://doi.org/10.1016/0016-7037(76)90092-2, 1976.

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U.: Analyzing Inverse Problems with Invertible Neural Networks, https://doi.org/10.48550/ARXIV.1808.04730, 2018.

Beckwith, M., Bates, E., Gillah, A., and Carslaw, N.: NO_2 hotspots: Are we measuring in the right places?, Atmospheric Environment: X, 2, 100 025, https://doi.org/10.1016/j.aeaoa.2019.100025, 2019.

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y., and Wagner, T.: The Mainz profile algorithm (MAPA), Atmospheric Measurement Techniques, 12, 1785–1806, https://doi.org/10.5194/amt-12-1785-2019, 2019.

Bergstra, J. and Bengio, Y.: Random Search for Hyper-Parameter Optimization, Journal of Machine Learning Research, 13, 281–305, http://jmlr.org/papers/v13/bergstra12a.html, 2012.

Berkhout, A., Gast, L., van der Hoff, G., Swart, D., Hoed, M., and Allaart, M.: Atmospheric NO_2 profiles measured with lidar during the CINDI-2 campaign, EPJ Web of Conferences, 176, 10 002, https://doi.org/10.1051/epjconf/201817610002, 2018.

Bieser, J., Aulinger, A., Matthias, V., Quante, M., and van der Gon, H.: Vertical emission profiles for Europe based on plume rise calculations, Environmental Pollution, 159, 2935–2946, https://doi.org/10.1016/j.envpol.2011.04.030, 2011.

Boersma, K. F., Jacob, D. J., Trainic, M., Rudich, Y., DeSmedt, I., Dirksen, R., and Eskes, H. J.: Validation of urban NO_2 concentrations and their diurnal and seasonal variations observed from the SCIAMACHY and OMI sensors using in situ surface measurements in Israeli cities, Atmospheric Chemistry and Physics, 9, 3867–3879, https://doi.org/10.5194/acp-9-3867-2009, 2009.

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.: Comparison of airborne measurements of NO, NO_2, HONO, NO_y, and CO during FIREX-AQ, Atmospheric Measurement Techniques, 15, 4901–4930, https://doi.org/10.5194/amt-15-4901-2022, 2022.

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.: GHOST: A globally harmonised dataset of surface atmospheric composition measurements, Earth System Science Data Discussions, 2024, 1–137, https://doi.org/10.5194/essd-2023-397, 2024.

Bösch, T.: Detailed analysis of MAX-DOAS measurements in Bremen: Spatial and temporal distribution of aerosols, formaldehyde and nitrogen dioxide, Ph.D. thesis, Universität Bremen, 2018.

Cao, E. L.: National ground-level NO_2 predictions via satellite imagery driven convolutional neural networks, Frontiers in Environmental Science, 11, https://doi.org/10.3389/fenvs.2023.1285471, 2023.

Chan, K. L., Khorsandi, E., Liu, S., Baier, F., and Valks, P.: Estimation of Surface NO_2 Concentrations over Germany from TROPOMI Satellite Observations Using a Machine Learning Method, Remote Sensing, 13, 969, https://doi.org/10.3390/rs13050969, 2021.

Chowdhury, S., Haines, A., Klingmüller, K., Kumar, V., Pozzer, A., Venkataraman, C., Witt, C., and Lelieveld, J.: Global and national assessment of the incidence of asthma in children and adolescents from major sources of ambient NO_2, Environmental Research Letters, 16, 035 020, https://doi.org/10.1088/1748-9326/abe909, 2021.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Crippa, M., Guizzardi, D., Oreggioni, G., Muntean, M., and Schaaf, E.: EDGARv5.0 Air Pollutant Emissions [Data Set], https://doi.org/10.1594/PANGAEA.921922, 2020.

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.: Comparing Sentinel-5P TROPOMI NO_2 column observations with the CAMS regional air quality ensemble, Geoscientific Model Development, 16, 509–534, https://doi.org/10.5194/gmd-16-509-2023, 2023.

Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., Martini, F. S., 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., Villegas, C. R. R., Kolb, C. E., Molina, L. T., and Molina, M. J.: Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment, Atmospheric Chemistry and Physics, 7, 2691–2704, https://doi.org/10.5194/acp-7-2691-2007, 2007.

Elshorbany, Y. F., Steil, B., Brühl, C., and Lelieveld, J.: Impact of HONO on global atmospheric chemistry calculated with an empirical parameterization in the EMAC model, Atmospheric Chemistry and Physics, 12, 9977–10 000, https://doi.org/10.5194/acp-12-9977-2012, 2012.

Emmons, L. K., Schwantes, R. H., Orlando, J. J., Tyndall, G., Kinnison, D., Lamarque, J.-F., 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.: The Chemistry Mechanism in the Community Earth System Model Version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, https://doi.org/10.1029/2019ms001882, 2020.

Eskes, H., van Geffen, J., Sneep, M., Apituley, A., and Veefkind, J.: Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Nitrogendioxide, Royal Netherlands Meteorological Institute, 2019.

European Environment Agency: Air Quality e-Reporting [Data Set], https://www.eea.europa.eu/data-and-maps/data/aqereporting-8, last access: 10th March 2024.

European Environment Agency: EMEP/EEA air pollutant emission inventory guidebook 2023, https://doi.org/10.2800/795737, 2023.

Faustini, A., Rapp, R., and Forastiere, F.: Nitrogen dioxide and mortality: review and meta-analysis of long-term studies, European Respiratory Journal, 44, 744–753, https://doi.org/10.1183/09031936.00114713, 2014.

Fayt, C., Friedrich, M., and Hendrick, F.: 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, 2021.

Friedrich, M. M., Rivera, C., Stremme, W., Ojeda, Z., Arellano, J., Bezanilla, A., García-Reynoso, J. A., and Grutter, M.: NO_2 vertical profiles and column densities from MAX-DOAS measurements in Mexico City, Atmospheric Measurement Techniques, 12, 2545–2565, https://doi.org/10.5194/amt-12-2545-2019, 2019.

Gardner, M. and Dorling, S.: Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London, Atmospheric Environment, 33, 709–719, https://doi.org/https://doi.org/10.1016/S1352-2310(98)00230-1, 1999.

