university of groningen / faculty of science and engineering / bernoulli institute

Forecasting Smog Clouds With Deep Learning: A Proof-Of-Concept

Bachelor's Project Thesis

Valentijn Oldenburg*, s4372948, v.w.oldenburg@student.rug.nl
Supervisor: J.D. Cardenas-Cartegena, M.Sc.

Abstract: Air pollution and smog carry correlations to numerous pervasive health effects. Given the risks, foreseeing toxic pollutant levels poses a vital challenge that, upon resolution, enacts a framework for life-saving decisions. Data-driven deep learning (DL) methods offer a novel approach to air quality prediction, yet their potential for modelling a combined set of smog-related pollutants with recurrent neural nets (RNNs) remains unexplored. In this proof-of-concept study, we conduct multivariate timeseries forecasting of nitrogen dioxide (NO_2), ozone (O_3), and (fine) particulate matter (PM_{10} & $PM_{2.5}$) with meteorological covariates between two locations in the Netherlands using various DL architectures, with a focus on RNNs with long short-term memory (LSTM) and gated recurrent unit (GRU) memory cells. In particular, we propose an integrated, hierarchical model architecture inspired by air pollution dynamics and atmospheric science that employs multi-task learning (MTL) and is benchmarked by unidirectional and fully-connected models. Results demonstrate that, above all, the hierarchical GRU proves itself as a competitive and efficient method for forecasting smog-related pollutants.

1 Introduction

Air pollution stands as a critical global challenge to humanity (UN, 2015). The rise of large-scale combustion and anthropogenic polluting activities has led to significant increases in air pollutant concentrations over the last century, leaving a heavy burden on human health (Kampa and Castanas, 2008). An unmistakable manifestation of these developments is the occurrence of smog: a noxious mixture of air pollutants that obstructs visibility and severely impairs human health in various ways (Brunekreef and Holgate, 2002). Given the detrimental effects, it is imperative to be able to predict when harmful pollutant levels might occur.

This research proposes different methods to gain insight into air pollutant levels through timeseries forecasting and the application of multiple deep neural network (DNN) architectures, notably including recurrent neural networks (RNNs). The following subsections will elaborate on the motivation for the research, the state-of-the-art, and its contributions.

1.1 Motivation

The presence of hazardous atmospheric chemicals characterises the phenomenon of air pollution. Although a number of physical activities (volcanoes, fire, etc.) may release different pollutants, anthropogenic activities are the head cause of environmental air pollution (Kampa and Castanas, 2008).

Adverse air pollution effects can range from skin irritation and difficulty in breathing to an increased risk of cardiac and respiratory illnesses, cancer, and mortality overall (Brunekreef and Holgate, 2002; Kampa and Castanas, 2008; Wong et al., 2008; Orellano et al., 2020). A recent addition is its direct link to COVID-19 morbidity and severity (Zorn et al., 2024). Furthermore, as stated in Lim et al. (2012), air pollution ranks high in the general disease burden attributable to environmental factors, with 3.1 million deaths in 2012 and 3.1% of disability-adjusted life years worldwide.

Since smog is primarily a form of air pollution characterised by elevated levels of specific pollutants, it carries comparable dangers, if not heightened ones. Moreover, distinctive smog features, such as reduced visibility and eye irritation, can exacerbate the risks.

Fundamentally, the air pollution problem and the extent to which it spreads is evident. Indeed, the air and its contaminants are everywhere and will remain inevitably inherent to human-nature interaction in the future. Positive notions present themselves, nonetheless: (1) humans, being the primary polluters, also possess the opportunity to act as the "primary purifiers"; and (2) comprehensive fundamental problem knowledge (refer, e.g.,

*Contact v.w.oldenburg@student.rug.nl for code access.

to Vallero (2014)) offers positive prospects for further advancing research. The subsequent step is to gather insight into the dynamics in which air pollution, smog, and their patterns reside: the weather system.

Despite the notable complexity of the weather system, (preliminary) insight into its dynamics and the ability to accurately forecast hazardous situations would provide a framework for making impactful decisions. An apparent use of this framework would be to prevent acute health damage from episodes of, for instance, heavy smog or hurricanes/cyclones. Another compelling goal is to get to the very source of the problem and deliver a more systematic policy-based counterforce to pollution hotspots as, e.g., Li et al. (2017b) showed that air pollution change is heavily policy-driven. A demonstrative example is a "congestion tax", where air pollution reduction resulted in, among other successes, a direct decrease in acute asthma in young children (Johansson et al., 2009)).

This problem motivates the development and application of data-driven forecasting models based on multiple DNN architectures, using contaminant and meteorological data to simulate and predict air pollution and smog. In particular, we consider the modelling of nitrogen dioxide (NO_2), ozone (O_3), and (fine) particulate matter (PM_{10} & $PM_{2.5}$) with various meteorological covariates as a first proof-of-concept (PoC). By employing these weather-predictive methods, this study aims to contribute incrementally to understanding air pollution dynamics and enhance environmental conditions for improved public health.

1.2 State-of-the-art

Traditional weather systems have evolved into sophisticated models that approximate real-world weather dynamics with increasing precision (Alley et al., 2019). The systems apply numerical weather prediction (NWP), a now ubiquitous, though computationally costly, numerical method grounded on physical first-principles (Bauer et al., 2015). While applying purely natural laws as boundary conditions for predictions is theoretically possible, it presents challenges in practice: the weather system is everywhere and contains numerous complex processes that make it computationally infeasible to provide these predictions with more than a highly simplified, parameterised value. Moreover, the non-linear dynamics, exemplified by the chaotic behaviour of turbulent flow, make predictions at high resolution—spatially, temporally, and/or across variables—a lasting challenge.

The emergence of data-driven methods presents a novel approach to abstracting physical processes embedded in the weather system. Machine learning (ML) models are adept at recognising complex patterns within large datasets with unparalleled efficiency—patterns that, speculatively, may represent relationships and correlations between atmospheric variables and influences not yet understood by traditional physics.

A recently undertaken application of large-scale data-driven deep learning (DL) weather forecasting is FourCastNet by Pathak et al. (2022). FourCastNet generates global forecasts orders of magnitude faster than traditional NWP with comparable or better accuracy. It herewith demonstrates the potential of data-driven methods to make significant progress in weather forecasting without explicitly considering the underlying (known) physical processes and equations. Implications are reducing costs of the traditional NWP and, more importantly, reducing the opportunity cost of inaccurate forecasts.

FourCastNet, specifically, combines the Fourier neural operator (FNO) learning approach, a vision transformer (ViT) architecture, and a dataset consisting of several atmospheric variables into predictions of precipitation, wind speed, and surface water vapour (Pathak et al., 2022). In the latter respect, utilising DL and many variables to predict few, FourCastNet is consistent with the endeavour of this study. Its scope does not, however, encompass predicting components directly related to air pollution or smog.

More closely related state-of-the-art studies (see Masood and Ahmad (2021) for a review) that distinctly forecast air pollution are Freeman et al. (2018) and Tao et al. (2019). The former performs a forecast of surface O_3 levels using an RNN with long short-term memory (LSTM); its approach takes as input hourly-sampled meteorological data and O_3 itself, outputting a multivariate 72-hour horizon forecast. The latter, Tao et al. (2019), respectively, highlights several prediction methods with a particular emphasis on a new method, a composition of 1D convnets and the bidirectional gated recurrent unit (GRU), for a multivariate short-term prediction of $PM_{2.5}$. Both studies are consistent and relevant to the purpose of this PoC in that they use RNNs (i.e. an LSTM and GRU architecture), take meteorological covariates as inputs, and consequently predict air pollution. Nonetheless, as much as O_3 and $PM_{2.5}$ are influential elements, a more complete air pollution and smog prediction requires consideration of a broader and, above all, combined set of air contaminants.

1.3 Contributions

Acknowledging the recent developments (Masood and Ahmad, 2021) and state-of-the-art, the LSTM and GRU establish themselves as the appropriate

choice for modelling the sequential series of components in ambient (polluted) air. This insight steers us towards contributing an attempt at getting further command of the air pollution problem through a PoC of smog modelling with LSTM and GRU models. Specifically, the combined modelling of contaminants NO_2, O_3, $PM_{2.5}$, and PM_{10} is considered.

Ultimately, this research addresses the question: "To what extent are models with the LSTM and GRU architecture capable of the multivariate timeseries forecasting of smog-related air components?" It is found that the LSTM and GRU can indeed accurately forecast smog-related air components, thus providing an effective method for modelling and forecasting pollutants.

2 Theoretical framework

The theoretical framework of this study explores air pollution, timeseries forecasting, and recurrent neural nets (RNNs), providing a comprehensive perspective on smog forecasting through deep learning.

2.1 Air pollution

Since the advent of the industrial revolution, increasing pollutant and trace gas concentrations have fundamentally changed the air quality (Fowler et al., 2020). A by-product of growing urban populations is spatial areas with a high concentration of polluting activities (Elsom, 2014), which can accumulate atmospheric pollution, forming "clouds of smog" of either photochemical or sulphurous kind (Haagen-Smit, 1952; Davis, 2002). The prevalence of smog will not subside anytime soon (as, for instance, estimates predict the world's urban population to grow from ±55% in 2018 to ±68% in 2050 (Ritchie and Roser, 2018)), nor will air pollution in general. Moreover, the same applies to the health effects.

2.1.1 Health implications

Besides the aforementioned general adverse health effects, air components are associated with adverse effects individually. Kampa and Castanas (2008) list, among others, gaseous pollutants and respirable Particulate Matter (PM) as two significant categories of pollutants. Out of these categories, the health effects of tropospheric ozone (O_3) (Amann, 2008; Malley et al., 2017), nitrogen dioxide (NO_2) (Eisner et al., 2010; Faustini et al., 2013), and PM with an aerodynamic diameter of $\leq 10\,\mu m$ (PM_{10}) and $\leq 2.5\,\mu m$ ($PM_{2.5}$) (Kampa and Castanas, 2008; Chen and Lippmann, 2009; Mašková et al., 2015) have been extensively studied and Appendix A discusses them in more detail.

Table 2.1: World Health Organization (WHO) air quality guidelines (AQGs) from 2021, where the AQG levels describe recommendations for maximum average pollutant concentrations on different timescales (WHO, 2021).

	Averaging time	*AQG level* ($\mu g\,m^{-3}$)
NO_2	24-hour	25
	Annual	10
O_3	8-hour	100
	Peak-season[a]	60
PM_{10}	24-hour	45
	Annual	15
$PM_{2.5}$	24-hour	15
	Annual	5

[a]Indicates the average daily maximum 8-hour mean in the consecutive months with the highest six-month running average

Following the impacts of the four, together with the acknowledgement that many components in the weather system affect air quality and its cleanliness, the scope of this research will focus specifically on the contaminants NO_2, O_3, $PM_{2.5}$, and PM_{10}.

Furthermore, this enumeration of negative consequences prompts consideration of the threshold at which concerns should arise. The WHO guidelines drawn in 2021 are presented in Table 2.1, with the maximum average air quality guidelines (AQGs), or exposure limits, enlisted for the four pollutants at issue. The guidelines, nevertheless, represent a predominantly practical target; concentrations fluctuate freely and dynamically and are already harmful at lower values (WHO, 2021).

2.1.2 Atmospheric interactions

The very reality of contaminant concentrations changing over time naturally focuses attention on the question of how these changes originate and evolve. Whereas the origins and sources of pollution are reasonably well understood (Vallero, 2014; Saxena and Naik, 2018), much is still unknown about its dynamics and how it evolves—hence, the subject of this research. Especially from a ML perspective, understanding the specific relationships between variables, or features, is critical for efficient learning (Li et al., 2017a).

How pollutants evolve is partly explicable by their interaction with their environment. As a result, a combined modelling of atmospheric variables can be justified. The following paragraphs briefly introduce the pollutants' interconnectivity and atmospheric interaction.

Foremost, the chemical interrelation of NO_2 and O_3. Nitrogen oxides (NO_x), a mixture of the colourless nitric oxide (NO) and reddish-brown, pungent

NO_2, are (mostly anthropogenically-generated) primary pollutants. When in the presence of certain volatile organic compounds (VOCs) or another initiator or catalyst, NO can oxidate into NO_2 (Atkinson, 2000). The photodecomposition of NO_2

$$NO_2 + hv(\lambda < 430\text{nm}) \rightarrow NO + O, \quad (2.1)$$

where h is Planck's constant and v is the light frequency ($\frac{c}{\lambda}$), will, in turn, initiate the sequential formation of secondary pollutant O_3

$$O + O_2 \xrightarrow{M} O_3, \quad (2.2)$$

where M is any third stabilising molecule for the excited intermediate formed by adding O and O_2 (Finlayson-Pitts and Pitts Jr, 1993). It follows that, by exclusive consideration and assumption of the presence of photons with adequate energy, NO_2 appears to affect the atmospheric quantity of O_3 positively.

When no (or less) solar energy is present (e.g., at night), NO_2 remains stable and reacts with O_3 to form nitrate radical (NO_3) (Finlayson-Pitts and Pitts Jr, 1993):

$$NO_2 + O_3 \rightarrow NO_3 + O_2, \quad (2.3)$$

lowering the concentrations of NO_2 and O_3 (at least for now). Therefore, owing to the chemical interdependence of O_3 and NO_x, the levels of O_3 and NO_2 are inextricably linked (Clapp, 2001). Besides, we already observe that one cannot dissociate pollutant concentrations from atmospheric (and cosmic) influences.

Continuing, PM is either emitted directly into the atmosphere (primary) or formed later (secondary) and is subject to air transport, cloud processing, and removal from the atmosphere (Seinfeld and Pandis, 2016). PM_{10} and $PM_{2.5}$ are interconnected as seen empirically (Velders et al., 2015) and naturally (Rhodes and Seville, 2024), given that their distinction is their size. With its relatively large size, PM itself experiences negligible chemical reactivity with the atmosphere compared to minor compounds such as NO_2 and O_3. Nonetheless, noting its susceptibility to transport, PM and other pollutants alike are responsive to airflow and dispersion in their ambient air environment—an environment amidst all meteorological influences, without yet considering factors such as geology and topology. Therefore, many, at least implicit, parameters are required to model PM and other air components reliably.

In short, NO_2, O_3, PM_{10}, and $PM_{2.5}$ are subject to influences from all dimensions and thus can be broadly modelled: for modelling pollution movements, pollution can be assumed to behave as air. Furthermore, pollutants are "internally" affected by each other and externally by the atmosphere, warranting a multivariate, integrated modelling approach.

2.2 Timeseries forecasting

A timeseries is a series of data points indexed along the time dimension. In the context of measurements, a timeseries is typically sampled at uniform intervals, constituting a discretised representation of a continuous temporal process. Timeseries can have a finite and infinite length, although they are finite for ML modelling purposes. Formally, finite length timeseries are sequences that start at time $n = 0$ and run until a maximum time $n_{\max}$:

$$(\mathbf{x}(n))_{n_{\max} \in [0,\infty]} = \mathbf{x}(0), \mathbf{x}(1), ..., \mathbf{x}(n_{\max}), \quad (2.4)$$

where $\mathbf{x}(n)$ is the data point at time n. The finite sequence can describe an interval of any dynamical system, thereby potentially communicating meaningful information about the temporal evolution of such system.

Timeseries analysis involves methods to extract statistics and characteristics from these sequences. Examples of uses are dynamical pattern generation (e.g. automatic music generation), pattern detection (e.g. nuclear fusion plasma instability detection), system modelling (e.g. aircraft behaviour modelling), and *timeseries forecasting*.

In timeseries forecasting, forecasts are made using timeseries analysis based on data comprising one or more timeseries. Forecasts can be univariate, i.e. with one involved data source (a weather sensor, for example,) or they can be multivariate, where forecasts of a given variable, or variables, depend, at least partly, on one or more additional predictor variables (Chatfield, 2000). Adding exogenous variables to the modelling problem can theoretically boost forecasting accuracy by revealing relevant, previously unknown information about the underlying process. On the other hand, handling the added complexity requires sufficient flexibility and power (see Sfetsos and Coonick (2000), for instance).