Ghahremanloo, M., Lops, Y., Choi, Y., and Yeganeh, B.: Deep Learning Estimation of Daily Ground-Level NO_2 Concentrations From Remote Sensing Data, Journal of Geophysical Research: Atmospheres, 126, https://doi.org/10.1029/2021jd034925, 2021.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and Eder, B.: Fully coupled "online" chemistry within the WRF model, Atmospheric Environment, 39, 6957–6975, https://doi.org/10.1016/j.atmosenv.2005.04.027, 2005.

He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, https://doi.org/10.48550/ARXIV.1502.01852, 2015.

Hersbach, H. and Dee, D.: ERA5 reanalysis in production, ECMWF Newsletter 147, 2017.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Hönninger, G., von Friedeburg, C., and Platt, U.: Multi axis differential optical absorption spectroscopy (MAX-DOAS), Atmospheric Chemistry and Physics, 4, 231–254, https://doi.org/10.5194/acp-4-231-2004, 2004.

Ialongo, I., Virta, H., Eskes, H., Hovila, J., and Douros, J.: Comparison of TROPOMI/Sentinel-5 Precursor NO_2 observations with ground-based measurements in Helsinki, Atmospheric Measurement Techniques, 13, 205–218, https://doi.org/10.5194/amt-13-205-2020, 2020.

Jesemann, A.-S., Matthias, V., Böhner, J., and Bechtel, B.: Using Neural Network NO_2-Predictions to Understand Air Quality Changes in Urban Areas — A Case Study in Hamburg, Atmosphere, 13, 1929, https://doi.org/10.3390/atmos13111929, 2022.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., and Kim, S.: Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia, Environmental Pollution, 288, 117 711, https://doi.org/10.1016/j.envpol.2021.117711, 2021.

Kerkweg, A. and Jöckel, P.: The 1-way on-line coupled atmospheric chemistry model system MECO(n) – Part 1: Description of the limited-area atmospheric chemistry model COSMO/MESSy, Geoscientific Model Development, 5, 87–110, https://doi.org/10.5194/gmd-5-87-2012, 2012.

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., kyun Kim, S., 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.: New Era of Air Quality Monitoring from Space: Geostationary Environment Monitoring Spectrometer (GEMS), Bulletin of the American Meteorological Society, 101, E1–E22, https://doi.org/10.1175/bams-d-18-0013.1, 2020.

Krol, M., Houweling, S., Bregman, B., van den Broek, M., Segers, A., van Velthoven, P., Peters, W., Dentener, F., and Bergamaschi, P.: The two-way nested global chemistry-transport zoom model TM5: algorithm and applications, Atmospheric Chemistry and Physics, 5, 417–432, https://doi.org/10.5194/acp-5-417-2005, 2005.

Kuhn, L., Beirle, S., Kumar, V., Osipov, S., Pozzer, A., Bösch, T., Kumar, R., and Wagner, T.: 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, 185–217, https://doi.org/10.5194/acp-24-185-2024, 2024.

Kuik, F., Lauer, A., Churkina, G., van der Gon, H. A. C. D., Fenner, D., Mar, K. A., and Butler, T. M.: 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, 4339–4363, https://doi.org/10.5194/gmd-9-4339-2016, 2016.

Kuik, F., Kerschbaumer, A., Lauer, A., Lupascu, A., von Schneidemesser, E., and Butler, T. M.: 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, 8203–8225, https://doi.org/10.5194/acp-18-8203-2018, 2018.

Lamsal, L. N., Martin, R. V., van Donkelaar, A., Steinbacher, M., Celarier, E. A., Bucsela, E., Dunlea, E. J., and Pinto, J. P.: Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/https://doi.org/10.1029/2007JD009235, 2008.

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,

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

C., Uhlmannsiek, K., Wagner, T., Ruhtz, T., Eskes, H., Bohn, B., Santana Diaz, D., Abuhassan, N., Schüttemeyer, D., and Burrows, J. P.: 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, 1357–1389, https://doi.org/10.5194/amt-16-1357-2023, 2023.

Li, B., Hu, Q., Gao, M., Liu, T., Zhang, C., and Liu, C.: Physical informed neural network improving the WRF-CHEM results of air pollution using satellite-based remote sensing data, Atmospheric Environment, 311, 120 031, https://doi.org/10.1016/j.atmosenv.2023.120031, 2023.

Liu, F., Beirle, S., Zhang, Q., Dörner, S., He, K., and Wagner, T.: NO_x lifetimes and emissions of cities and power plants in polluted background estimated by satellite observations, Atmospheric Chemistry and Physics, 16, 5283–5298, https://doi.org/10.5194/acp-16-5283-2016, 2016.

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.: An improved TROPOMI tropospheric NO_2 research product over Europe, Atmospheric Measurement Techniques, 14, 7297–7327, https://doi.org/10.5194/amt-14-7297-2021, 2021.

Manders, A. M. M., Builtjes, P. J. H., Curier, L., van der Gon, H. A. C. D., Hendriks, C., Jonkers, S., Kranenburg, R., Kuenen, J. J. P., Segers, A. J., Timmermans, R. M. A., Visschedijk, A. J. H., Kruit, R. J. W., van Pul, W. A. J., Sauter, F. J., van der Swaluw, E., Swart, D. P. J., Douros, J., Eskes, H., van Meijgaard, E., van Ulft, B., van Velthoven, P., Banzhaf, S., Mues, A. C., Stern, R., Fu, G., Lu, S., Heemink, A., van Velzen, N., and Schaap, M.: Curriculum vitae of the LOTOS–EUROS (v2.0) chemistry transport model, Geoscientific Model Development, 10, 4145–4173, https://doi.org/10.5194/gmd-10-4145-2017, 2017.

Meng, F., Zhang, Y., Kang, J., Heal, M. R., Reis, S., Wang, M., Liu, L., Wang, K., Yu, S., Li, P., Wei, J., Hou, Y., Zhang, Y., Liu, X., Cui, Z., Xu, W., and Zhang, F.: Trends in secondary inorganic aerosol pollution in China and its responses to emission controls of precursors in wintertime, Atmospheric Chemistry and Physics, 22, 6291–6308, https://doi.org/10.5194/acp-22-6291-2022, 2022.

Menut, L., Bessagnet, B., Briant, R., Cholakian, A., Couvidat, F., Mailler, S., Pennel, R., Siour, G., Tuccella, P., Turquety, S., and Valari, M.: The CHIMERE v2020r1 online chemistry-transport model, Geoscientific Model Development, 14, 6781–6811, https://doi.org/10.5194/gmd-14-6781-2021, 2021.

Mills, I. C., Atkinson, R. W., Kang, S., Walton, H., and Anderson, H. R.: Quantitative systematic review of the associations between short-term exposure to nitrogen dioxide and mortality and hospital admissions, BMJ Open, 5, e006 946, https://doi.org/10.1136/bmjopen-2014-006946, 2015.

Naeger, A. R., Newchurch, M. J., Moore, T., Chance, K., Liu, X., Alexander, S., Murphy, K., and Wang, B.: Revolutionary Air-Pollution Applications from Future Tropospheric Emissions: Monitoring of Pollution (TEMPO) Observations, Bulletin of the American Meteorological Society, 102, E1735–E1741, https://doi.org/10.1175/bams-d-21-0050.1, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems 32, 32, 8024–8035, https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, https://doi.org/10.48550/ARXIV.1201.0490, 2012.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Peng, S., Giron, C., Liu, G., d'Aspremont, A., Benoit, A., Lauvaux, T., Lin, X., de Almeida Rodrigues, H., Saunois, M., and Ciais, P.: High-resolution assessment of coal mining methane emissions by satellite in Shanxi, China, iScience, 26, 108 375, https://doi.org/10.1016/j.isci.2023.108375, 2023.

Platt, U. and Stutz, J.: Differential Optical Absorption Spectroscopy, Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-540-75776-4, 2008.

Poraicu, C., Müller, J.-F., Stavrakou, T., Fonteyn, D., Tack, F., Deutsch, F., Laffineur, Q., Van Malderen, R., and Veldeman, N.: Cross-evaluating WRF-Chem v4.1.2, TROPOMI, APEX, and in situ NO_2 measurements over Antwerp, Belgium, Geoscientific Model Development, 16, 479–508, https://doi.org/10.5194/gmd-16-479-2023, 2023.

Raissi, M., Perdikaris, P., and Karniadakis, G.: 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, https://doi.org/10.1016/j.jcp.2018.10.045, 2019.

Riess, T. C. V. W., Boersma, K. F., Van Roy, W., de Laat, J., Dammers, E., and van Vliet, J.: 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, 5287–5304, https://doi.org/10.5194/amt-16-5287-2023, 2023.

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding, World Scientific, https://doi.org/10.1142/3171, 2000.

Ruder, S.: An overview of gradient descent optimization algorithms, https://doi.org/10.48550/ARXIV.1609.04747, 2016.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323, 533–536, https://doi.org/10.1038/323533a0, 1986.

Ruppert, D.: Trimming and Winsorization, Wiley StatsRef: Statistics Reference Online, https://doi.org/10.1002/9781118445112.stat01887, 2014.

Sayeed, A., Choi, Y., Jung, J., Lops, Y., Eslami, E., and Salman, A. K.: A Deep Convolutional Neural Network Model for Improving WRF Simulations, IEEE Transactions on Neural Networks and Learning Systems, 34, 750–760, https://doi.org/10.1109/tnnls.2021.3100902, 2023.

Schmidhuber, J.: Deep learning in neural networks: An overview, Neural Networks, 61, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, 2015.

Schofield, R., Connor, B., Kreher, K., Johnston, P., and Rodgers, C.: The retrieval of profile and chemical information from ground-based UV-visible spectroscopic measurements, Journal of Quantitative Spectroscopy and Radiative Transfer, 86, 115–131, https://doi.org/https://doi.org/10.1016/S0022-4073(03)00278-4, 2004.

Shah, V., Jacob, D. J., Li, K., Silvern, R. F., Zhai, S., Liu, M., Lin, J., and Zhang, Q.: 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, 1483–1495, https://doi.org/10.5194/acp-20-1483-2020, 2020.

Shapley, L. S.: Notes on the N-Person Game — II. The Value of an N-Person Game, RAND Corporation, https://doi.org/10.7249/rm0670, 1951.

Sluis, W. W., Allaart, M. A. F., Piters, A. J. M., and Gast, L. F. L.: The development of a nitrogen dioxide sonde, Atmospheric Measurement Techniques, 3, 1753–1762, https://doi.org/10.5194/amt-3-1753-2010, 2010.

Stark, H., Möller, H., Courrèges-Lacoste, G., Koopman, R., Mezzasoma, S., and Veihelmann, B.: The Sentinel-4 mission and its implementation, Proceedings of the ESA Living Planet Symposium, Edinburgh, 2013.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Steinbacher, M., Zellweger, C., Schwarzenbach, B., Bugmann, S., Buchmann, B., Ordóñez, C., Prevot, A. S. H., and Hueglin, C.: Nitrogen oxide measurements at rural sites in Switzerland: Bias of conventional measurement techniques, Journal of Geophysical Research, 112, https://doi.org/10.1029/2006jd007971, 2007.

Su, J., McCormick, M. P., Johnson, M. S., Sullivan, J. T., Newchurch, M. J., Berkoff, T. A., Kuang, S., and Gronoff, G. P.: Tropospheric NO_2 measurements using a three-wavelength optical parametric oscillator differential absorption lidar, Atmospheric Measurement Techniques, 14, 4069–4082, https://doi.org/10.5194/amt-14-4069-2021, 2021.

Tack, F., Merlaud, A., Iordache, M.-D., Pinardi, G., Dimitropoulou, E., Eskes, H., Bomans, B., Veefkind, P., and Van Roozendael, M.: Assessment of the TROPOMI tropospheric NO_2 product based on airborne APEX observations, Atmospheric Measurement Techniques, 14, 615–646, https://doi.org/10.5194/amt-14-615-2021, 2021.