In data-driven, inductive air pollution forecasting, approaches thus far can be mainly categorised into statistical methods, shallow ML methods, and DL methods. Of the former, examples are non-linear regression (Baker and Foley, 2011), autoregressive integrated moving average (ARIMA) modelling (Prybutok et al., 2000; Vlachogianni et al., 2011), and single exponential smoothing (SES) (Gardner Jr, 2006). Exemplary is that, SES, with the inherent delay in its computational mechanism, cannot conform to irregular non-linear patterns(—which is also illustrated in Appendix B.4). Likewise, purely statistical methods tend to have serious shortcomings for forecasting the evolution of highly non-linear complex systems (Cheng et al., 2015). Shallow ML methods, exemplified by support vector machines (SVMs) (Chuentawat and Kan-ngan, 2018) and artificial neural networks (ANNs) (Elangasinghe et al., 2014), prove to be more accomodating to the complexity of air pollution data, as

reviewed in Masood and Ahmad (2021) and Cabaneros et al. (2019). Moreover, the authors highlight DL methods' ascendency to establishing themselves as the best-performing AI-based technique.

This, in conjunction with RNNs generally performing competitively within timeseries forecasting (Hewamalage et al., 2021), e.g. seen in the winning method of the M4 competition (Makridakis et al., 2020), consequently brings RNNs to the forefront for the application of air pollution timeseries forecasting.

2.3 Recurrent Neural Networks

Recurrent neural networks (RNNs) host cyclical connection pathways that, mathematically speaking, represent not functions but dynamical systems (DSs). RNNs have a network state $\mathbf{x}(n)$ allowing some earlier input $\mathbf{u}(n')$ to leave its traces on output $\mathbf{y}(n)$, and, therewith, the current state is influenced by past states and input. In practical terms, RNNs are tailored for sequential, fixed order data, such as timeseries. Illustratively, when data is fed non-sequentially, e.g. today's weather prior to yesterday's, the internal state becomes confounded.

The fact that RNNs represent, or are, DSs unlocks the possibility to model the many DSs inherent in reality. RNNs have proven highly adept for various applications where sequential data, often temporal in nature, is involved (Salehinejad et al., 2017; Hewamalage et al., 2021).

Formally, the transition of an RNN network state is given by the update equations:

$$\mathbf{x}(n) = \sigma(W\mathbf{x}(n-1) + W^{\text{in}}\mathbf{u}(n), \tag{2.5}$$

$$\mathbf{y}(n) = f(W^{\text{out}}\mathbf{x}(n)), \tag{2.6}$$

where $n = 0, 1, 2, ..., n_{\max}$ are the time steps, W is a matrix containing the connections weights, W^{in} and W^{out} contain the weights from/to the input/output neurons, σ is a sigmoid function, and f a function wrapping the readout $W^{\text{out}}\mathbf{x}(n)$ (Jaeger, 2023b). In particular the activation function σ, which introduces non-linearity to the evolution of the internal state (2.5), enables RNNs to capture (long-term) non-linear dependencies in the data.

RNNs train through backpropagation through time (BPTT). The idea of BPTT is to unfold the network through time, identically replicating the net for each timestep and rewiring the network-internal connections forward. A consequence is that with many timesteps, say of depth T, the error gradient is back-propagated through all those T unfolded layers. After many derivative passes, this causes the vanishing gradient problem (Hochreiter, 1998)—a well-known difficulty for effective RNN training (Pascanu et al., 2013). Since the gradient vanishes (or explodes) progressively, the probability of successfully encoding long-term dependencies decreases rapidly (Bengio et al., 1994).

Long short-term memory (LSTM) networks were introduced by Hochreiter, Schmidhuber, and Gers with the intention to solve this problem (Hochreiter, 1991; Hochreiter and Schmidhuber, 1997; Gers et al., 2000). They proposed a self-connected linear unit, the LSTM *memory cell*, with a constant error flow: in the absence of new input or error signals to the cell, the local error backflow remains constant, neither growing nor decaying (Gers et al., 2000). Thus, with the LSTM, the gradient is independent of T.

Memory cells possess a gating mechanism facilitating this. While many variations to this mechanism exist (Greff et al., 2017), it generally contains the following elements. The foundation is formed by the cell state and hidden state. Its input and output are regulated by the input gate and output gate, respectively, and a linear self-loop is controlled by the forget gate. All the gating units have a sigmoid non-linearity, while the input unit can have any squashing non-linearity (Goodfellow et al., 2016). These elements, along with the introduced non-linearities, allow for adaptive information flow modulation, enabling selective processing and long-term retention of non-linear patterns.

A more recently proposed recurrent unit is the gated recurrent unit (GRU) by Cho et al. (2014). The GRU uses a similar approach to solving the vanishing gradient problem but simpler. It contains only two gates, the reset gate and update gate, making it easier to compute (and implement). The former controls the degree to which the previous hidden state influences the current, and the latter combines the LSTM input and forget gate into one. Its performance has shown to be on par with the LSTM, and, in some cases, can outperform it in terms of convergence in CPU time and in terms of parameter updates and generalisation (Chung et al., 2014).

Since their introduction, gated RNNs achieved most RNN breakthroughs (Yu et al., 2019). As seen in Section 1.2, the gating mechanism also proves itself in air pollution-related applications (Freeman et al., 2018; Tao et al., 2019). Still, these studies predicted one contaminant only, while LSTMs are proven to be adequate for multivariate data (Che et al., 2018).

At a higher level, beyond the configurations of individual gates or neurons, is where discussion of such multivariate data can begin. A way to strike a balance between a multivariate forecast and an individual one is through non-homogeneous hierarchical neural circuits and architectures, also called hierarchical neural nets (HNNs). HNNs consist of a number of loosely-coupled subnets, arranged in layers, where each subnet is intended to capture

specific aspects of the input data (Mavrovouniotis and Chang, 1992). In HNNs, a priori knowledge can be embodied in the neuronal arrangement, thereby steering the model in a preferred direction. Such a balance lends itself particularly well to air pollution data, as will be discussed in Section 3.3, continuing per the methods section discussed next.

3 Methods

In this study, we model pollutants using meteorological covariates along a one-dimensional spatial trajectory, or, in other words, we use pollutant and meteorological data at point A to predict pollutants at point B. The ensuing sections describe and inspect the data employed, explain the preprocessing, introduce the models, outline the training process, and present the used evaluations.

3.1 Data

The proposed forecasting experiment uses hourly-sampled data from 2016 to 2023 (RIVM, 2024; KNMI, 2024), which is available under an initiative of the Dutch government and the Dutch national meteorological service, the Royal Netherlands Meteorological Institute (KNMI). The data is accreditated under NEN-EN-ISO/IEC 17025 standards and is technically and substantively validated (and possibly rejected) before release (KNMI, 2023a).

Where traditional weather prediction methods, such as NWP, get data from ground sensors, buoys, air balloons, satellites, weather radars, and even commercial planes—to name a few—the data for this experiment is measured solely by automatic weather stations (AWSs) that make synoptic readings (KNMI, 2023b). By synchronising the readings with respect to time, this approach makes the data consistent and adequate for spatiotemporal air quality monitoring and analysis. Moreover, uniformly-discretised data is particularly convenient for timeseries forecasting.

3.1.1 Spatiotemporal context

This experiment involves forecasting with data from two locations[1], a source location (A) and a target location (B). The source location is in Utrecht, the Netherlands, and here, pollutant data is combined with meteorologically related covariates to forecast pollutant data at the relatively northwestern target location in Breukelen. Their relative positions are best illustrated visually; see Figure 3.1.

Given the distinct locations, the key assumption is that environmental conditions in the different locations are related—both for pooling the predictive data at A and for traversing the approx. 15 km between A and B. Besides the non-intuitive scale of weather phenomena, geographical and empirical arguments can underpin this premise. First, geographically, the Netherlands has low elevation and flat topography, thus not confounding the (modelled) airflows; second, analyses, such as Karaca et al. (2009) and Glavas and Sazakli (2011), show long-range, international-scale PM, NO_x, and O_3 travel and influences, thereby illustrating the scales. Given the large, open scales, however, a flip side is that a sample of reality from merely one location, i.e. the sensors at A, cannot, explicitly nor implicitly, capture all required information for a complete forecast (at B). More data from more locations, e.g. from sensors surrounding B, could theoretically narrow this disparity, but this is beyond the scope of this PoC.

Figure 3.1: Utrecht area with markers indicating the AWS locations. Located at the bottom right, in De Bilt, is the source meteorological sensor (52°06'06.4"N, 5°10'42.1"E); slightly to its west, in Utrecht, the source pollutant sensor (52°06'18.1"N, 5°07'28.1"E). Situated approx. 15 km north-west of those, near Breukelen, is the target sensor (52°12'05.5"N, 4°59'14.8"E).

As for the more immediate surroundings of A and B, they are located around a city with relatively fruitful surrounding nature and little industry. Pollutant sensor A, however, is located in a densely populated district right next to a four-lane access road, and pollutant sensor B is located right next to the ten-lane A2-motorway (see Figure 3.1), a crowded connector between Utrecht and Amsterdam. For both locations, this inevitably means that anthropogenic patterns, such as morning and evening commutes, will show in the data.

In addition to where to sample, there is a choice of when. Recent years proved relatively turbulent with accelerating climate change and a pandemic, implying that atmospheric distributions may shift from year to year. An uncontested example is

[1]The two-site approach was adopted in anticipation of extending the scope of this study with physics-informed ML, which is discussed in Section 6.1 on future research.

Table 3.1: Predictive variables and initially considered meteorological variables (in alphabetic order). Some units are multiplied by 0.1 for simplification without losing significance.

Variable	*Unit*
Nitrogen dioxide (NO_2)	$\mu g\,m^{-3}$
Ozone (O_3)	$\mu g\,m^{-3}$
PM $\leq 10\,\mu m$ (PM_{10})	$\mu g\,m^{-3}$
PM $\leq 2.5\,\mu m$ ($PM_{2.5}$)	$\mu g\,m^{-3}$
Air pressure (AP)	0.1 hPa
Dew point temperature (DPT)	0.1 °C
Global radiation (GR)	$J\,cm^{-2}$
Maximum wind gust (MWG)	$0.1\,m\,s^{-1}$
Mean wind direction (MWD)	$0 - 360°$
Mean wind speed (MWS)	$0.1\,m\,s^{-1}$
Precipitation amount (PA)	0.1 mm
Precipitation duration (PD)	0.1 h
Sunshine duration (SD)	0.1 h
Temperature (T)	0.1 °C

the COVID-19 outbreak in the Spring of 2020, where the initial severe lockdown measures led to a significant reduction in Dutch NO_2, O_3, PM_{10}, and $PM_{2.5}$ concentrations (Velders et al., 2021). Another more regular anomaly is New Year's, as seen in Appendix B.4, Figure B.2. Accordingly, any such factor or outlier should be considered in data selection. As such, New Year's data is preemptively discarded, as are, unavoidably, data around the pandemic outbreak, including the other years for seasonal uniformity: only August to December will be considered. The following section expands on this narrowing and selection of the dataset by inspection.

3.1.2 Inspection

Table 3.1 details the predictive variables used in this experiment, along with the initially considered meteorological parameters. The rationale and relevance of the meteorological parameters in the context of pollution prediction are discussed in Appendix B.1, coupled with an inspection. The listed meteorological variables were—except some individual disputes—available for all years.

Unfortunately, this did not hold for the pollutant variables, which were either wholly unmeasured or rejected for some years. Bijma (2012) clarifies the missing data: single values can be rejected in case of equipment failure and certain sensor-unfavourable weather conditions such as condensation in PM-sensors, and, crucially, whole years can be rejected during the yearly validation prior to release. As a result, O_3 showed absent for 2016 and 2019—and these years, in turn, are excluded.

To get an impression of the to-be-forecasted data, the four predictive variables are plotted over one week in Figure 3.2. Foremost, referencing Table 2.1, we can conclude that AGQs are not always met, even here. Furthermore, noteworthy atmospheric interactions can be discerned in the plot.

It is apparent that PM concentrations are at times negative (and can be even more negative, see Table 3.2). This can be explained by measurement inaccuracies depending on varying environmental factors (e.g. condensation in PM equipment), internal variations in the measurement apparatus, and variations in the calibrators (KNMI, 2023a). Thus, all readings are equally correct and negative values are left uncorrected to avoid a positive bias.

Secondly, a pattern shows for NO_2 and O_3. The conditions and interactions discussed in Sec-

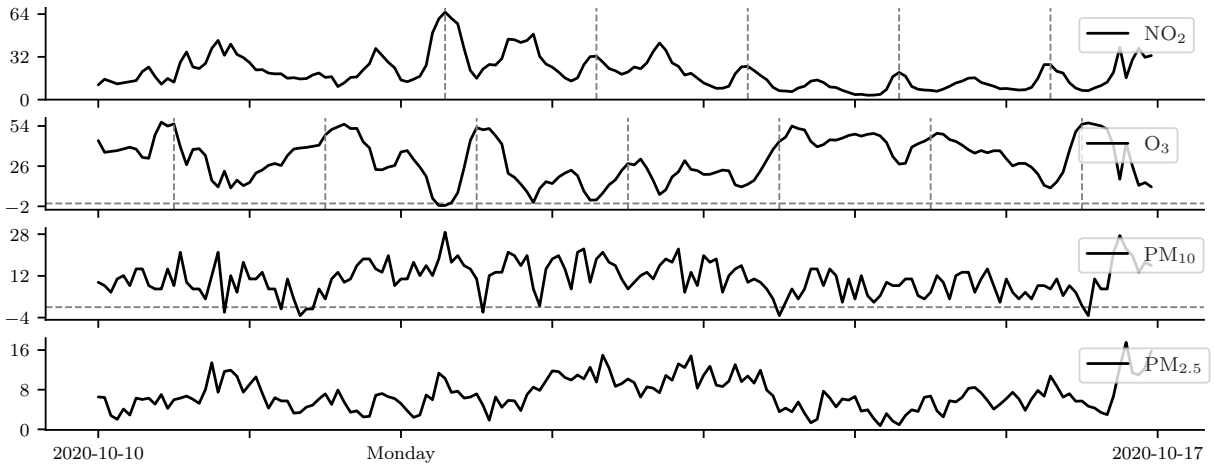


Figure 3.2: The pollutants in $\mu g\,m^{-3}$ as a function of time, for a week in October in Utrecht. Time-axis markers indicate start-of-day, midnight. Crossings of the horizontal dashed lines indicate negative observations. The PMs do not exhibit obvious patterns. O_3 shows a diurnal pattern with midday peaks. Less obvious and apparent every weekday, starting on Monday, are slight NO_2-morning-peaks preceding the O_3-midday-peaks, both indicated by vertical dashed lines.

tion 2.1.2 and 3.1.1 are reflected in the graph on working days: NO_2 concentrations rise with morning commute (Shi et al., 2018), photodisassociation occurs, and the formation of O_3 is initiated (Finlayson-Pitts and Pitts Jr, 1993). In this sense, although data points such as traffic volume, industry volume, and irregularities like Saharan dust are not explicitly included in the data, they can be implicitly reflected nonetheless, serving the practical purpose and, in parallel, the simplicity of the model.

Aside from a visual representation of the data, an intrinsically essential part of modelling is the data's distributions. In a training-validation-test split, each of the distributions should be reasonably mutually consistent for a sound test of a modelling technique (see, e.g., Duan et al. (2023)). A classic indicator of this is the stationarity of the data, and a temporal shift in case of violation thereof.

Where values in nature will naturally evolve around equilibria (think of diffusion or entropy), this is not guaranteed over shorter timescales. With recent accelerating climate change (Lee et al., 2023), it is a fair bet that year-to-year distributions will undergo a temporal shift. Kernel density estimates (KDEs) for NO_2, O_3, PM_{10}, and $PM_{2.5}$ over the years 2017, 2018, 2020, 2021, 2022, and 2023 are presented Figure 3.3 to investigate this further.

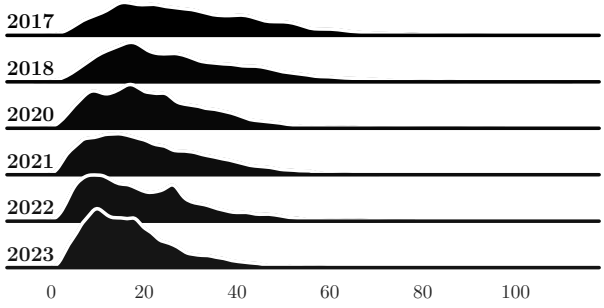
Figure 3.3a shows that year-on-year, NO_2 exhibits an on-average negative trend, i.e. increasing air quality. For O_3, in Figure 3.3b, the contrary can be seen: the values are increasing on average. The PMs in Figures 3.3c and 3.3d appear reasonably stable. The most apparent temporal shifts—those for NO_2 and O_3—are not readily explained. For the purpose of modelling, it remains true that it is wise to capture varied distributions in the training validation, and testing set.