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.: Sentinel-5P TROPOMI NO_2 retrieval: impact of version v2.2 improvements and comparisons with OMI and ground-based data, Atmospheric Measurement Techniques, 15, 2037–2060, https://doi.org/10.5194/amt-15-2037-2022, 2022.

van Geffen, J., Eskes, H. J., Boersma, K., and Veefkind, J.: TROPOMI ATBD of the total and tropospheric NO_2 data products, Royal Netherlands Meteorological Institute, 2022.

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.: 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, https://doi.org/10.1016/j.rse.2011.09.027, 2012.

Villena, G., Bejan, I., Kurtenbach, R., Wiesen, P., and Kleffmann, J.: Interferences of commercial NO_2 instruments in the urban atmosphere and in a smog chamber, Atmospheric Measurement Techniques, 5, 149–159, https://doi.org/10.5194/amt-5-149-2012, 2012.

Visser, A. J., Boersma, K. F., Ganzeveld, L. N., and Krol, M. C.: European NO_x emissions in WRF-Chem derived from OMI: impacts on summertime surface ozone, Atmospheric Chemistry and Physics, 19, 11 821–11 841, https://doi.org/10.5194/acp-19-11821-2019, 2019.

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.: NO_2 lidar profile measurements for satellite interpretation and validation, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/10.1029/2009jd012441, 2009.

Zhang, C., Liu, C., Li, B., Zhao, F., and Zhao, C.: Spatiotemporal neural network for estimating surface NO_2 concentrations over north China and their human health impact, Environmental Pollution, 307, 119 510, https://doi.org/10.1016/j.envpol.2022.119510, 2022.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

Table A1. Overview of NitroNet's hyperparameters

Hyperparameter name	Sampling range	Optimal value
Hidden layers	4 - 8	8
Neurons per layer	100 - 400	326
Activation function	ReLU, PReLU, CELU, GELU, SELU	PReLU
Loss function	MSE, L_1, smooth L_1 [(1)], RMSLE	L_1
Batch size	2^9 - 2^{13}	2^{11}
Optimizer	NAdam, AdamW [(2)]	NAdam
Learning rate	10^{-6} - 10^{-3}	$3.4 \cdot 10^{-4}$
Batch normalization	True, False	False
Dropout probability [(3)]	0 - 0.15	0
Δ_{VCD} [(4)]	0 - 1	0.2
Δ_{PBLH} [(4)]	0 - 1	0.1

For a combined reference of these terms, see Schmidhuber (2015) and Paszke et al. (2019).
(1) see the PyTorch documentation: https://pytorch.org/docs/stable/generated/torch.nn.SmoothL1Loss.html
(2) see the PyTorch documentation: https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html
(3) Original range was 0 - 0.5, but training diverged for runs with dropout probability > 0.15.
(4) see sect. 3.3

Appendix A: Hyperparameter study

The hyperparameter study for NitroNet is based on 300 different model versions. The model configurations were sampled randomly ("random search", see Bergstra and Bengio (2012)). An overview of the hyperparameters and their respective sampling range can be found in table A1. Stochastic Gradient Descent (SGD) was not used, because all training runs using SGD diverged. The Adam optimizer was found to be out-classed by NAdam and AdamW early-on and subsequently omitted from the study. Figure A1 shows the results of the hyperparameter study in a parallel coordinate view. The validation MAPE, which is used as a performance metric to compare the model configurations, ranges from $\sim 10\ \%$ - $30\ \%$. This demonstrates, that a hyperparameter search can potentially improve the neural network's performance by up to a factor of 3, making it an essential step in the development of NitroNet.

Appendix B: Feature relevance analysis

In order to gain more insight into how the neural network of NitroNet operates, a feature relevance analysis was conducted. The goal is to quantify, how strongly each input variable contributes to the overall model performance. The standard method is to compute the *Shapley scores* of the input variables (see Shapley (1951)). The Shapley score of the i-th input variable x_i is

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

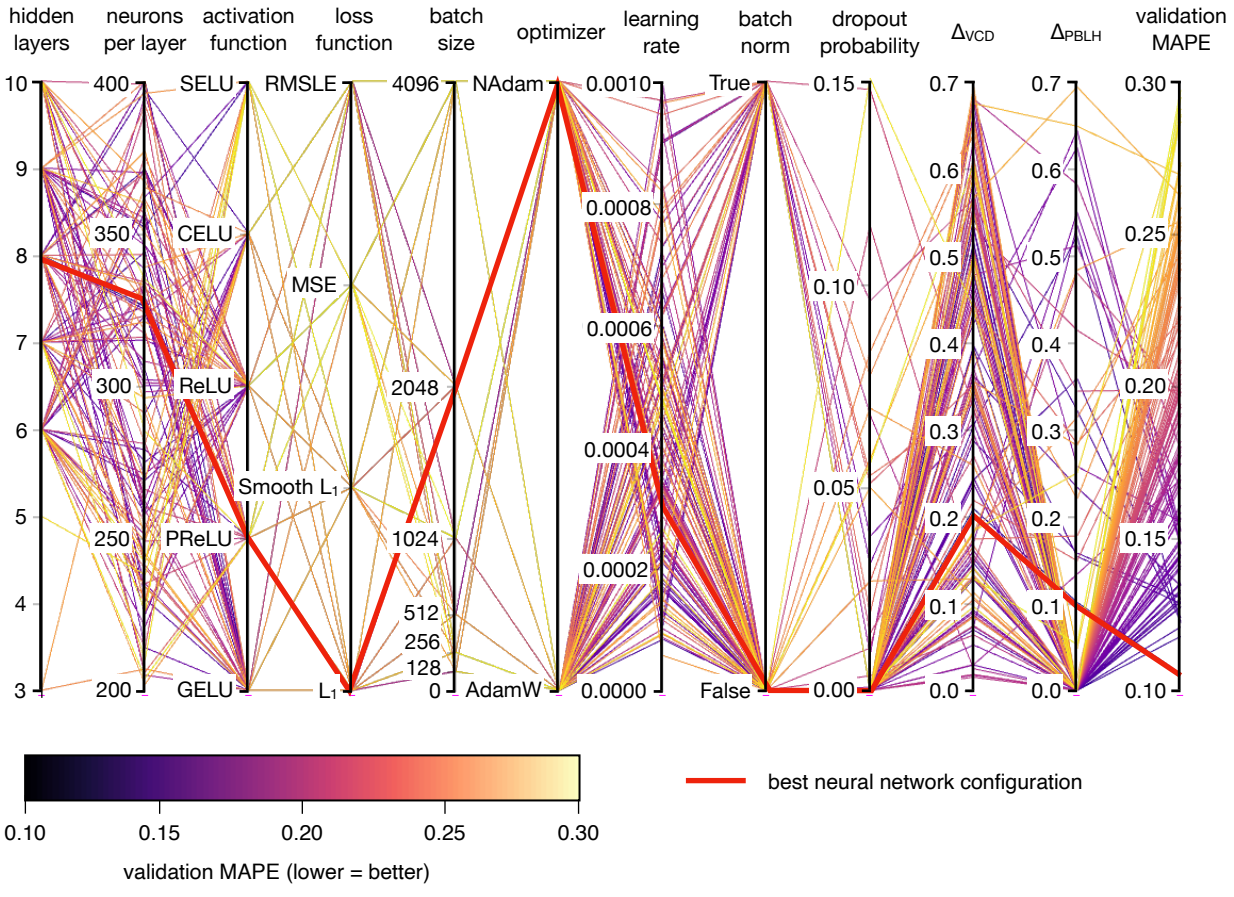

Figure A1. Results of the hyperparameter study in a parallel coordinate view. Each hyperparameter is represented by one vertical axis ("hidden layers", "neurons per layer", ...). Each variant of the neural network is represented by a contiguous line, intersecting the vertical axes at the network's hyperparameter values. The last vertical axis shows the MAPE achieved on the validation set, which acts as the metric for the selection of the best neural network configuration (lower = better). The optimal configuration is drawn as a thick red line.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

defined as

$$R_i = \sum_{S \subseteq P \setminus \{x_i\}} \frac{|S|!\,(|P| - |S| - 1)!}{|P|!} \left(f\left(S \cup \{x_i\}, y_{\text{true}}, y_{\text{pred}}\right) - f\left(S, y_{\text{true}}, y_{\text{pred}}\right) \right) \tag{B1}$$

where P denotes the set of all input variable variables, and $|\cdot|$ the set cardinality. $f(I, y_{\text{true}}, y_{\text{pred}})$ is a function of choice, which acts as a measure for model performance by comparison of the ground truth y_{true} vs. the model's predictions y_{pred}, using either all input variables (i.e. $I = S \cup \{x_i\}$) versus omitting variable x_i (i.e. $I = S$). Omission of an input variable x_i is simulated by replacing its values with with random samples from the validation set (approximating a sample drawn from the prior probability distribution of x_i.). The "feature relevance" F_i is obtained by normalization of the Shapley scores, i.e. $F_i = R_i / \sum_i R_i$. The following further premises were made:

1. We define

$$f = \frac{\text{RMSE}(I, y_{\text{true}}, y_{\text{pred}}) - \text{RMSE}(S = \emptyset, y_{\text{true}}, y_{\text{pred}})}{\text{RMSE}(S = P, y_{\text{true}}, y_{\text{pred}}) - \text{RMSE}(S = \emptyset, y_{\text{true}}, y_{\text{pred}})} \tag{B2}$$

   i.e. we we use a scaled RMSE to measure model performance. The "uninformed" case (omitting all input variables, $I = \emptyset$) equates to a model performance of $f = 0$, and the "fully informed" case (omitting none of the input variables, $I = P$) equates to a model performance of $f = 1$. Subsequently, all Shapley scores lay in the interval [0, 1].

2. Because the sum in eq. (B1) iterates over a power set of large cardinality, not all summands can be evaluated. Instead, R_i is approximated by computing random summands of eq. (B1) until the overall distribution of the feature relevances has converged.

3. Certain input variables are grouped together (e.g. the group "wind" contains all wind speed variables and does not discriminate between u and v direction).

The feature relevance can also be computed separately for each vertical layer. The resulting feature relevance profiles are shown in Fig. B1. We draw the following conclusions:

1. The NO_2 VCD is generally the most important input variable from 0 to 1500 m altitude.

2. The feature relevance of the PBLH peaks at $\sim$ 1800 m, which corresponds to the average PBLH value in WRF-2019. Because the NO_2 profiles show strong gradients at the top of the PBLH, this feature relevance profile shape is expected.

3. The NO_2 concentrations above the PBL are known to be low and weakly correlated to satellite observations. Here, the model performance is dominated by the input groups "surface class" and "tropospheric AMF", which the neural network most likely uses to predict average NO_2 profile estimates, based on coarse general constraints (e.g. "over water", "rural land", "urban land").

4. At the surface, there is a trade-off between the feature relevance of emission data and the NO_2 VCD. This confirms that emission data are a valuable addition to NitroNet, as they can improve the model performance by almost 20 %.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