From this, one may also conclude that the universal principle of more data being favourable certainly holds in this matter as well: a data sample of merely a few years would make the dissimilarities too significant and hamper modelling. Ideally, training, validation, and testing sets should cover multiple (time)points to mitigate disparities caused by shifts, and thereby improve modelling potential. Now, we will look at how the inspected data is prepared for modelling.

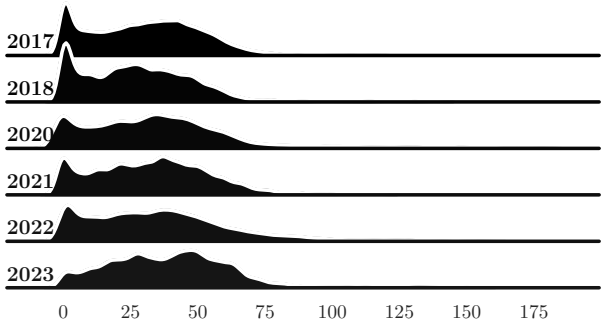
3.2 Preprocessing

Preprocessing starts with tidying the raw data, followed by a train-validation-test split, feature engineering, normalisation, and ends with generating (input, output)-pairs.

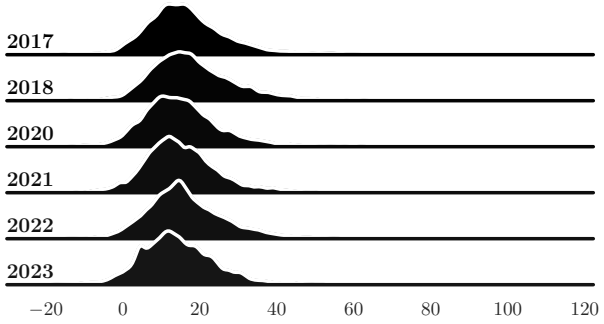
Firstly, the raw data was cleaned to make it utilisable, for example by solving erronous (split) rows and columns, converting encodings, extracting



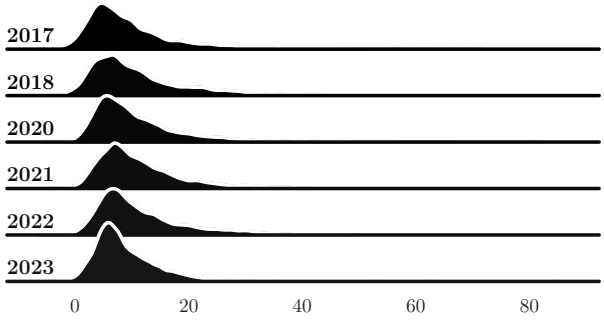
(a) KDEs of NO_2-concentrations. Concentrations at the higher end show a downward trend, where the low 2023 peak is exemplary of the overall downward trend.



(b) KDEs of O_3-concentrations. The mean seems to drift rightward, with values distinctly peaking in 2023.



(c) KDEs of PM_{10}-concentrations. Its values remain approximately stable over the years and seem normally distributed.



(d) KDEs of $PM_{2.5}$-concentrations. Values seem stable over time, although a peak is visible in 2023.

Figure 3.3: Ridge kernel density estimation (KDE) plots for the concentrations ($\mu g\, m^{-3}$) of, from top to bottom, NO_2, O_3, PM_{10} and $PM_{2.5}$ in Utrecht from August 1st to December 30th for the years 2017 to 2023, excluding 2019.

metadata, and the exclusion of data disqualified in previous the section (i.e., the years 2016 and 2019, and New Year's days). Next, there were missing values (Table B.2). These were assumed to be missing completely at random (MCAR) and were imputed by linear interpolation

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}, \tag{3.1}$$

where x and y lie in the intervals (x_0, x_1) and (y_0, y_1), respectively.

Secondly, the tidy data is split into a training, validation, and testing set. Granting the heterogeneous nature of the data from year to year, but also the fact that forecasting the future using information from the future is fallacious, a sampling balance has to be struck. This is where each year's five-month cut/interval comes in handy: the validation and testing sets can be safely sampled from multiple years, resulting in the following training/validation/testing distribution.

Training set. August 1st up to and including December 30th of 2017, 2018, and 2020, and August 1st up to and including November 18th of 2021 and 2022.

Validation set. November 19th up to and including December 9th of 2021 and 2022, and August 1st up to including October 2nd of 2023.

Testing set. December 9th up to and including December 30th of 2021 and 2022, and October 3rd up to and including December 4th of 2023.

By percentage: [76.3%/11.9%/11.9%]. By taking the validation and testing set from multiple years, and by including earlier months from 2023, a balanced dataset is created, taking into account both month-on-month and year-on-year changes. Now, the validation set is stashed away until training, and the testing set until evaluation.

Thirdly, the newly acquired training set undergoes feature selection. As described in Hall (1999), good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other. Thus, to assess the features—the pollutants and meteorological parameters listed in Table 3.1—their intercorrelations are assessed and compared to a treshold r_{th} using the absolute value Pearson correlation coefficient r_{xy}:

$$r_{xy} = \left| \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \right|, \tag{3.2}$$

where, given paired data $(x_i, y_i)_{i=1}^{n}$, n is the sample size, and $\bar{x}$, $\bar{y}$ are their sample means. It must be noted, however, that the calculation assumes linear relationships, heteroskedasticity, and a Gaussian distribution—all of which are not necessarily accounted for in the present data (see Figure 3.3, for instance). In the future, alternate methods could be investigated. For now, the correlations are plotted in Figure 3.4(—for their exact numerical coefficients, see Table B.1).

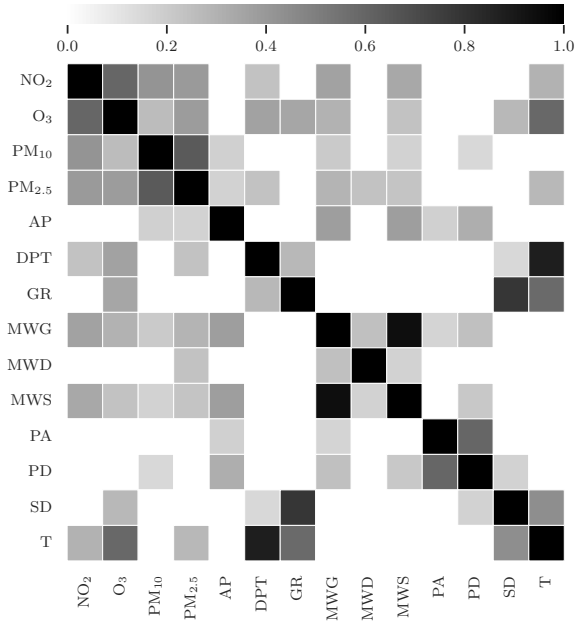


Figure 3.4: Coefficient matrix for the initially considered features. A threshold $\mathbf{r_{th}}$ for the absolute Pearson coefficient is set at $\mathbf{r_{th} = 0.15}$. When not met, the entry remains white.

As for the pollutants, the matrix expectedly reveals interconnections both amid the gaseous pollutants and amid the PMs, as suggested in Section 2.1, as well as between both groups—which is not surprising since they mostly emanate from the same sources. Also, several meteorological variables show strong interrelations, such as $r_{\text{PA,PD}} = 0.60$, $r_{\text{SD,GR}} = 0.79$, $r_{\text{MWS,MWG}} = 0.94$. The features precipitation amount and duration, besides being mutually correlated, show correlations that fail to meet $r_{\text{th}} = 0.15$, explicable by the fact that, in the data, it is dry 85.94% of the time (see Figure B.1 for an impression), and so the values are often zero, distorting the calculation. Nevertheless, these features are excluded. The second pair, sunshine duration and global radiation, as expected also show strong correlation. Since both generally do not correlate strongly with the pollutants, only sunshine duration was chosen to remain because of its linkage with O_3. At last, in view of MWG's high intercorrelation with MWS, $r_{\text{MWS,MWG}} = 0.94$, the former is excluded.

Further, there are mean wind direction and air pressure. The naturally noisy mean wind direction is poorly correlated but can be kept thanks to denoising applied through the 24-hour moving average. Finally, air pressure is also lowly correlated but stays with an eye on an upcoming project extension; see Section 6.1. In conclusion, precipi-

Table 3.2: Minimum and maximum pollutant concentrations in the training set, used for min-max normalisation. Minima reach below zero due to measuring uncertainties (Bijma, 2012).

	Minimum ($\mu g\,m^{-3}$)	*Maximum* ($\mu g\,m^{-3}$)
NO_2	-0.28	107.45
O_3	-2.22	180.51
PM_{10}	-19.89	379.47
$PM_{2.5}$	-4.93	79.71

tation amount and duration, global radiation, and maximum wind gust are discarded in the feature selection. These adjustments are incorporated into the datasets, and 10 suitable features are available.

Fourth, normalisation. Normalising promotes generalisation, stabilises gradients and the learning process, and can produce faster convergence (Ioffe and Szegedy, 2015). The selected features are normalised to a range of $[0, 1]$ with min-maxscaling

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{3.3}$$

where x_{min} and x_{max} are the minimum and maximum value for each feature in the training set. Table 3.2 shows the extremes. Interestingly, the values extend to (well) below zero, and also above those observable in Figure 3.3. Yet, as mentioned earlier in Section 3.1, the values have already been checked for outliers prior to publishing.

Fiftly, and lastly in preparing the data for the models, pair generation. In order to obtain static input-output pairs from a discretised hourly-sampled temporal training sequence, one segments the input timeseries $(\mathbf{u}(n))_{n\in[0,N_u]}$ for Utrecht, and input timeseries $(\mathbf{y}(n))_{n\in[0,N_y]}$ for Breukelen with $N_u = N_y$, into sliding windows of length $l_{\text{in}} = 72$ and $h = 24$, respectively, obtaining input-output pairs $(\mathbf{u}_i, \mathbf{y}_i)_{i=1,...,P}$ consisting of input

$$\mathbf{u}_i = (\mathbf{u}(n_i), ..., (\mathbf{u}(n_i + l_{\text{in}})), \tag{3.4}$$

and output

$$\mathbf{y}_i = (\mathbf{y}(n_i + \delta), ..., (\mathbf{y}(n_i + \delta + h)), \tag{3.5}$$

where n_i represents the starting index of the i-th pair, P denotes the number of pairs as defined by $P = \left\lfloor \frac{N_u+1}{\Delta n} \right\rfloor$ with sampling step size Δn, and δ is defined as $\delta = l_{in} - 24 + 1$, meaning $\mathbf{y}_i$'s output is considered from the 48th hour on, plus a 1-hour window for the spatial prediction from $\mathbf{u}_i$ to $\mathbf{y}_i$. To expand on the latter, and as seen in (2.5) and (2.6), RNNs process values one-by-one, which for this case means for l_{in} iterations—δ, however, selects only the last 24 readouts for predicting and, thus, learning (facilitated by the loss function (3.6) discussed in Section 3.4). Step size Δn is set at $\Delta n = 24$ for computational efficiency, and because trial-and-error testing showed no or minor upside to a smaller Δn.

Essentially, this means that for each pair, an hour of pollutant concentrations at B will be predicted 24 times in sequence, with the preceding hours of A available as the ground for prediction. Thus, the data sets the models up to learn to model the pollutants using their covariates for the spatial prediction task from Utrecht to Breukelen.

Prior to generation of $(\mathbf{u}, \mathbf{y})$-pairs, the training set contained 676 days of data, which after sampling yielded $P = 656$ pairs (less than 676 due to $\Delta n = 24$ and boundary conditions), equivalent to 535,296 data points in each forward pass of the models during training—models described in the following section. (For more data statistics, see Appendix B.5).

3.3 Model architecture

The multivariate one-dimensional forecasting task of smog clouds, i.e., modelling the four pollutants from Utrecht to Breukelen, is taken on using six models: an ordinary multi-layer perceptron (MLP), a hierarchical MLP (HMLP), an LSTM and GRU, and, as main contenders, a hierarchical LSTM (HLSTM) and GRU (HGRU). This section will outline the modelling types and set-ups, followed by their hyperparameter optimization procedure.

3.3.1 Types

To begin with, as the implementation of the somewhat non-traditional hierarchical neural nets (HNNs) in high-level Python libraries akin to TensorFlow (Abadi et al., 2016) is not necessarily accessible (Fontenot et al., 2022), the models are implemented with PyTorch (Paszke et al., 2019).

Then, as touched upon in Section 2.3, the MLP models approximate not DSs but functions. Where RNNs have a state $\mathbf{x}(n)$ allowing some earlier input $\mathbf{u}(n')$ to leave its traces on output $\mathbf{y}(n)$, MLPs learn to approximate a (nonlinear) function $f : \mathbb{R}^{L^0} \to \mathbb{R}^{L^k}$, where L^0 and L^k represent the neurons in the input- and output layer, and lack an explicit mechanism for retaining sequences over extended periods. In practice, they cannot utilise the sequence-spanning BPTT; they propagate errors solely through the network. Hence, the MLP and HMLP are less suited for this task than RNNs and serve as benchmarks.

In terms of their specific architecture, the input and output layer are of size $L^0 = 10$ (ten features) and $L^k = 4$ (four predictive variables). Unidirectional layers knit these together. For the MLP, these are standard fully-connected layers. Its counterpart, the HMLP, is of the type of hierarchical

models—a term introduced in Section 2.3 which describes non-homogeneous, modular neural circuits. HNNs can, depending on the task, perform multi-task learning (MTL), a method whose principle goal is to improve generalisation performance (Caruana, 1997).

Furthermore, with HNNs, the hierarchical organisational structure is in hands of the model designer and offers an opening for a priori knowledge to be embodied in the neuronal arrangement as inductive bias or regularising factor, guiding the model in a preferred direction. In the context of this study, we aim to predict the four pollutants, each of which can be regarded as a distinct subtask. Recognising both the intercorrelations of the pollutants (as depicted, for example, in Figure 3.4) and the fact that they all live a life of their own, it seems reasonable to mirror this reality in a model's architecture. To achieve this, we employ one shared layer to establish shared representation and subsequently partition the network flow into a modular branch per subtask to reduce the interference between tasks. This design, including this nuanced regularising factor, confers HNNs a hypothetical advantage over fully-connected nets, which neglect an explicit internal-external balance.

Next are the RNNs. The RNNs use the PyTorch implementation of LSTM and GRU memory cells introduced in Section 2.3. The fully-connected RNNs are similar to the MLP, except for their gating mechanisms and recurrent synaptic connections, and vice versa for the hierarchical RNNs in relation to the HMLP.

Following up on identifying model types, hyperparameters reveal in more detail how these types are shaped into complete architectures. The following sections discuss how they are established.

3.3.2 Hyperparameter optimization

Hyperparameters can be used to control the behaviour of a learning algorithm and are not adapted by the algorithm itself (Goodfellow et al., 2016). For the DL models at hand, examples are the number of hidden layers and hidden units, the learning rate, choice of optimizer, and the regularisation term. An overview of the used hyperparameters per model can be found in Appendix D—this section will chiefly explore the methodology behind their selection.

To determine their values, a hyperparameter search procedure consisting of a grid search and cross-validation (CV) schemes is used. This procedure aims to find a hyperparameter configuration c with a minimised loss, while also testing c's generalisation capabilities. The loss is calculated on distinct validation sets generated by CV from all available training data, i.e. a concatenation of the training and validation set, to test this generalisation performance and prevent overfitting. Nested within the hyperparameter search and within the CV scheme, is the models' training algorithm, which, together with the loss, is specified in the next subsection, Section 3.4.

Table 3.3: List of the hyperparameters included in the grid search. They include: number of hidden units and hidden layers, learning rate for the fully-connected nets, learning rates for the shared and branched parts of the modular nets, and the weight decay. Values per model are presented in Appendix D, Table D.1.