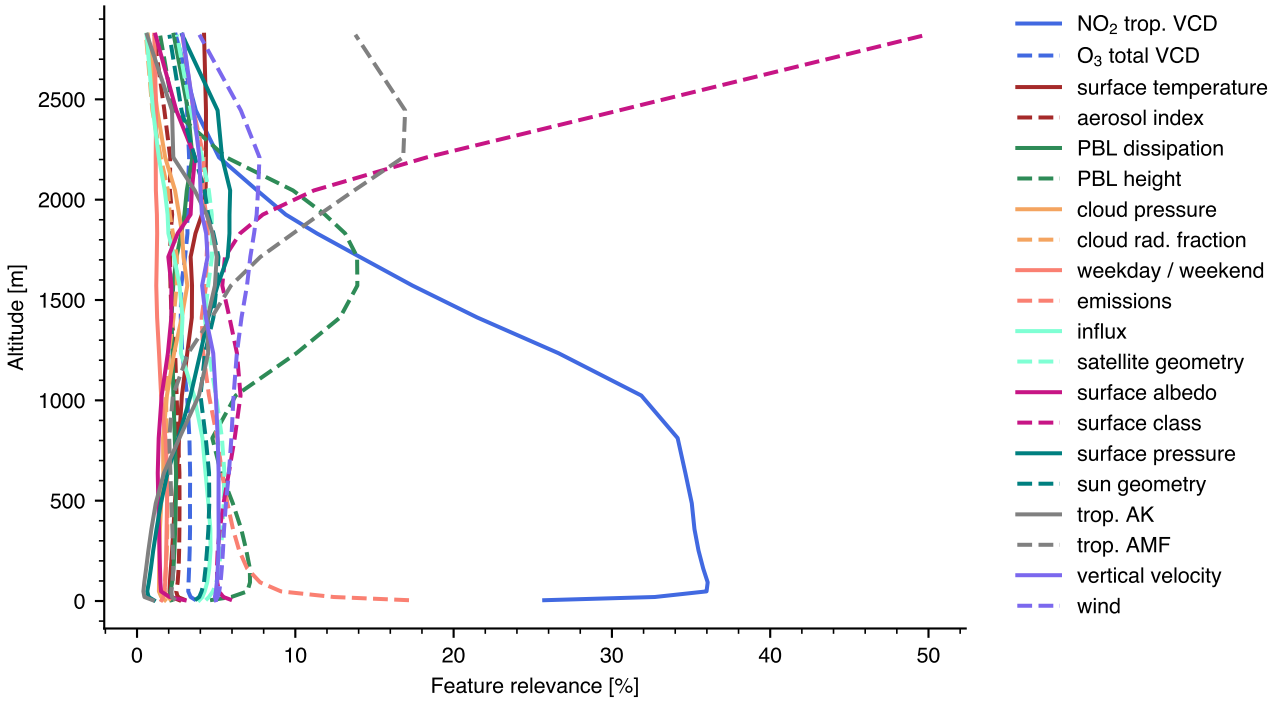

Figure B1. Vertically resolved feature relevance analysis of the NitroNet model.

The feature relevance of the emission data is further demonstrated in Fig. B2. Comparing Fig. B2a and B2b shows, that when no emission data is used, NitroNet's prediction of the NO_2 surface concentration is essentially proportional to the NO_2 VCD. Once emissions are added as input (see Fig. B2c), the distribution of predicted surface concentrations becomes significantly more complex: High values suddenly occur despite of comparably low VCDs (e.g. in the cities of Hamburg and Berlin, Germany) and fine-scale infrastructure, such as car highways connecting cities, becomes visible.

Appendix C: Additional figures

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

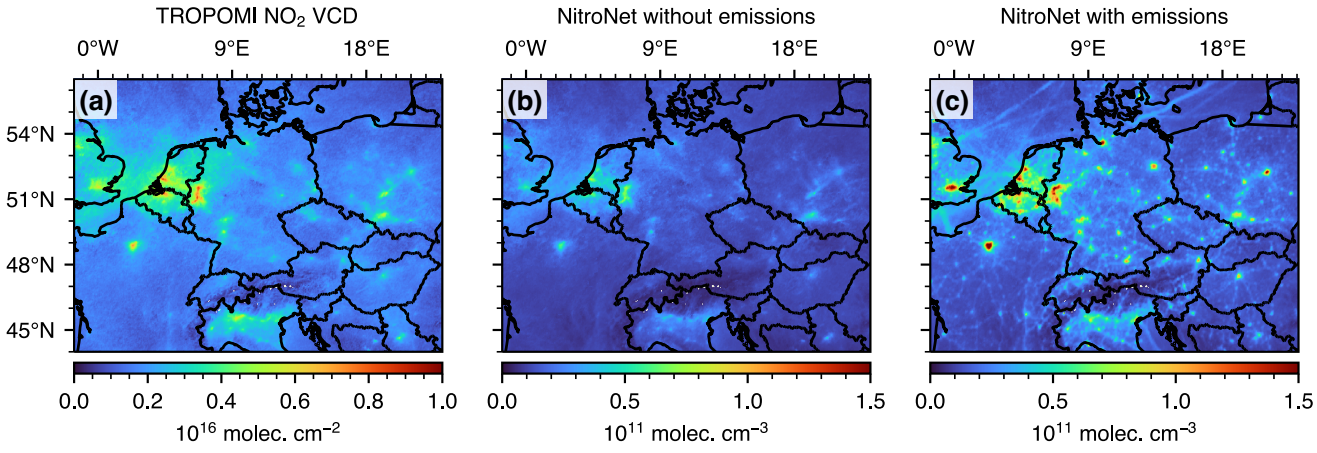

Figure B2. Demonstration of the "emissions" feature relevance. Subplot **(a)** shows the monthly-mean NO_2 VCD from TROPOMI (May, 2019). Subplots **(b)** and **(c)** show the corresponding NO_2 surface concentration from NitroNet with all emissions turned off / on, respectively.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

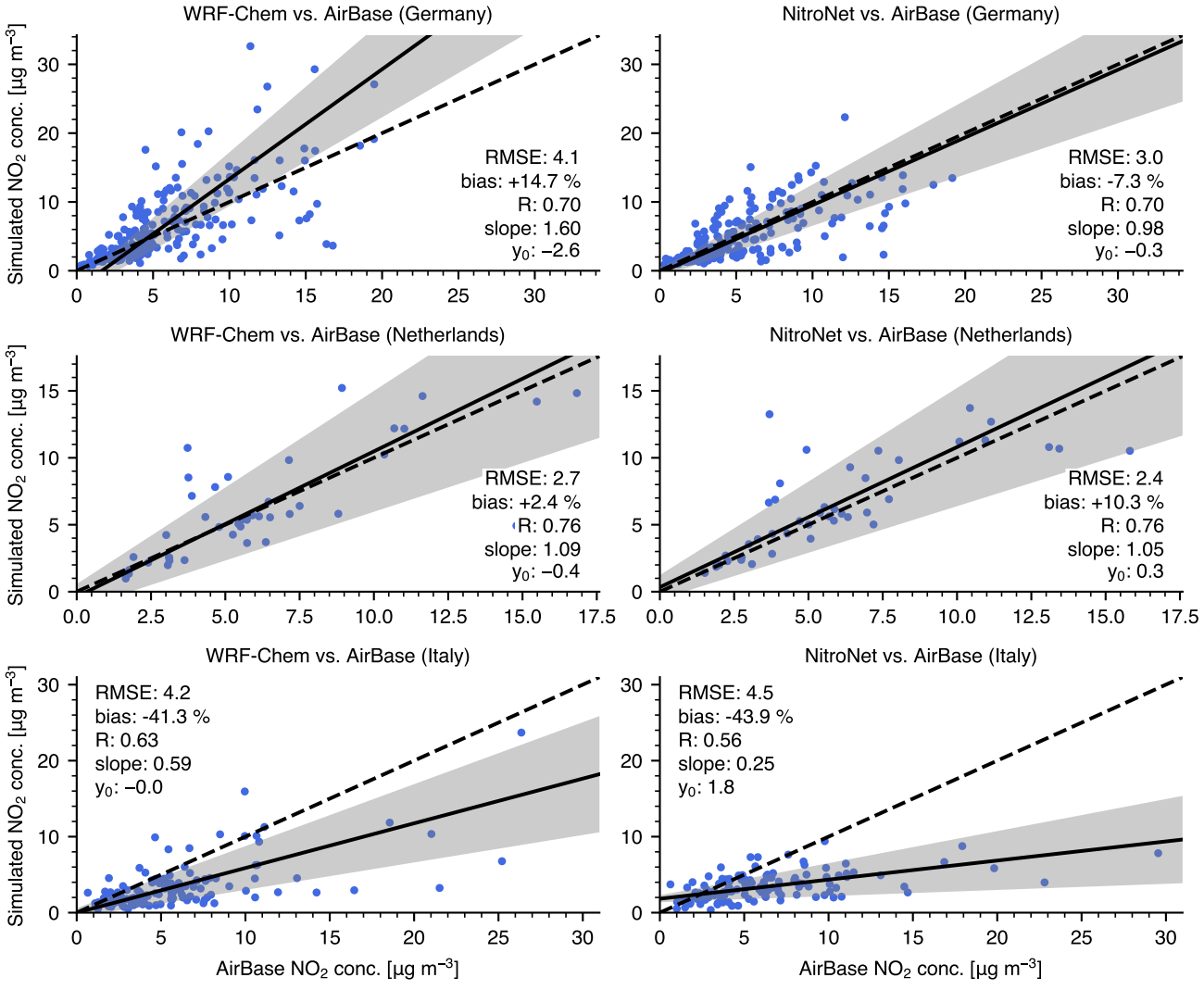

Figure C1. Scatter plots of the data shown in Fig. 5 restricted to individual countries (Germany, Netherlands, and Italy).

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

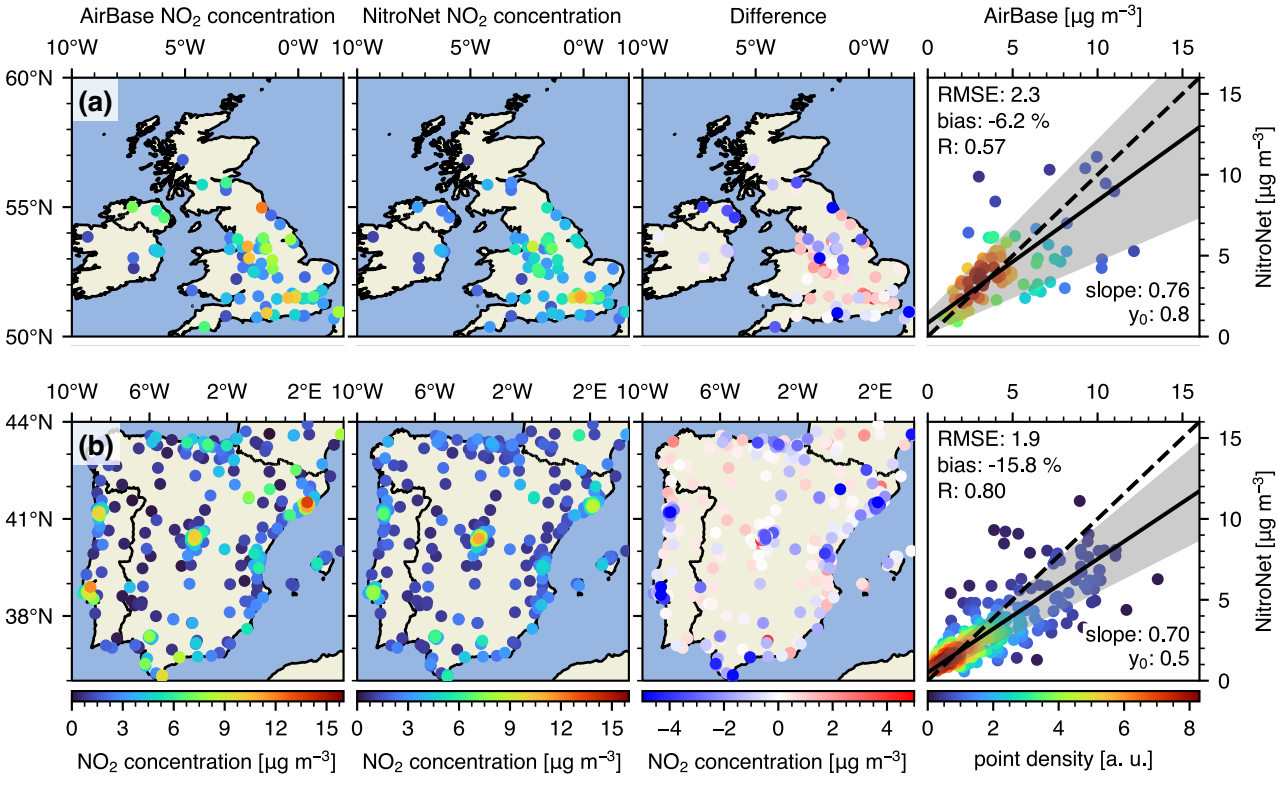

Figure C2. Like Fig. 11, but with urban stations included. RMSE and intercept are displayed in μg m^{-3}.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