Hyperparameter	*Symbol*
Hidden layers	k
Hidden units	L^{κ}
Learning rate	μ
Learning rate, shared	μ_{shared}
Learning rate, branches	μ_{branch}
Weight decay	λ

Continuing, the traditional grid search was used because of its ease suiting this PoC; it essentially does a brute-force search through the parameter space H. Here, H is defined as the Cartesian product of the finite sets S containing possible values for each parameter. Because H grows exponentially, a large S is not feasible, and smaller S are already computationally expensive. Hence, as measures, some initial trial runs were executed to get a feel of which options to include and the many models were computed on an HPC cluster.

Then, CV is run for each c, where c is a unique configuration within H. For the stateless MLPs, regular k-fold cross-validation, with $k = 5$, is used—with the perk of maximal data usage. RNNs, conversely, do have a state and allow memory trace of past sequences, as aforementioned in Section 2.3. A variation of k-fold CV, called sliding window CV, accommodates this: it samples training and validation sets—with, in contrast to pair generation, superposition of intervals—using a sliding window approach, thus not allowing validation of trained models with out-of-sample data directly from the past. One drawback of this scheme is that not all data is available during each iteration. Another scheme, expanding-window CV, would partly solve this, but was found biased due to distribution changes over time (see Figure 3.3) and consequently abandoned. And so, for every fold, a model is trained and validated using different data and the average validation risk is calculated.

With these schemes, grid search with k-fold CV for the MLPs and sliding-window CV for the RNNs, values for the hyperparameters listed in Table 3.3 were determined—forging the model types into architectures. (Appendix D presents complete model architecture summaries.)

3.4 Training

In this section, we explain the training procedure used during hyperparameter optimization and the final training itself by defining the optimization goal and method, followed by some anti-overfitting measures. The final models are trained using the training and validation set created in Section 3.2.

To approximate a model m_θ parametrised by tuneable parameters collected in a vector θ, given a search space $\theta \in \Theta$ of target models Θ within the same architecture; training pairs $(\mathbf{u}_i, \mathbf{y}_i)_{i=1,\dots,P}$; and a loss function L defined as

$$L = \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\theta^2, \tag{3.6}$$

where MSE denotes mean squared error, n denotes sample amount, y the ground truth, and $\hat{y}$ the prediction—one has to solve the optimization problem

$$\theta_{\text{opt}} = \underset{\theta\in\Theta}{\text{argmin}}\frac{1}{P}\sum_{i=1,\dots,P} L(m_\theta(x_i), y_i), \tag{3.7}$$

where θ_{opt} denotes a model with a minimised empirical risk (Jaeger, 2023a,b). MSE punishes extreme values quadratically more, suiting the context of air pollution where extremes are of greater concern.

With an initial model $\theta^{(0)}$, initialised by the PyTorch-default Kaiming initialisation (He et al., 2015), optimization of $\theta^{(n)}$ is performed by the *Adam* (ADAptive Momentum) optimizer (Kingma and Ba, 2014). Adam differentiates itself from, e.g., stochastic gradient descent (SGD) by using momentum (Sutskever et al., 2013) and adaptive learning rates per parameter while only requiring first-order gradients and little memory (Kingma and Ba, 2014). This study uses its implementation in PyTorch. An implemented, assisting add-on is a scheduler: it reduces the learning rate by a factor of 0.1 when the validation loss stagnates for a set number of epochs (defined per model in Table D.2).

Adam does its work everytime a batch B of 16 $(\mathbf{u}, \mathbf{y})$-pairs is passed. $|B| = 16$ was plainly adopted from Masters and Luschi (2018), who found smaller batch sizes ($2 \leq |B| \leq 32$) to provide benefits in terms of convergence stability and overall test performance for a given number of epochs. Batches are randomly sampled (while in sequence order) from the available pairs, introducing stochasticity (and efficiency over, e.g., one-by-one calculation). Internally, this adds the batch dimension to the pairs, creating the tensors $[|B|, l_{in}, L^0]$ for $\mathbf{u}$ and $[|B|, h, L^k]$ for $\mathbf{y}$. When $\mathbf{u}$ is fed, the models spit out forecasts $\mathbf{y}'$ in the form of such tensor, which is subsequently compared to the ground truth $\mathbf{y}$ yielding the loss withwhich θ can be updated.

Updating θ, however, proceeds quite differently for the two main types of models. Whilst for the fully-connected models, this proceeds as usual with one optimizer updating θ, the modular models require a different approach. As they essentially consist of multiple core components (one shared layer, four branches) with different search spaces and convergence qualities, the process capitalises on this: all five components have their own optimizer and matched scheduler. In addition, they have two separate (initial) learning rates, as seen in Table 3.3 and Table D.1. Distributing the learning tasks helps each model part stably reach an optimum.

Lighting this in terms of implementation, the shared and branched parts do epochs in turns, seeing all the batches separately while the other is *frozen*. Frozen, as in, the parameters cannot update but can infer. A con here is efficiency: the batches are passed through the model once more for every epoch.

When zooming out and looking at when learning should finalise, early stopping comes in: it finishes training when for some number of epochs (defined in Table D.2) the validation loss does not decrease by $\geq$ 1e−5 (not zero to prevent endless minor updates). Another anti-overfitting measure, or regulariser, is the $L2$-norm added to (3.6). As effect, larger weights are penalised and smaller weights are encouraged, preventing some set of weights dominating the model.

In summary, the training process seeks to find an optimal set of model weights θ_{opt}, and to regularise, the learning process early stops, the batches introduce stochasticity, the regularisation term balances weight values, and the hierarchical nets incorporate MTL. With these regularisation steps, the training procedure should yield models fulfilling the ultimate objective of generalisation, tested in next section.

3.5 Evaluation

For evaluation of the models, the held-out test set is used. The test set is unseen by the models before evaluation to properly assess their generalisation capabilities. The predictions were first post-processed using inverse minmax-scaling and the extremes in Table 3.2, sampled in batches without shuffling to eliminate any randomness, and then evaluated using the root mean squared error (RMSE) and symmetric mean percentage error (sMAPE) metric, which both serve a different interpretation of model performance. In addition, the inference speed of each model is evaluated, as this is one of the unique advantages of data-driven methods over first-principle methods like NWP.

The RMSE, defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.8}$$

provides a measure of the average magnitude of the error with, due to its squaring operation, larger errors getting penalised more. This fits the context of smog modelling, where higher values are especially harmful.

The other metric, the sMAPE:

$$\text{sMAPE} = \frac{2}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)} \times 100, \qquad (3.9)$$

is an accuracy measure based on percentage (or relative) errors, providing a scale-independent, well-interpretable metric. The sMAPE complements the RMSE by taking into account the individual and different distributions of the pollutants—which can be visually seen in Figure 3.3. Therefore, the metric allows for a fair relative comparison of the models' performance.

Finally, it is worth emphasising that, generally speaking, the RMSE serves a practical purpose because it tells about the deviation in $\mu g\,m^{-3}$ and has a quadratically progressive penalty. The sMAPE mainly fulfils a "scientific" purpose due to the possibility of comparing models, though the two metrics are not mutually exclusive. This idea acts as a guide to interpret the results meaningfully.

4 Results

Following training, the models are evaluated on out-of-sample data. It is found that the models provide an effective method for the modelling and forecasting of the pollutants. Quantitative results by RMSE and sMAPE are listed in Table 4.1.

Considering the subtask-specific lowest sMAPE values, NO_2 is predicted most accurately. Following NO_2 is O_3, then $PM_{2.5}$, and the models were least successful in predicting PM_{10}. (Discussion of possible explanations for this is held in next section, Section 5.) Nonetheless, the lowest sMAPEs, as well as the RMSEs—which are primarily generated by the HGRU—confirm the models' suitability for forecasting the pollutants at B using data at A. Meanwhile, the models also differed in performance.

Consistent with expectation, the MLPs achieve the highest errors on average. Measured by RMSE, the non-hierarchical fully-connected RNNs perform predominantly better than the MLPs, but also utilise many parameters to do so. Measured by sMAPE, they do too, despite HMLP's sMAPE ($M = 46.274, SD = 45.344$) being slightly inferior to the LSTM's ($M = 46.321, SD = 46.515$): a paired t-test with $\alpha < .05$ suggests there is no sufficient evidence to reject the null hypothesis of no difference, $t(8927) = 1.011, p = 3.12\text{e}{-1}$. Worth mentioning is that this is the only metric-model-model combination found not to be statistically significant, see Appendix E, Table E.1, E.2, E.3.

Furthermore, the GRU yields the lowest errors of the non-hierarchical RNN models. The HLSTM ranks second, and, as per RMSE and sMAPE, the HGRU performs best, establishing the hierarchical models as the top performers. A paired t-test confirms the HGRU's ($M_{\text{RMSE}} = 5.468, SD_{\text{RMSE}} = 4.906, M_{\text{sMAPE}} = 44.519, SD_{\text{sMAPE}} = 44.519$) significant predictive ability on the testing set over the HLSTM ($M_{\text{RMSE}} = 5.633, SD_{\text{RMSE}} = 4.935, M_{\text{sMAPE}} = 44.981, SD_{\text{sMAPE}} = 45.850$) both by RMSE ($t(8927) = -5.922, p = 3.30\text{e}{-9}$) and sMAPE ($t(8927) = -2.855, p = 4.32\text{e}{-3}$), as well as on the other models (see Appendix E). Moreover, for all individual pollutant subtasks by RMSE, and most by sMAPE, the HGRU exhibits the highest predictive precision, where it is only surpassed repeatedly with the sMAPE of the $PM_{2.5}$-subtask.

A visual comparison of predicted and observed concentrations for all pollutants is shown in Figure 4.1(—in Appendix E, Figure E.9, E.10, E.11, and E.12, non-combined depictions are presented). Interestingly, all models show a negative bias. Similarly here, the HGRU shows the least deviation from the diagonal, reflecting a closer agreement between forecast and ground truth. Its superiority is seen, for example, in the upper-right of the subplots, among the higher-concentration data points. Here, the MLP (taken for contrast) consistently underestimates all, while the HGRU is tighter on the diagonal, showing its flexibility. Above all, however, is its relative performance in Figure E.11.

A most natural representation of the HGRU's forecasts is with a lineplot, shown in Figure 4.2. Consistent with the numerical interpretation of the RMSE and sMAPE, the patterns of NO_2 and O_3 seem to be most closely captured. The PMs show more short-term fluctuations, which are infrequently caught. This proves the most challenging with PM_{10}. Altogether, it can be stated that the HGRU is well equipped to use data at A for forecasting at B. Additional forecasts are presented in Appendix E: a 24-hour combined forecast (Figure E.1), two-week combined (Figure E.2), and individual (Figure E.3, E.4, E.5, E.6, E.7, E.8).

Furthermore, in terms of efficiency, the inference speed t_{inf} of the models, as also seen in Table 4.1, shows that efficiency is high: a 24-hour prediction is generated with negligible delay on a relatively inefficient processor. Counterintuitively, the model with the most parameters is the quickest, though the margins are small. As also discussed in Pathak et al. (2022) and Section 1.2, the speeds, apart from the initial training cost, highlight the (operational) advantage of DL models over traditional first-principle methods: they are orders of magnitude faster and more efficient. Last to note on efficiency is that the best-performing models, the HLSTM and HGRU, require significantly fewer pa-

Table 4.1: Results of each model, evaluated and compared on performance (RMSE and sMAPE) and efficiency (inference speed and parameter count). The error metrics are calculated per pollutant and combined, with the lowest error bolded. Inference speed $t_{\mathbf{inf}}$ is the time for one inference of one 24-hour lead time prediction (processed on an Intel Core i7-8565U CPU, 8GB RAM, 64-bit OS).

Models	Performance										Efficiency	
	RMSE ($\mu g\,m^{-3}$)					**sMAPE** (%)					$t_{\mathbf{inf}}$ (s)	**Param #**
	NO_2	O_3	PM_{10}	$PM_{2.5}$	Total	NO_2	O_3	PM_{10}	$PM_{2.5}$	Total		
MLP	6.63	7.53	7.82	4.85	6.71	35.89	41.90	65.24	53.15	49.04	0.0272	17 604
HMLP	5.99	6.83	7.95	4.62	6.35	31.84	39.44	65.42	**48.00**	46.27	0.1352	15 620
LSTM	5.97	6.39	7.48	4.32	6.04	32.09	38.09	63.40	51.70	46.32	0.0144	572 640
HLSTM	5.36	6.57	6.60	4.00	5.63	**28.53**	38.83	60.80	51.76	44.98	0.0187	72 244
GRU	6.01	6.18	6.84	3.94	5.74	32.62	38.46	61.15	49.67	45.47	0.0479	363 360
HGRU	**5.35**	**6.01**	**6.59**	**3.92**	**5.47**	28.78	**36.97**	**59.92**	52.40	**44.52**	0.0774	74 948

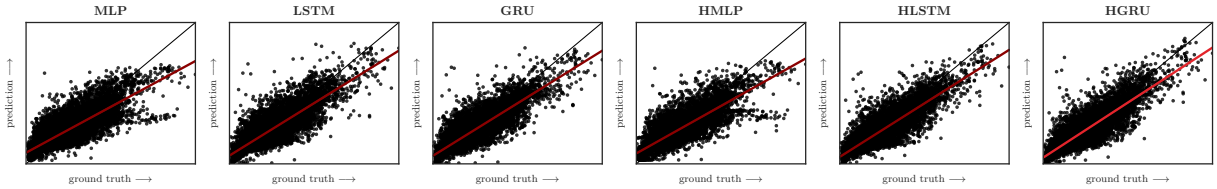


Figure 4.1: Each datapoint for all pollutants in the evaluation set scattered, with the ground truth as x and the prediction as y. Precisely correct predictions are along the diagonal. The maroon trend line gives a visual indication of performance. As indicated by the red line, the HGRU performs optimally. A non-combined depiction per pollutant is shown in Figure E.9, E.10, E.11, and E.12.

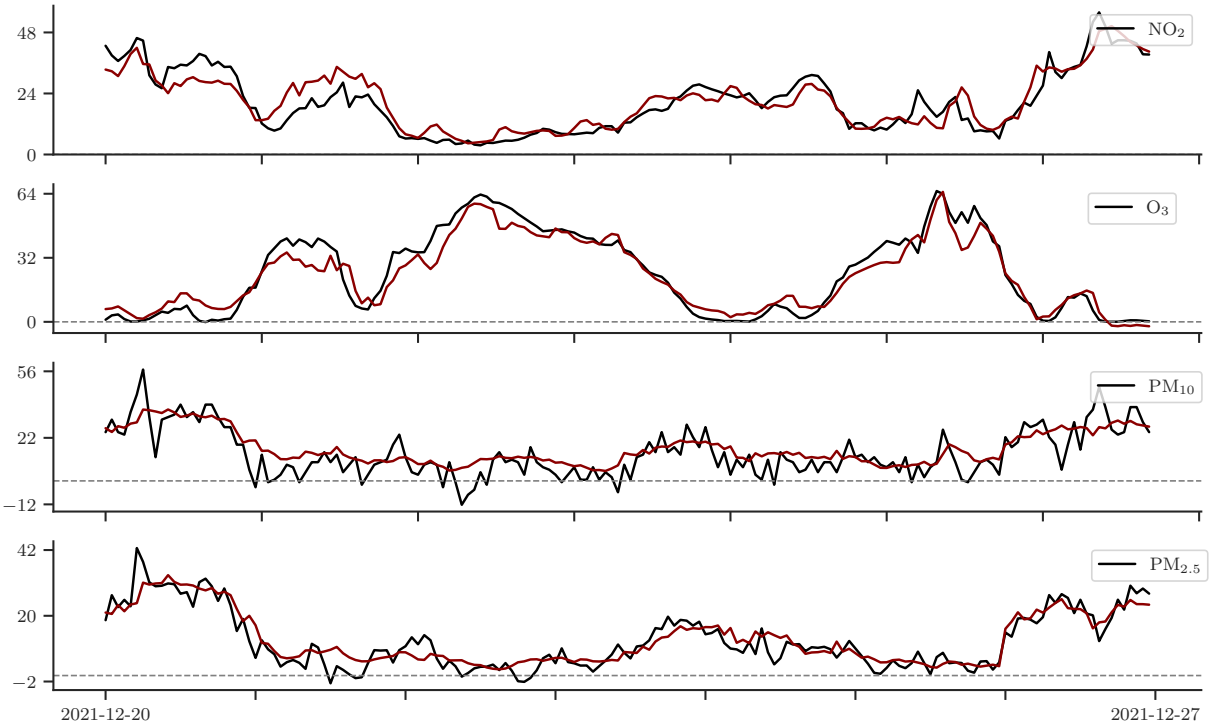


Figure 4.2: Hierarchical GRU (HGRU) forecasts for NO_2, O_3, PM_{10}, and $PM_{2.5}$ taken for a week (= seven 24-hour forecasts) from the evaluation set. Black indicates the ground truth and maroon the forecasts. Dashed lines indicate zero. More (HGRU) forecasts are presented in Appendix E.

rameters (due to reduced parameter sharing) than the non-hierarchical RNNs as well.