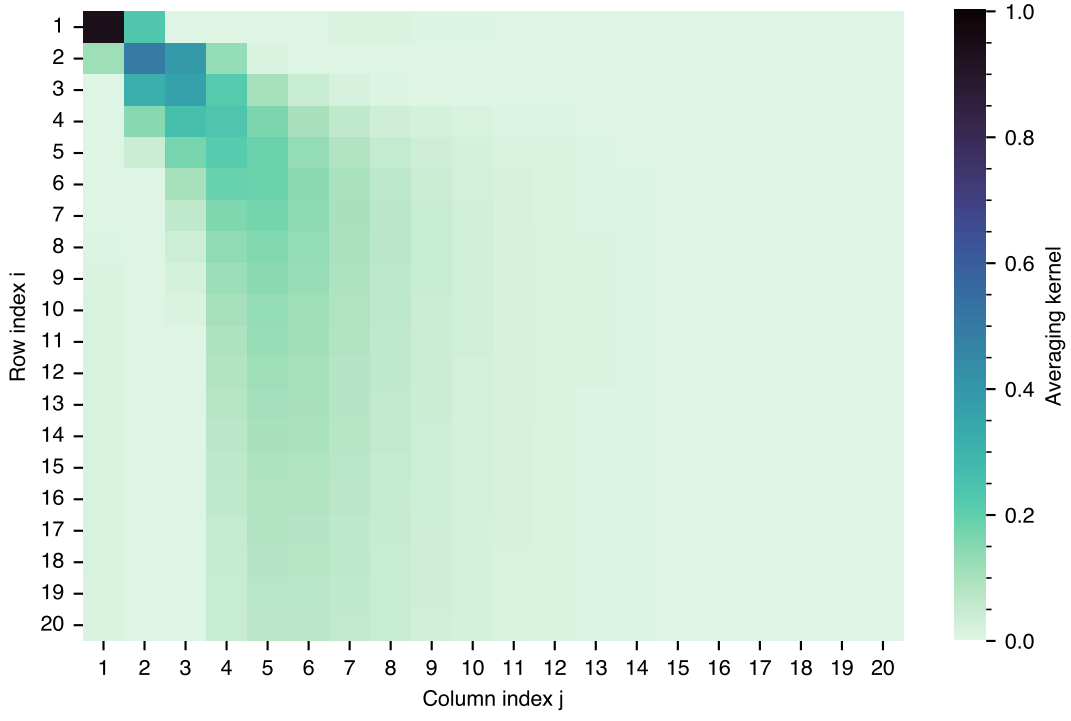

Figure C3. Monthly-mean averaging kernel matrix from the FRM_4DOAS instrument in Heidelberg, May 2022. The rows and columns are ordered such that index 1 represents the lowest layer of the retrieval, and index 20 the highest.

https://doi.org/10.5194/egusphere-2024-1196
Preprint. Discussion started: 21 May 2024

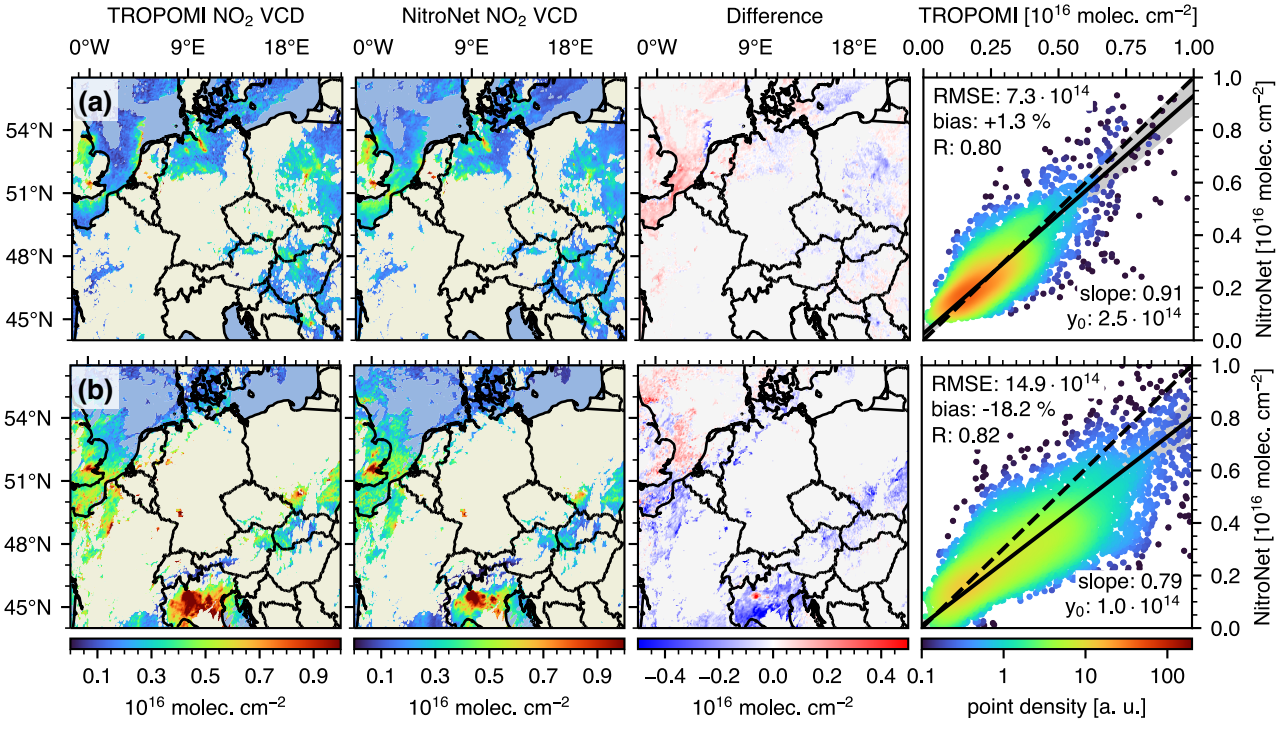

Figure C4. Like Fig. 7a, but for two single summer and winter days. **(a)** shows data of 2022-05-05. **(b)** shows data of 2021-11-05.