5 Discussion

This section discusses interpretations of the results, the project's limitations, and implications.

5.1 Interpretations

First and with a practical purpose, comparing the optimal RMSEs with Table 2.1 reveals that the guidelines are multiple times higher. The forecasts can, thus, already be deemed theoretically functional if deployed for the signalling of guideline exceedances.

In the context of sMAPE scores, gaseous pollutants outperform PM, with NO_2 being the most accurately predicted. As for the difference in sMAPE between primary pollutant NO_2 and secondary pollutant O_3, Lee et al. (2014) suggest that ambient primary pollutants depend primarily on their respective source emissions whilst secondary pollutants depend not only on the primary emissions but also on the availability of atmospheric oxidants. Consequently, the prediction of NO_2 is inherently more tractable than that of O_3. The sMAPE scores of NO_2 of O_3 reflect this difference in predictability.

Furthermore, the sMAPEs may appear high but require a more nuanced interpretation. A look at Figure 4.2 (and at the forecasts in Appendix E) gives away that forecasts and ground truths of the gaseous pollutants synchronise to a fair degree. For the PMs, however, the graphed forecasts are more indicative of an average, showing a lack of retrieval of sharply non-linear patterns, which is especially noticeable for peaks in PM_{10}. The higher errors of the PMs can be explained by their more "jagged," more frequently fluctuating temporal evolution being more challenging to capture—see, e.g., Figure E.1 and E.2 for the PMs versus the gaseous pollutants on different timescales. These issues are less evident for the smoother NO_2 and O_3.

Also, Figure 4.1 showed that all models suffer a negative bias on the testing set. Despite precautionary measures, the data shifts illustrated in Figure 3.3 hinder the models' performance here. Figure E.11 and E.12, for instance, where the RNNs outperform the MLPs and the hierarchical RNNs particularly excel, would support this by pointing out that the RNNs' long-term memory traces and the hierarchical models' nuanced regularisation make them more robust against a data shift—as opposed to the vanilla MLP. Possible future refinements to accommodate this problem include diversifying the training date through data augmentation, which was recently applied to weather forecasting by Cheon et al. (2024) and generally has been shown to improve robustness and generalisation (Bengio et al., 2011), or by incorporating models that carry more inductive bias.

Nonetheless, given that the discussed forecasts are performed from 15 km away over one dimension within a three-dimensional complex system, but also given measurement inaccuracies (KNMI, 2023a) and imprecisely interpolated percentages shown in Appendix B.3—that both render the data more impure and complicate predictions, the scores evince a compelling degree of accuracy.

Regarding the models individually, a few points stood out. Firstly, the LSTMs across many training runs proved markedly more resistant to initiating and completing convergence. Both Table D.1 and the total required epochs for the HLSTM in Figure C.1 highlight this. This is consistent with the theory in Section 2.3: the GRUs indeed converge faster than the LSTMs—not to mention the lower errors achieved by the GRUs. Secondly, it stood out that MLPs are performing above expectations, considering they do not have access to long-term pattern retention. Their accuracy may suggest that the significance of long-term dependencies is overestimated or that the instantaneous nature of weather dynamics is underestimated. Nonetheless, precisely the improvements that move away from mostly "averaging" forecasts; see the PM predictions in Figure E.3 and Figure E.4 for example, are the most difficult and demanding.

5.2 Limitations and implications

This research's limitations are summarised by simplifying measures to keep it within a scope appropriate for a PoC and by conceptually inherent limitations. Simplifications provide openings for easy improvements and are discussed in Section 6.1. Notable inherent limitations of this study include: the data being limited to merely two sensors, which fails to honour the complexity (e.g. the multidimensionality, emission sources, geographical features) of the modelled system; the non-stationarity of the data not being explicitly taken into account neither in preprocessing nor in model design; and modelling at a location where the air pollution and smog clouds problem is almost absent, thus limiting the direct impact.

Following are implications. Recurrent deep architectures offer a promising addition or augmentation to traditional NWP, given their adequacy and efficiency. Additionally, the dataset's minimal transformation makes real-time and continuous predictions possible. Moreover, this study provides a basis for further exploration into DL's potential in pollution forecasting and its PIML-expansion-tuned design for extending the project with PIML. Both are explained in more detail in Section 6.1.

6 Conclusions

In this paper, multivariate timeseries forecasting of smog clouds, represented by NO_2, O_3, PM_{10}, and $PM_{2.5}$ concentrations, using RNNs is conducted. Specifically, meteorological and pollution data at A is used to forecast air pollution levels at B. The most sophisticated models, the HLSTM and HGRU, are benchmarked with unidirectional and fully-connected DL architectures.

The research question, "To what extent are models with the LSTM and GRU architecture capable of the multivariate timeseries forecasting of smog-related air components?" is answered by the fact that the models are indeed highly adequate. Results demonstrate that, above all, the HGRU is suitable and competitive at this task. Reasons include the sequence-processing prowess of RNNs, a GRU's simplicity, and an integrated design streamlined to the very nature of the pollutants.

To sum up, our study contributes a PoC of smog cloud modelling using RNNs, providing a basis for advancements in pollution and weather forecasting to improve future public health.

6.1 Future research

This study opens up possibilities for numerous future adaptations, firstly due to its proof-of-concept nature and secondly because of the anticipatory design choices made for the extension to physics-informed neural nets (PINNs) with methods from PIML. This section discusses them in that order.

Improvements omitted for the sake of simplicity in the realm of data include: a refined interpolation method more fitting (see Figure E.2 for an illustration) to non-linear data such as polynomial or sinusoidal regression (Freeman et al. (2018) employs an advanced example); adding a significant constituent to sulphurous smog (Davis, 2002), sulphur dioxide (SO_2), as a predictive variable; utilising data from regions with more volatile pollution spectra and persistent smog, such as Medellín, Colombia (Peláez et al., 2020) and Beijing, China (Li et al., 2017b), to assess the models' compatibility there and to amplify the popultation health impacts; and consideration of stationarity.

If the data is non-stationary, data decomposition with a trend and seasonal component can transform the data to stationarity and additionally shift the models' focus towards more "buried" patterns, possibly previously obscured by the removed components. Another motive to do this is that seasonality and trends are inherently present precisely in nature, amplifying the effect.

Besides, while data decomposition may not necessarily solve the problem of the data shifts, it can help mitigate it (Hyndman and Athanasopoulos, 2018). Another way to make the models more robust to unseen data is via data augmentation or simply by acquiring more data from additional years or sensors. When scaling to more sensors or locations, the prediction domain becomes two-dimensional, the data can be represented as a graph, and graph neural networks (GNNs) (Scarselli et al., 2008) can be employed as well, aiming to enhance accuracy by leveraging spatial relationships in the data, while also aligning better with real-world conditions.

6.1.1 Physics-informed ML

PIML, where both scientific knowledge and ML are integrated in a synergistic manner (Karniadakis et al., 2021; Willard et al., 2022), is gaining increasing popularity within computational fluid dynamics (CFD) because of its data efficiency and stabler predictions over purely data-driven, inductive modelling (Kochkov et al., 2021; Sharma et al., 2023). Particularly relevant to this study are PINNs, neural networks trained to solve supervised learning tasks while respecting any given laws of physics described by general nonlinear partial differential equations (PDEs), as introduced by Raissi et al. (2019). Several anticipatory design choices made during the project have prepared for a prospect extension to PINNs, which will be explained briefly.

Qualities of PIML attractive to the modelling problem include greater physical consistency, improved data efficiency, and better generalisation (Kashinath et al., 2021). Because of first principles' invariance to distributions, an injection of such principles would help combat the in Figure 3.3 seen data shifts present in pollution data and also to help the nets, which in plain form are known to be poorly calibrated for out-of-distribution data (Guo et al., 2017), generalise to unseen situations, such as locations unseen in the data, or literal unseen situations caused by e.g., accelerating climate change (Lee et al., 2023). In addition, considering the possible high-stake environments pollution forecasting might be used for, O'Driscoll et al. (2019) prove, based on mean-variance portfolio theory and bias-variance trade-off analysis, that hybridised models have reduced model risk, increasing user trust and helping in model adaptation, the overarching goal.

When viewing ML and traditional workflow as a spectrum, a balance has to be struck between their pros and cons. On the ML side—where this study provides a start—a relatively non-invasive yet influential addition is adding a physics term to the loss function, combining (3.6) and a prediction by a one-dimensional depiction of the Navier-Stokes equations

$$\rho(\frac{\partial c}{\partial t} + v\frac{\partial c}{\partial x}) = -\frac{\partial p}{\partial x}, \tag{6.1}$$

where c denotes concentration and x denotes along

the (perhaps, discretised for computational feasibility) spatial axis between A and B. Assumptions include the airflow (which contains the pollutants) to be Newtonian, incompressible, isothermal, etc. This method leaves the architectures unchanged and encodes the physics directly into the loss function, giving an accessible yet effective method for hybridising the strengths of data-driven and first-principle modelling.

Acknowledgements

This endeavor would not have been possible without my supervisor Juan Diego and I thank him for his invaluable patience, feedback and discussions. Thanks should also go to the Center for Information Technology of the University of Groningen for providing access to the Hábrók HPC cluster NVIDIA V100 GPUs, offering an efficient and sustainable alternative to running the models locally.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Richard B. Alley, Kerry A. Emanuel, and Fuqing Zhang. Advances in weather prediction. *Science*, 363(6425):342–344, January 2019. ISSN 1095-9203. doi:10.1126/science.aav7274. URL `http://dx.doi.org/10.1126/science.aav7274`.

Markus Amann. *Health risks of ozone from long-range transboundary air pollution*. WHO Regional Office Europe, 2008.

R Atkinson. Atmospheric chemistry of vocs and nox. *Atmospheric Environment*, 34 (12–14):2063–2101, 2000. ISSN 1352-2310. doi:10.1016/s1352-2310(99)00460-4. URL `http://dx.doi.org/10.1016/s1352-2310(99)00460-4`.

Kirk R. Baker and Kristen M. Foley. A nonlinear regression model estimating single source concentrations of primary and secondarily formed pm2.5. *Atmospheric Environment*, 45 (22):3758–3767, July 2011. ISSN 1352-2310. doi:10.1016/j.atmosenv.2011.03.074. URL `http://dx.doi.org/10.1016/j.atmosenv.2011.03.074`.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, September 2015. ISSN 1476-4687. doi:10.1038/nature14956. URL `http://dx.doi.org/10.1038/nature14956`.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, March 1994. ISSN 1941-0093. doi:10.1109/72.279181. URL `http://dx.doi.org/10.1109/72.279181`.

Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger-Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172. JMLR Workshop and Conference Proceedings, 2011.

Jan Bijma. *Nauwkeurigheid van operationele temperatuurmetingen*. KNMI, 2012.

Bert Brunekreef and Stephen T Holgate. Air pollution and health. *The Lancet*, 360 (9341):1233–1242, October 2002. ISSN 0140-6736. doi:10.1016/s0140-6736(02)11274-8. URL `http://dx.doi.org/10.1016/s0140-6736(02)11274-8`.

Sheen Mclean Cabaneros, John Kaiser Calautit, and Ben Richard Hughes. A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling and Software*, 119:285–304, September 2019. ISSN 1364-8152. doi:10.1016/j.envsoft.2019.06.014. URL `http://dx.doi.org/10.1016/j.envsoft.2019.06.014`.

Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

Chris Chatfield. *Time-series forecasting*. CRC press, 2000.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1), April 2018. ISSN 2045-2322. doi:10.1038/s41598-018-24271-9. URL `http://dx.doi.org/10.1038/s41598-018-24271-9`.

Lung Chi Chen and Morton Lippmann. Effects of metals within ambient air particulate matter (pm) on human health. *Inhalation Toxicology*, 21(1):1–31, January 2009. ISSN 1091-7691. doi:10.1080/08958370802105405. URL `http://dx.doi.org/10.1080/08958370802105405`.

Changqing Cheng, Akkarapol Sa-Ngasoongsong, Omer Beyca, Trung Le, Hui Yang, Zhenyu (James) Kong, and Satish T.S. Bukkapatnam. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. *IIE Transactions*, 47 (10):1053–1071, January 2015. ISSN 1545-8830. doi:10.1080/0740817x.2014.999180. URL http://dx.doi.org/10.1080/0740817x.2014.999180.

Minjong Cheon, Daehyun Kang, Yo-Hwan Choi, and Seon-Yu Kang. Advancing data-driven weather forecasting: Time-sliding data augmentation of era5. *arXiv preprint arXiv:2402.08185*, 2024.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Ronnachai Chuentawat and Yosporn Kan-ngan. The comparison of pm2.5 forecasting methods in the form of multivariate and univariate time series based on support vector machine and genetic algorithm. In *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, July 2018. doi:10.1109/ecticon.2018.8619867. URL http://dx.doi.org/10.1109/ecticon.2018.8619867.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

L Clapp. Analysis of the relationship between ambient levels of o3, no2 and no as a function of nox in the uk. *Atmospheric Environment*, 35(36):6391–6405, December 2001. ISSN 1352-2310. doi:10.1016/s1352-2310(01)00378-8. URL http://dx.doi.org/10.1016/s1352-2310(01)00378-8.

DL Davis. A look back at the london smog of 1952 and the half century since. *Environmental Health Perspectives*, 110(12), December 2002. ISSN 1552-9924. doi:10.1289/ehp.110-a734. URL http://dx.doi.org/10.1289/ehp.110-a734.

Wenying Duan, Xiaoxi He, Lu Zhou, Lothar Thiele, and Hong Rao. Combating distribution shift for accurate time series forecasting via hypernetworks. In *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, January 2023. doi:10.1109/icpads56603.2022.00121. URL http://dx.doi.org/10.1109/icpads56603.2022.00121.

Mark D Eisner, Nicholas Anthonisen, David Coultas, Nino Kuenzli, Rogelio Perez-Padilla, Dirkje Postma, Isabelle Romieu, Edwin K Silverman, and John R Balmes. An official american thoracic society public policy statement: Novel risk factors and the global burden of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 182(5): 693–718, 2010.

M.A. Elangasinghe, N. Singhal, K.N. Dirks, J.A. Salmond, and S. Samarasinghe. Complex time series analysis of pm10 and pm2.5 for a coastal site using artificial neural network modelling and k-means clustering. *Atmospheric Environment*, 94:106–116, September 2014. ISSN 1352-2310. doi:10.1016/j.atmosenv.2014.04.051. URL http://dx.doi.org/10.1016/j.atmosenv.2014.04.051.

Derek Elsom. *Smog alert: managing urban air quality*. Routledge, 2014.

Annunziata Faustini, Regula Rapp, and Francesco Forastiere. Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. *ISEE Conference Abstracts*, 2013(1), September 2013. ISSN 1078-0475. doi:10.1289/isee.2013.p-2-05-23. URL http://dx.doi.org/10.1289/isee.2013.p-2-05-23.

BJ Finlayson-Pitts and JN Pitts Jr. Atmospheric chemistry of tropospheric ozone formation: scientific and regulatory implications. *Air & Waste*, 43(8):1091–1100, 1993.

Rick Fontenot, Joseph Lazarus, Puri Rudick, and Anthony Sgambellone. Hierarchical neural networks (hnn): Using tensorflow to build hnn. *SMU Data Science Review*, 6(2):4, 2022.

David Fowler, Peter Brimblecombe, John Burrows, Mathew R Heal, Peringe Grennfelt, David S Stevenson, Alan Jowett, Eiko Nemitz, Mhairi Coyle, Xuejun Liu, et al. A chronology of global air quality. *Philosophical Transactions of the Royal Society A*, 378(2183):20190314, 2020.

Brian S. Freeman, Graham Taylor, Bahram Gharabaghi, and Jesse Thé. Forecasting air quality time series using deep learning. *Journal of the Air and Waste Management Association*, 68(8):866–886, May 2018. ISSN 2162-2906. doi:10.1080/10962247.2018.1459956. URL http://dx.doi.org/10.1080/10962247.2018.1459956.

Everette S Gardner Jr. Exponential smoothing: The state of the art—part ii. *International journal of forecasting*, 22(4):637–666, 2006.

Fred Gelbard and John H Seinfeld. The general dynamic equation for aerosols. theory and application to aerosol formation and growth. *Journal of Colloid and Interface Science*, 68(2):363–382, 1979.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

Sotirios D Glavas and Eleni Sazakli. Ozone long-range transport in the balkans. *Atmospheric environment*, 45(8):1615–1626, 2011.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10): 2222–2232, October 2017. ISSN 2162-2388. doi:10.1109/tnnls.2016.2582924. URL `http://dx.doi.org/10.1109/tnnls.2016.2582924`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

A. J. Haagen-Smit. Chemistry and physiology of los angeles smog. *Industrial and Engineering Chemistry*, 44(6):1342–1346, June 1952. ISSN 1541-5724. doi:10.1021/ie50510a045. URL `http://dx.doi.org/10.1021/ie50510a045`.

Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427, January 2021. ISSN 0169-2070. doi:10.1016/j.ijforecast.2020.06.008. URL `http://dx.doi.org/10.1016/j.ijforecast.2020.06.008`.

Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1):31, 1991.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

James R Holton and Gregory J Hakim. *An introduction to dynamic meteorology*. Academic press, 2012.

Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

J.G. Irwin and M.L. Williams. Acid rain: Chemistry and transport. *Environmental Pollution*, 50(1–2):29–59, 1988. ISSN 0269-7491. doi:10.1016/0269-7491(88)90184-4. URL `http://dx.doi.org/10.1016/0269-7491(88)90184-4`.

Herbert Jaeger. Lecture Notes: Machine Learning. `https://www.ai.rug.nl/minds/uploads/LN_ML_RUG.pdf`, 2023a. Accessed on 18-04-2023.

Herbert Jaeger. Lecture Notes: Neural Networks. `https://www.ai.rug.nl/minds/uploads/LN_NN_RUG.pdf`, 2023b. Accessed on 02-07-2023.

Christer Johansson, Lars Burman, and Bertil Forsberg. The effects of congestions tax on air quality and health. *Atmospheric Environment*, 43(31):4843–4854, October 2009. ISSN 1352-2310. doi:10.1016/j.atmosenv.2008.09.015. URL `http://dx.doi.org/10.1016/j.atmosenv.2008.09.015`.

Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.

Ferhat Karaca, Ismail Anil, and Omar Alagha. Long-range potential source contributions of episodic aerosol events to pm10 profile of a megacity. *Atmospheric Environment*, 43(36): 5713–5722, November 2009. ISSN 1352-2310. doi:10.1016/j.atmosenv.2009.08.005. URL `http://dx.doi.org/10.1016/j.atmosenv.2009.08.005`.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu

Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, Ashesh Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

KNMI. Luchtkwaliteitsdata als csv bestanden, versie 7, 2023a. URL `https://data.rivm.nl/data/luchtmeetnet/readme.pdf`. Accessed on 14/10/2023.

KNMI. Uitleg over automatisch weerstations, 2023b. URL `https://www.knmi.nl/kennis-en-datacentrum/uitleg/automatische-weerstations`. Accessed on 14/10/2023.

KNMI. KNMI dataplatform, 2024. URL `https://dataplatform.knmi.nl/`. Accessed on 15/03/2024.

Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barrett, et al. *Climate change 2023: synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change*. The Australian National University, 2023.

Hyung Joo Lee, Choong-Min Kang, Brent A. Coull, Michelle L. Bell, and Petros Koutrakis. Assessment of primary and secondary ambient particle trends using satellite aerosol optical depth and ground speciation data in the new england region, united states. *Environmental Research*, 133:103–110, August 2014. ISSN 0013-9351. doi:10.1016/j.envres.2014.04.006. URL `http://dx.doi.org/10.1016/j.envres.2014.04.006`.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017a.

Sike Li, Kuishuang Feng, and Mengxue Li. Identifying the main contributors of air pollution in beijing. *Journal of Cleaner Production*, 163:S359–S365, October 2017b. ISSN 0959-6526. doi:10.1016/j.jclepro.2015.10.127. URL `http://dx.doi.org/10.1016/j.jclepro.2015.10.127`.

Stephen S Lim, Theo Vos, Abraham D Flaxman, Goodarz Danaei, Kenji Shibuya, Heather Adair-Rohani, Mohammad A AlMazroa, Markus Amann, H Ross Anderson, Kathryn G Andrews, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2224–2260, 2012.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, January 2020. ISSN 0169-2070. doi:10.1016/j.ijforecast.2019.04.014. URL `http://dx.doi.org/10.1016/j.ijforecast.2019.04.014`.

Christopher S. Malley, Daven K. Henze, Johan C.I. Kuylenstierna, Harry W. Vallack, Yanko Davila, Susan C. Anenberg, Michelle C. Turner, and Mike R. Ashmore. Updated global estimates of respiratory mortality in adults 30+ years of age attributable to long-term ozone exposure. *Environmental Health Perspectives*, 125(8), August 2017. ISSN 1552-9924. doi:10.1289/ehp1390. URL `http://dx.doi.org/10.1289/ehp1390`.

Adil Masood and Kafeel Ahmad. A review on emerging artificial intelligence (ai) techniques for air pollution forecasting: Fundamentals, application and performance. *Journal of Cleaner Production*, 322:129072, November 2021. ISSN 0959-6526. doi:10.1016/j.jclepro.2021.129072. URL `http://dx.doi.org/10.1016/j.jclepro.2021.129072`.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

Micheal L Mavrovouniotis and S Chang. Hierarchical neural networks. *Computers & chemical engineering*, 16(4):347–369, 1992.

Ludmila Mašková, Jiří Smolík, and Petr Vodička. Characterisation of particulate matter in different types of archives. *Atmospheric Environment*, 107:217–224, April 2015. ISSN 1352-2310. doi:10.1016/j.atmosenv.2015.02.049. URL `http://dx.doi.org/10.1016/j.atmosenv.2015.02.049`.

Patrick O'Driscoll, Jaehoon Lee, and Bo Fu. Physics enhanced artificial intelligence. *arXiv preprint arXiv:1903.04442*, 2019.

Pablo Orellano, Julieta Reynoso, Nancy Quaranta, Ariel Bardach, and Agustin Ciapponi. Short-term exposure to particulate matter (pm10 and pm2.5), nitrogen dioxide (no2), and ozone (o3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment International*, 142:105876, September 2020. ISSN 0160-4120. doi:10.1016/j.envint.2020.105876. URL http://dx.doi.org/10.1016/j.envint.2020.105876.

Riccardo Orioli, Giuseppe Cremona, Luisella Ciancarella, and Angelo G. Solimini. Association between pm10, pm2.5, no2, o3 and self-reported diabetes in italy: A cross-sectional, ecological study. *PLOS ONE*, 13 (1):e0191112, January 2018. ISSN 1932-6203. doi:10.1371/journal.pone.0191112. URL http://dx.doi.org/10.1371/journal.pone.0191112.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Luisa María Gómez Peláez, Jane Meri Santos, Taciana Toledo de Almeida Albuquerque, Neyval Costa Reis Jr, Willian Lemker Andreão, and Maria de Fátima Andrade. Air quality status and trends over large cities in south america. *Environmental Science & Policy*, 114:422–435, 2020.

Victor R. Prybutok, Junsub Yi, and David Mitchell. Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentrations. *European Journal of Operational Research*, 122(1):31–40, April 2000. ISSN 0377-2217. doi:10.1016/s0377-2217(99)00069-7. URL http://dx.doi.org/10.1016/s0377-2217(99)00069-7.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. 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, February 2019. ISSN 0021-9991. doi:10.1016/j.jcp.2018.10.045. URL http://dx.doi.org/10.1016/j.jcp.2018.10.045.

Martin J Rhodes and Jonathan Seville. *Introduction to particle technology*. John Wiley & Sons, 2024.

Hannah Ritchie and Max Roser. Urbanization. *Our world in data*, 2018.

RIVM. RIVM luchtmeetnet datasets, 2024. URL https://data.rivm.nl/data/. Accessed on 15/03/2024.

Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

Pallavi Saxena and Vaishali Naik. *Air pollution: sources, impacts and controls*. Cabi, 2018.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

John H Seinfeld and Spyros N Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016.

A. Sfetsos and A.H. Coonick. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy*, 68(2):169–178, February 2000. ISSN 0038-092X. doi:10.1016/s0038-092x(99)00064-x. URL http://dx.doi.org/10.1016/s0038-092x(99)00064-x.

Pushan Sharma, Wai Tong Chung, Bassem Akoush, and Matthias Ihme. A review of physics-informed machine learning in fluid mechanics. *Energies*, 16(5):2343, February 2023. ISSN 1996-1073. doi:10.3390/en16052343. URL http://dx.doi.org/10.3390/en16052343.

Kai Shi, Baofeng Di, Kaishan Zhang, Chaoyang Feng, and Laurence Svirchev. Detrended cross-correlation analysis of urban traffic congestion and no 2 concentrations in chengdu. *Transportation Research Part D: Transport and Environment*, 61:165–173, June 2018. ISSN 1361-9209. doi:10.1016/j.trd.2016.12.012. URL http://dx.doi.org/10.1016/j.trd.2016.12.012.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Qing Tao, Fang Liu, Yong Li, and Denis Sidorov. Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru. *IEEE Access*, 7:76690–76698, 2019. ISSN 2169-3536. doi:10.1109/access.2019.2921578. URL http://dx.doi.org/10.1109/access.2019.2921578.

UN. Transforming our world: the 2030 agenda for sustainable development, 2015. URL https://sdgs.un.org/2030agenda.

Daniel A Vallero. *Fundamentals of air pollution*. Academic press, 2014.

GJM Velders, JMM Aben, GP Geilenkirchen, HA Den Hollander, E van der Swaluw, WJ de Vries, and MC van Zanten. *Grootschalige concentratie-en depositiekaarten Nederland: Rapportage 2015*. Rijksinstituut voor Volksgezondheid en Milieu RIVM, 2015.

Guus J.M. Velders, Saskia M. Willers, Joost Wesseling, Sef van den Elshout, Eric van der Swaluw, Dennis Mooibroek, and Sjoerd van Ratingen. Improvements in air quality in the netherlands during the corona lockdown based on observations and model simulations. *Atmospheric Environment*, 247: 118158, February 2021. ISSN 1352-2310. doi:10.1016/j.atmosenv.2020.118158. URL http://dx.doi.org/10.1016/j.atmosenv.2020.118158.

A. Vlachogianni, P. Kassomenos, Ari Karppinen, S. Karakitsios, and Jaakko Kukkonen. Evaluation of a multiple regression model for the forecasting of the concentrations of nox and pm10 in athens and helsinki. *Science of The Total Environment*, 409(8): 1559–1571, March 2011. ISSN 0048-9697. doi:10.1016/j.scitotenv.2010.12.040. URL http://dx.doi.org/10.1016/j.scitotenv.2010.12.040.

WHO. *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, 2021.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37, 2022.

Stephen Wolfram and M Gad-el Hak. A new kind of science. *Appl. Mech. Rev.*, 56(2):B18–B19, 2003.

Chit-Ming Wong, Nuntavarn Vichit-Vadakan, Haidong Kan, and Zhengmin Qian. Public health and air pollution in asia (papa): A multicity study of short-term effects of air pollution on mortality. *Environmental Health Perspectives*, 116(9):1195–1202, September 2008. ISSN 1552-9924. doi:10.1289/ehp.11257. URL http://dx.doi.org/10.1289/ehp.11257.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31(7):1235–1270, July 2019. ISSN 1530-888X. doi:10.1162/neco_a_01199. URL http://dx.doi.org/10.1162/neco_a_01199.

Jelle Zorn, Mariana Simões, Guus J.M. Velders, Miriam Gerlofs-Nijland, Maciek Strak, José Jacobs, Marieke B.A. Dijkema, Thomas J. Hagenaars, Lidwien A.M. Smit, Roel Vermeulen, Lapo Mughini-Gras, Lenny Hogerwerf, and Don Klinkenberg. Effects of long-term exposure to outdoor air pollution on covid-19 incidence: A population-based cohort study accounting for sars-cov-2 exposure levels in the netherlands. *Environmental Research*, 252:118812, July 2024. ISSN 0013-9351. doi:10.1016/j.envres.2024.118812. URL http://dx.doi.org/10.1016/j.envres.2024.118812.

A Air pollution: individual health implications

This section goes into more detail on the health implications caused and carried by air pollution, providing a brief expansion and overview of the discussion introduced in Sections 1.1 and 2.1.1. Specifically, discussed is further problem contextualisation, individual health effects attributable to the modelled pollutants, and how those connect back to the project's motivation.

Context. Over the last century, pollutant and trace gas concentrations have increased substantially, fundamentally changing the air quality (Fowler et al., 2020). Major drivers of these increases are the arrival of industrialisation, large-scale fossil fuel combustion, and further urbanisation. The acceleration and expansion of production processes, coupled with the combustion of natural gases, acted as a catalyst for the interaction between different components and their concentrations in ambient air, effectively moving them from underground to the atmosphere. A by-product of urbanisation is spatial areas with a high concentration of polluting activities (Elsom, 2014), which can accumulate atmospheric pollution, forming "clouds of smog" of either sulphurous or photochemical kind (Davis, 2002; Haagen-Smit, 1952). The prevalence of smog will not subside anytime soon (as, for instance, estimates predict the world's urban population to grow from ±55% in 2018 to ±68% in 2050 (Ritchie and Roser, 2018)), nor will air pollution in general. Moreover, the same applies to the consequential health effects.

Implications. In continuation of the general descriptions of adverse health effects touched upon in the introduction, Section 1, individual air components are also associated with health effects. Kampa and Castanas (2008) list, among others, gaseous pollutants and respirable PM as two significant categories of air pollutants.

Of the gaseous pollutants category, O_3 annually accounts for 21,000 premature deaths in Europe (Amann, 2008) and more than 1.1 million deaths worldwide—that is more than 20% of all deaths attributed to respiratory diseases (Malley et al., 2017). For another prevalent gaseous pollutant, NO_2, analyses, such as Eisner et al. (2010) and Faustini et al. (2013), found NO_2 to play a causal role in mortality and development of chronic respiratory diseases.

The other category, PM, describes a group of pollutants consisting of an intricate and heterogeneous mix of tiny particles and droplets suspended in breathing air with an aerodynamic diameter of $\leq 10\,\mu m$ or, even smaller, $\leq 2.5\,\mu m$ for PM_{10} and $PM_{2.5}$, respectively (Mašková et al., 2015). Some particles are visible without any aid, such as dust, dirt, soot, or smoke, while others are too small to be seen and require an electron microscope for their detection. With their tiny sizes, they can reach deep into the respiratory system, all the way to the alveoli, and inflict substantial harm there (Kampa and Castanas, 2008)—harm caused by, for example, (heavy) metals present in the PM (Chen and Lippmann, 2009).

Another illustratory example is Orioli et al. (2018). They found that the gaseous pollutants nitrogen dioxide (NO_2) and O_3, and the PM-types PM_{10} and $PM_{2.5}$, all show a positive association with diabetes—a link that at first glance, one might not anticipate. Moreover, in some cases, smog can even cause harm to agricultural crops, as seen with a particular variety of intense smog that left a metallic sheen on the leaves of spinach, sugar beets, and endive (Haagen-Smit, 1952).

Motivation. Following the aforementioned impacts of all four components, together with the acknowledgement that many components in the weather system affect air quality and its cleanliness, the scope of this research focused specifically on the contaminants NO_2, O_3, $PM_{2.5}$, and PM_{10}. The motivation for modelling the four pollutants NO_2, O_3, PM_{10}, and $PM_{2.5}$ thus consisted of both their contribution to the formation of smog and their pervasive health effects.

B Data insights

This Appendix section provides some additional insight into the data by exploring the meteorogical variables and discussing the correlations of all features. In addition, an overview of the data availabilities is displayed, some data statistics are presented, and an extraordinary outlier is visualised.

B.1 Exploration of meteorological data

This subsection provides a visual overview of all the initially considered meteorological variables (Figure B.1), accompanied by an explanation of why these variables might be helpful for modelling smog clouds, i.e. the pollutants NO_2, O_3, PM_{10}, $PM_{2.5}$. Important to emphasise is that not all of these rationales have held up in feature selection (or, worded differently, showed in the data).

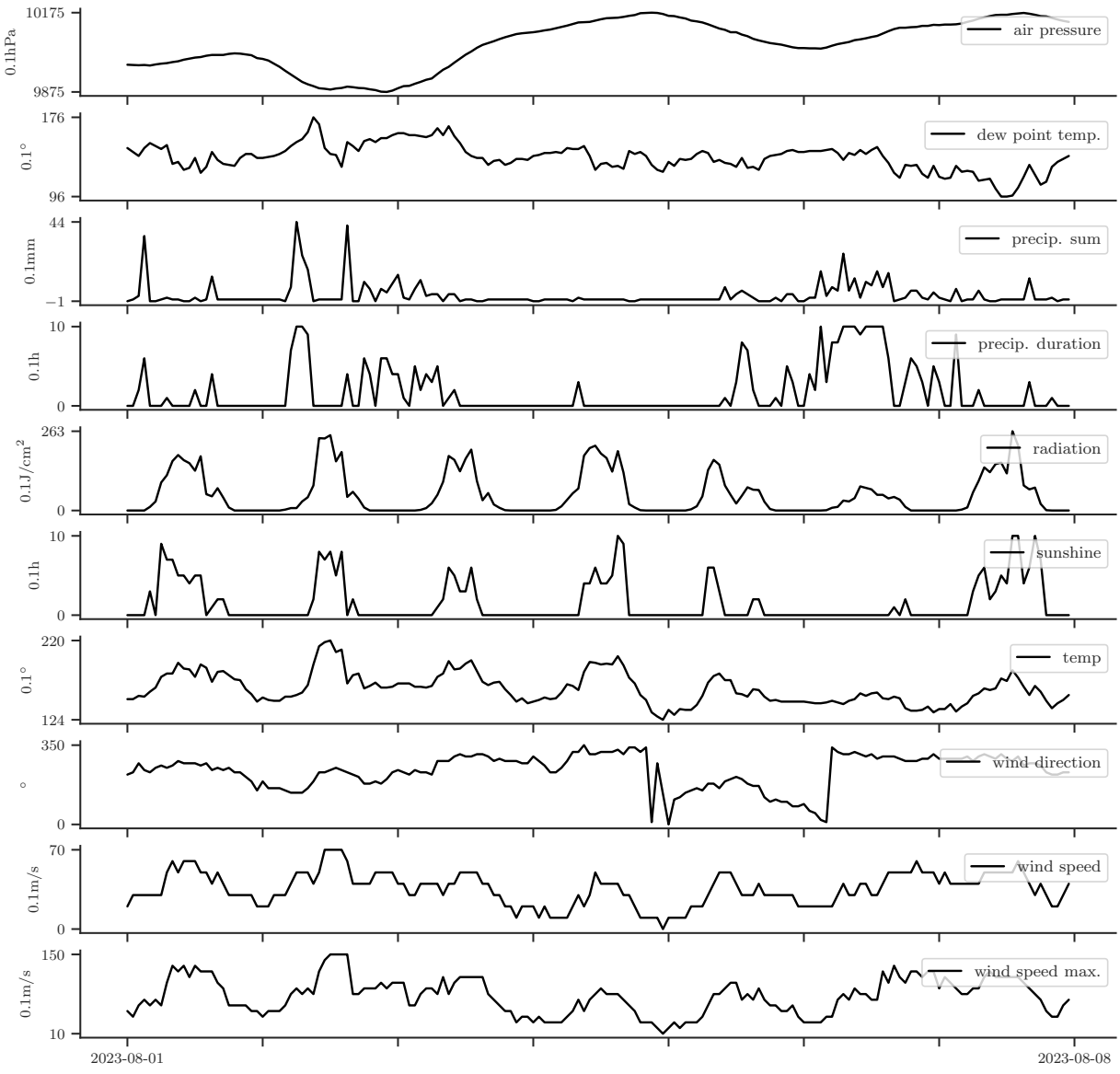


Figure B.1: Plots of all initially considered meteorological variables, before any preprocessing. From top to bottom, in alphabetic order: air pressure (0.1 hPa), dew point temperature (0.1 °C), precipitation sum (0.1 mm), precipitation duration (0.1 h), radiation ($J\,cm^{-2}$), sunshine (0.1 h), temperature (0.1 °C), wind direction ($0 - 360°$), wind speed ($0.1\,m\,s^{-1}$), and wind speed maximum ($0.1\,m\,s^{-1}$). For reference, these are also listed in Table 3.1.

It is important to stress that besides the individual characteristics of the pollutants, they are located in the tropospheric sky, and have such a low mass they can be assumed to behave as air in terms of their interaction with large-scale meteorological processes. We will go through the variables from top to bottom

to explain their relevance:

- **Air pressure** can indicate dispersion and transport of large air currents, for example by low and high-pressure areas, and thus also influence the air currents of pollutants (Holton and Hakim, 2012).
- **Dew point temperature, precipitation sum, and precipitation amount** are indicators of atmospheric moisture content levels. These levels can say something about, for example, the rate of condensation (which, via nucleation, can lead to the formation of fine aerosols (PM) (Gelbard and Seinfeld, 1979)), any scavenging and cleansing of the air by rainfall (lowering the pollutant concentrations) (Vallero, 2014), and the formation of acid rain (where acidic gases SO_2 and NO_x that are (related to) the predictive variables, get washed out, thus lowering the concentrations (Irwin and Williams, 1988)).
- **Global radiation and sunshine**, which signify the presence of solar energy in the form of photons, serve as fundamental drivers of low-entropy energy input on Earth. One direct way where this is observed, is in the photochemical processes discussed in Section 2.1.2.
- **Temperature** is an essential factor in chemical processes seen by its role as accelerator in the formation of secondary pollutants. In addition, temperature plays a role in atmospheric stability, with, for example, (suddenly) high temperatures signifying increased convective activity. Furthermore, temperature influences state changes, and is also tightly connected with global radiation and sunshine, therewith also indirectly contributing to their effects. For more context on atmospheric chemistry and physics, refer to the extensive Seinfeld and Pandis (2016).
- **Mean wind direction, mean wind speed, and maximum wind gust** all tell about the wind's properties, which in turn carries the pollutants through the atmosphere. In context of the experiment, wind direction, for example, tells about the relative directional relationship between A and B. Out of the pollutants, the wind especially plays a role for the PMs, as they have a bigger surface.

B.2 Feature correlations

In Table B.1, the exact (and non-absolute) coefficients—in accordance to the correlation matrix Figure 3.4 discussed in Section 3.2—are shown for each pair of features.

Table B.1: Diagonal correlation matrix for the initially considered variables, that is, before any feature selection. The non-absolute pearson correlation coefficient is shown for: nitrogen dioxide (NO_2), ozone (O_3), particulate matter $\leq$ 10 µm (PM_{10}), particulate matter $\leq$ 2.5 µm ($PM_{2.5}$), air pressure (AP), dew point temperature (DPT), global radiation (GR), maximum wind gust (MWG), mean wind direction (MWD), mean wind speed (MWS), precipitation amount (PA), precipitation duration (PD), sunshine duration (SD), and temperature (T), which are all also listed in Table 3.1 with their respective units. During preprocessing, the bolded GR, MWG, PA, and PD were discarded, and MWG was augmented with its 24-hour moving average. The preprocessing steps are explained in Section 3.2. The bolded entries correspond to coefficients named in the text.

	NO_2	O_3	PM_{10}	$PM_{2.5}$	AP	DPT	**GR**	**MWG**	MWD	MWS	**PA**	**PD**	SD	T
NO_2														
O_3	**-0.59**													
PM_{10}	0.42	-0.26												
$PM_{2.5}$	0.39	-0.38	**0.64**											
AP	0.15	-0.11	0.18	0.17										
DPT	-0.23	0.37	-0.12	-0.24	0.02									
GR	-0.14	0.35	-0.06	-0.08	0.12	0.27								
MWG	-0.36	0.30	-0.20	-0.29	-0.38	-0.01	0.13							
MWD	-0.03	0.14	-0.12	-0.24	-0.03	0.08	0.06	0.25						
MWS	-0.34	0.24	-0.17	-0.23	-0.38	-0.04	0.11	**0.94**	0.17					
PA	-0.06	0.04	-0.09	-0.09	-0.18	0.05	-0.07	0.17	0.05	0.12				
PD	-0.08	0.01	-0.16	-0.13	-0.32	0.01	-0.15	0.24	0.06	0.21	**0.60**			
SD	-0.07	0.28	-0.03	-0.06	0.14	0.15	**0.79**	0.07	0.03	0.06	-0.09	-0.17		
T	-0.30	**0.60**	-0.13	-0.28	0.06	**0.88**	**0.58**	0.10	0.09	0.07	-0.02	-0.09	0.44	

A introductory general notion is that the correlations are not necessarily high. There are various reasons for this. Firstly, the values—to speak in the context of modelling—have not yet been processed and are (still) noisy, thus distorting the calculation. This noise can be caused, for example, by inconsistencies in the measurement equipment, such as fluctuating calibration, or simply disturbing factors, such as condensation on PM sensors (KNMI, 2023b; Bijma, 2012). Second, the calculation itself is not entirely sound. The Pearson correlation coefficient (defined as (3.2)) makes a number of assumptions, such as linear relationships, heteroskedasticity, and normally distributed data, all of which do not necessarily apply to the highly non-linear and possibly turbulent weather data. Thirdly, the variables merely represent a small part of the dynamical system they are part of, and their interdependencies cannot be captured by simply a number. Theoretically, with infinite data, and therefore infinite variables, the interrelationships should be captured. However, this is beyond the practical possibilities(—see Wolfram and Gad-el Hak (2003) for a further discussion of this topic). Having said this, it is worth highlighting a few individual coefficients:

- A first coefficient is $r_{O_3,NO_2} = -0.59$, which shows a negative correlation. Section 2.1.2 discussed atmospheric interactions, including the interaction between these two components. During the day, when the sun rises, NO_2 photodisassociates, thereby triggering the sequential formation of O_3. (This also translates into the coefficient of temperature with O_3, $r_{O_3,T} = 0.60$). This inversely related interaction—see Figure 3.2—explains the negative correlation in this case. A sidenote here is, for the sake of completeness, that many more processes cooperate in the creation of this number.
- Another particularly important one to highlight is the coefficient between subjects of this study PM_{10} and $PM_{2.5}$, $r_{PM_{10},PM_{2.5}} = 0.64$. This is high because the definitions of both PMs, separated by their size, are parallel. They otherwise come from roughly the same sources and, apart from a definition shift due to agglomeration and fragmentation of particles, will exhibit similar behaviour.
- Additionally, some peculiarly high coefficients can be observed. Evidently, this is because these variables are very similar. Take for instance $r_{T,DPT} = 0.88$, both of which describe temperature in minimally different ways, apart from an influence of humidity. Indeed, there are several more where the same pattern—a pattern of intrinsic similarities—inflate the coefficient: mean wind speed with maximum wind gust $r_{MWS,MWG} = 0.94$, precipitation amount with precipitation duration $r_{PA,PD} = 0.60$, sunshine duration with global radiation $r_{SD,GR} = 0.79$, or temperature with global radiation $r_{T,GR} = 0.58$, all of which can be explained intuitively. However, the opposite is also true: most data points have little to do with each other and therefore show weak correlation.

Altogether, it can be said from an ML perspective that the correlations present between predictive variables and covariates are naturally favourable for accurate prediction of pollutants, as are low correlations between the covariates themselves (Hall, 1999). For this reason, a number of features were selected for final modelling, as discussed in Section 3.2.

B.3 Data availability

Here is a short summary of the availability of the data used in the experiment. Missing data was interpolated with linear interpolation, (3.1).

Table B.2: Data availability percentage for the modelled pollutants for each year. The meteorological abbrevations are defined in Table 3.1. The meteorological data was completely available for all years—for the pollutants, it was not. Given the strict procedures by the KNMI (KNMI, 2023a), this is no surprise.

	NO_2	O_3	PM_{10}	$PM_{2.5}$	AP	DP	MWD	MWS	SD	T
2017	97.64%	96.85%	97.62%	99.10%	100%	100%	100%	100%	100%	100%
2018	99.67%	98.52%	97.70%	99.29%	100%	100%	100%	100%	100%	100%
2020	98.60%	98.05%	99.34%	99.31%	100%	100%	100%	100%	100%	100%
2021	99.67%	98.55%	98.88%	99.29%	100%	100%	100%	100%	100%	100%
2022	96.90%	97.62%	95.56%	99.62%	100%	100%	100%	100%	100%	100%
2023	98.60%	97.15%	98.46%	97.97%	100%	100%	100%	100%	100%	100%

B.4 An outlier visualised

Shown is a plot illustrating two points. Firstly, that New Year's is an outlier and should be excluded. Secondly, that SES (Gardner Jr, 2006):

$$f_{t+1} = \alpha y_t + (1 - \alpha) f_t, \tag{B.1}$$

where α is a smoothing factor determining the balance between observed value y_t and the previous forecast f_t, carries an inherent delay in its computational mechanism, underpinning the notion made in Section 2.2.

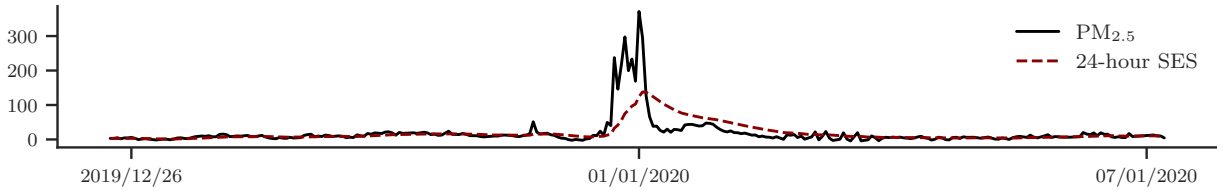


Figure B.2: Plot of hourly-sampled $PM_{2.5}$-concentration ($\mu g\, m^{-3}$) as a function of time and its 24-hour single exponential smoothing (SES) average, over a period around New Year's 2019. Notable is the striking—or, perhaps, concerning—peak around the 31st and 1st, when the fireworks and the concentrations take their flight. SES exhibits a delayed response to the abrupt peak.

B.5 Data allocation and quantity

This section provides transparency on how much data was used and in what proportions. Table B.3 shows the number of hours of data before pair generation, and Table B.4 the data after pair generation.

Table B.3: Hours of data for each feature per year in the training, validation, and testing sets before pair generation, illustrating the data balance between the different sets, and their amounts. Divide these by 24 for the amount of days. (Meteorological abbrevations are defined in Table 3.1).

	NO_2	O_3	PM_{10}	$PM_{2.5}$	AP	DP	MWD	MWS	SD	T
Train '17	3648	3648	3648	3648	3648	3648	3648	3648	3648	3648
Train '18	3648	3648	3648	3648	3648	3648	3648	3648	3648	3648
Train '20	3648	3648	3648	3648	3648	3648	3648	3648	3648	3648
Train '21	2640	2640	2640	2640	2640	2640	2640	2640	2640	2640
Train '22	2640	2640	2640	2640	2640	2640	2640	2640	2640	2640
Validation '21	504	504	504	504	504	504	504	504	504	504
Validation '22	504	504	504	504	504	504	504	504	504	504
Validation '23	1512	1512	1512	1512	1512	1512	1512	1512	1512	1512
Test '21	504	504	504	504	504	504	504	504	504	504
Test '22	504	504	504	504	504	504	504	504	504	504
Test '23	1512	1512	1512	1512	1512	1512	1512	1512	1512	1512

Table B.4: Table with numerical descriptions of the used datasets, after pair generation performed in Section 3.2 (with a Δn of only 24 hours). Due to the overlapping nature of the pair generation algorithm, "more" usable data was generated compared to the original data. The amount of pairs P is displayed, the total amount of hours, total datapoints, datapoints passed through the model as input u, and the ground truth y datapoints used for the loss function during training, giving an indication of the amount of computations needed for one training epoch. (With the "optimal" $\Delta n = 1$, the training set would grow to the impractical amount of 12,847,104 datapoints.)

	P	hrs_{total}	n_{total}	$n_{\mathbf{u}}$	$n_{\mathbf{y}}$
Training set	656	47 232	535 296	472 320	62 976
Validation set	93	6696	75 888	66 960	8928
Testing set	93	6696	75 888	66 960	8928

C Training insights

The subplots in Figure C.1 show the training and validation loss development during final training of the six models. Figure C.2 shows how both the shared and branched part of the HLSTM contributed to its training loss. See Table D.6 for the HLSTM's architecture summary.

Training the models took an hour maximum, using the hyperparameters listed in Table D.1 and D.2 and processed locally on an Intel Core i7-8565U CPU, 8GB RAM, 64-bit OS.

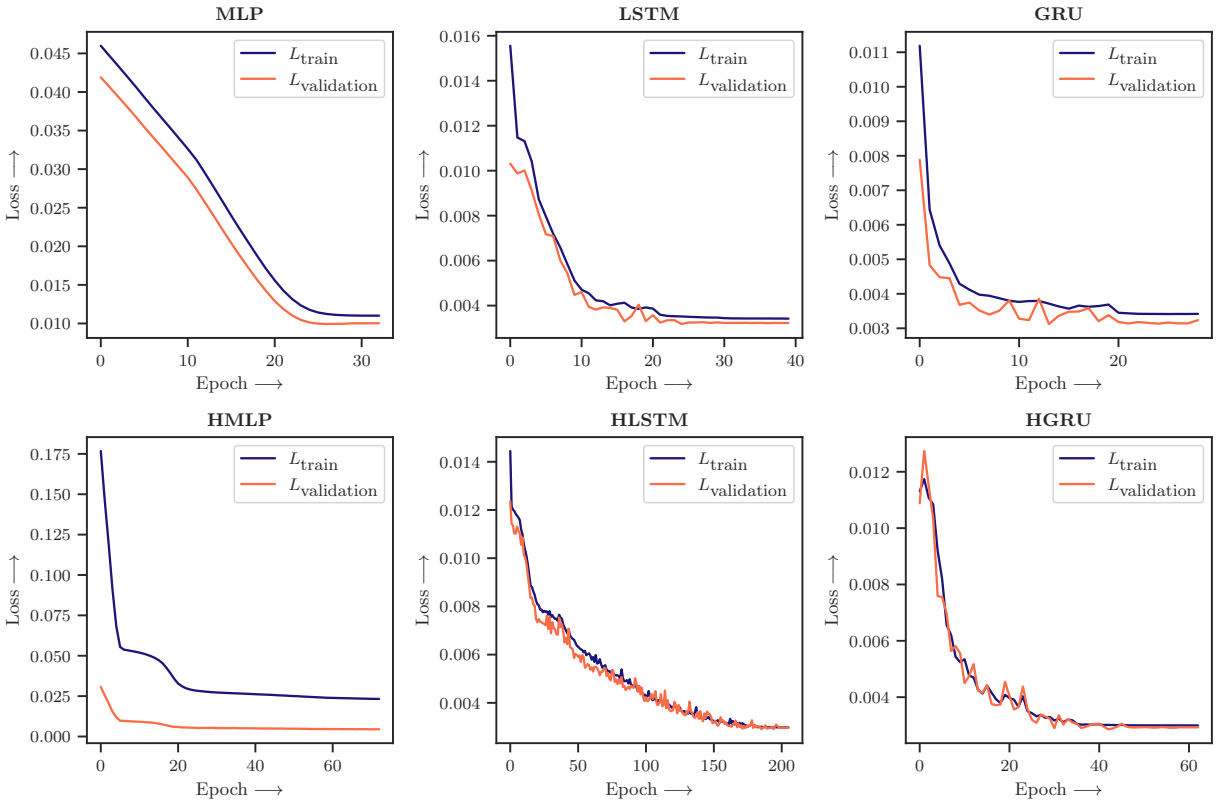


Figure C.1: Loss plots for all models, showing the training versus validation losses over epochs.

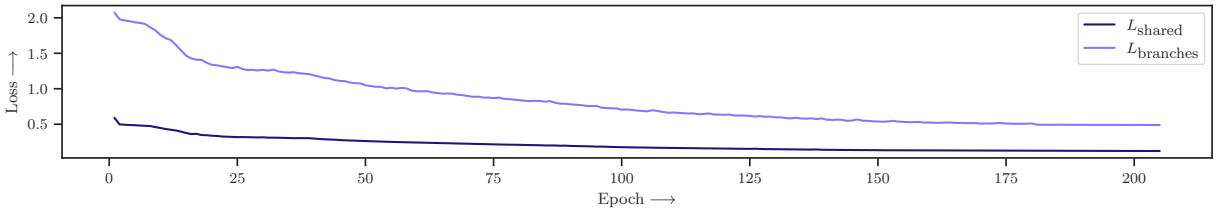


Figure C.2: Training loss plotted of the shared and branched part of the HLSTM. For illustrative purposes, the first epoch is left out from the plot. Both model parts have different complexities (see Section 3.3), causing their learning process to be different as well. The branches were more complex, causing its learning process to be less stable, visible by the small "bumps" in its descend.

D Architecture details

This Appendix section provides some additional detail on the employed architectures by specifying the used hyperparameters and model parameters. In all of the model summaries, pass size denotes the size of a forward/backward pass in megabytes (MB). All activation functions employed are ReLU, including the readout. The high number of parameters for the RNNs are caused by the BPTT procedure, with the number being less for the hierarchical recurrent nets due to reduced parameter sharing.

Table D.1: Overview of the hyperparameters that were determined through grid search and consequently used in the models. Their abbrevations are listed in Table 3.3. Also, note that the fully-connected models have just one learning rate, while the hierarchical models have two: one for their shared layer and one for the optimizers of each of their branches. Another thing to note here is that the ratio between these two, μ_{shared} and μ_{branch}, is equivalent to k. This was done, after lots of test runs, with the idea of a "power ratio" between the two: the branches needed a higher μ to let them converge in harmony with the shared layer. Another reason was that by interlinking the two, H was significantly reduced. A last thing to note is λ of the HLSTM being zero. During training, the HLSTM struggled to get momentum and to start learning, resulting in the hyperparameter search choosing a model with optimal flexibility—a regularisation term of zero.

	k	L^{κ}	μ	μ_{shared}	μ_{branch}	λ
MLP	4	64	$1e{-}5$			$1e{-}5$
HMLP	7	64		$1e{-}4$	$7e{-}4$	$1e{-}5$
LSTM	6	112	$1e{-}3$			$1e{-}6$
HLSTM	7	48		$1e{-}4$	$7e{-}4$	0
GRU	4	128	$1e{-}3$			$1e{-}5$
HGRU	4	64		$1e{-}3$	$4e{-}3$	$1e{-}7$

Table D.2: Overview of other training settings (or "hyperparameters") that were determined through trial-and-error (and not through exhaustive search). All models used the Adam optimizer, reduced their learning rates when the validation loss reached a plateau (ReduceLROnPlateau), had a batch size ($|B|$) of 16, and used $k = 5$ in their k-fold cross-validation schemes. The MLPs had a patience of 6 and the RNNs of 15 to accomodate for their differences in convergence speed.

	optimizer	$\mu_{\text{scheduler}}$	patience	$\lvert B\rvert$	k_{folds}
MLP	Adam	ReduceLROnPlateau	6	16	5
HMLP	Adam	ReduceLROnPlateau	6	16	5
LSTM	Adam	ReduceLROnPlateau	15	16	5
HLSTM	Adam	ReduceLROnPlateau	15	16	5
GRU	Adam	ReduceLROnPlateau	15	16	5
HGRU	Adam	ReduceLROnPlateau	15	16	5

Table D.3: MLP architecture summary.

#	Type	Output shape	Param #
	Input		0
1	Linear	[16, 64]	704
	ReLU	[16, 64]	–
2	Linear	[16, 64]	4,160
	ReLU	[16, 64]	–
3	Linear	[16, 64]	4,160
	ReLU	[16, 64]	–
4	Linear	[16, 64]	4,160
	ReLU	[16, 64]	–
5	Linear	[16, 64]	4,160
	ReLU	[16, 64]	–
6	Linear	[16, 4]	260
$\sum$			17,604
Input size			0.00 MB
Pass size			0.04 MB
Params size			0.02 MB

Table D.4: Hierarchical MLP architecture summary. At 3-1, the branches start, indicated by '*4'.

#	Type	Output shape	Param #
	Input		0
1	Linear	[16, 64]	704
	ReLU	[16, 64]	–
2	Linear	[16, 64]	4,160
	ReLU	[16, 64]	–
3-1 * 4	—Linear	[16, 16]	1040 * 4
	—ReLU	[16, 16]	–
3-2 * 4	—Linear	[16, 16]	272 * 4
	—ReLU	[16, 16]	–
3-3 * 4	—Linear	[16, 16]	272 * 4
	—ReLU	[16, 16]	–
3-4 * 4	—Linear	[16, 16]	272 * 4
	—ReLU	[16, 16]	–
3-5 * 4	—Linear	[16, 16]	272 * 4
	—ReLU	[16, 16]	–
3-6 * 4	—Linear	[16, 16]	272 * 4
	—ReLU	[16, 16]	–
3-7 * 4	—Linear	[16, 1]	17 * 4
$\sum$			15,620
Input size			0.00 MB
Pass size			1.12 MB
Params size			0.06 MB

Table D.5: LSTM architecture summary. The high amount of parameters is caused by the unfolding of the network over 72 timesteps.

#	Type	Output shape	Param #
	Input		0
	LSTM	[16, 24, 4]	–
1-1	—LSTM	[16, 72, 112]	561,792
1-2 * 24	—Linear	[16, 4]	452 * 24
$\sum$			572,640
Input size			0.05 MB
Pass size			1.04 MB
Params size			2.25 MB

Table D.6: Hierarchical LSTM architecture summary. At 3-1, the branches start, indicated by '*4'.

#	Type	Output shape	Param #
	Input		0
	LSTM	[16, 24, 1]	–
1	—LSTM	[16, 72, 48]	11,520
2	—LSTM	[16, 72, 48]	18,816
3-1 * 4	——LSTM	[16, 24, 1]	10,477 * 4
$\sum$			72,244
Input size			0.05 MB
Pass size			1.19 MB
Params size			0.29 MB

Table D.7: GRU architecture summary.

#	Type	Output shape	Param #
	Input		0
	GRU	[16, 24, 4]	–
1-1	—GRU	[16, 72, 128]	350,976
1-2 * 24	—Linear	[16, 4]	516 * 24
$\sum$			363,360
Input size			0.05 MB
Pass size			1.19 MB
Params size			1.41 MB

Table D.8: Hierarchical GRU architecture summary. At 3-1, the branches start, indicated by '*4'.

#	Type	Output shape	Param #
	Input		0
	GRU	[16, 24, 1]	–
1	—GRU	[16, 72, 64]	14,592
2	—GRU	[16, 72, 64]	24,960
3-1 * 4	——GRU	[16, 24, 1]	8,849 * 4
$\sum$			74,948
Input size			0.05 MB
Pass size			1.59 MB
Params size			0.30 MB

E Additional results

This section visualises the models' pollutant forecasting capabilities on out-of-sample data by:

- one pure 24-hour forecast of all models;
- one combined plot for two weeks, i.e. fourteen concatenated 24-hour forecasts, with all models;
- six individual plots, to further clarify each model's distinct performance; and
- four scatterplots giving an overview of each model's performance on the four subtasks.

In addition, paired t-tests are used to determine significant differences in the models' performance metrics (which can be found in Section 4, Table 4.1).

A thing to point out is that in the two-week plots for NO_2, a linearly interpolated part of the testing data is visible starting around a week into the plot, at December 16th, 2019.

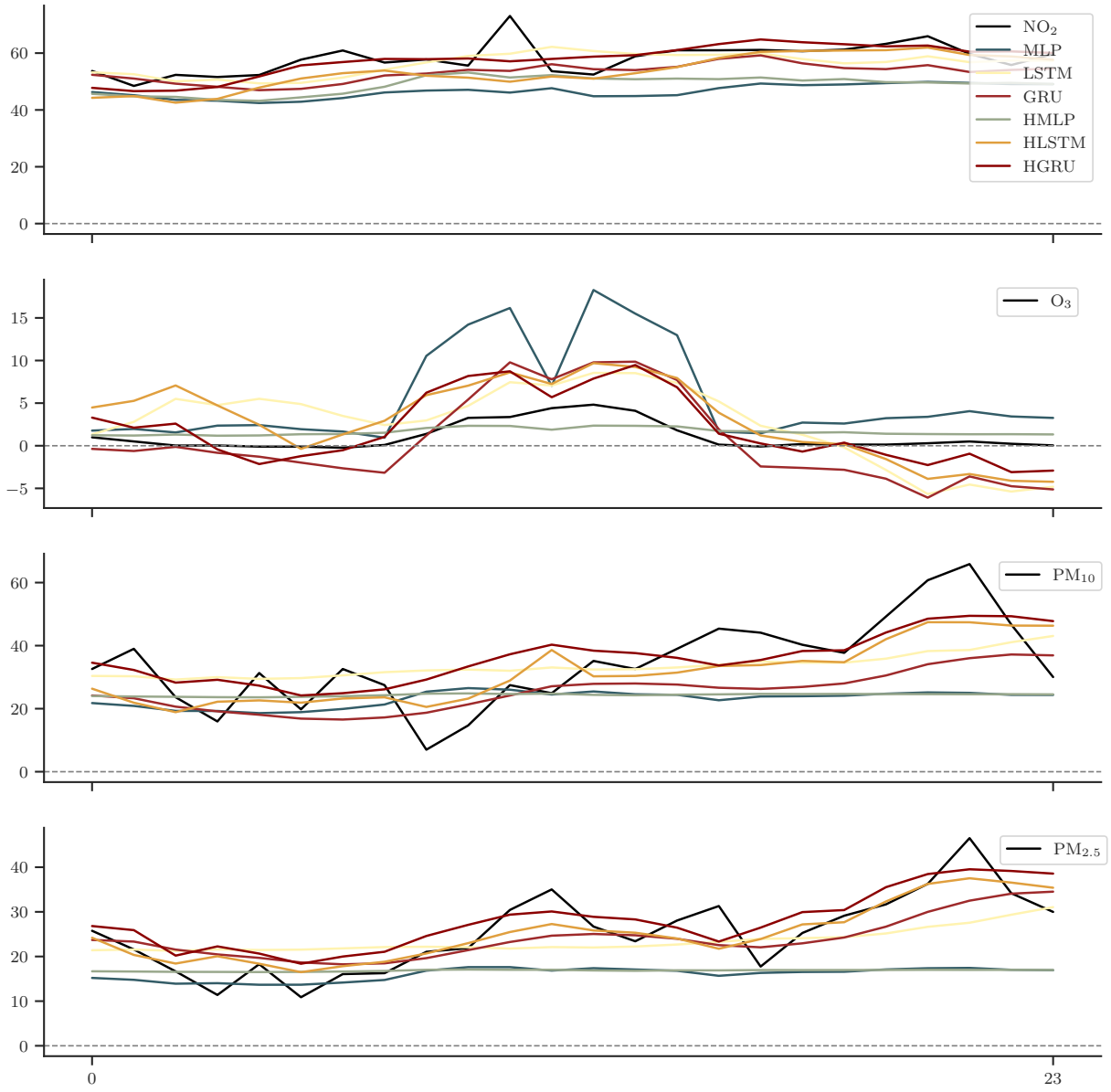


Figure E.1: Forecast for one day (December 29th, 2021) taken from the evaluation set for all models. Black indicates the ground truth and each colour, see the legend, the forecasts. A two-week forecast of all models is depicted in Figure E.2, individual forecasts are presented in Figure E.3, E.4, E.5, E.6, E.7, and E.8, and the corresponding numerical metrics are visible in Table 4.1.

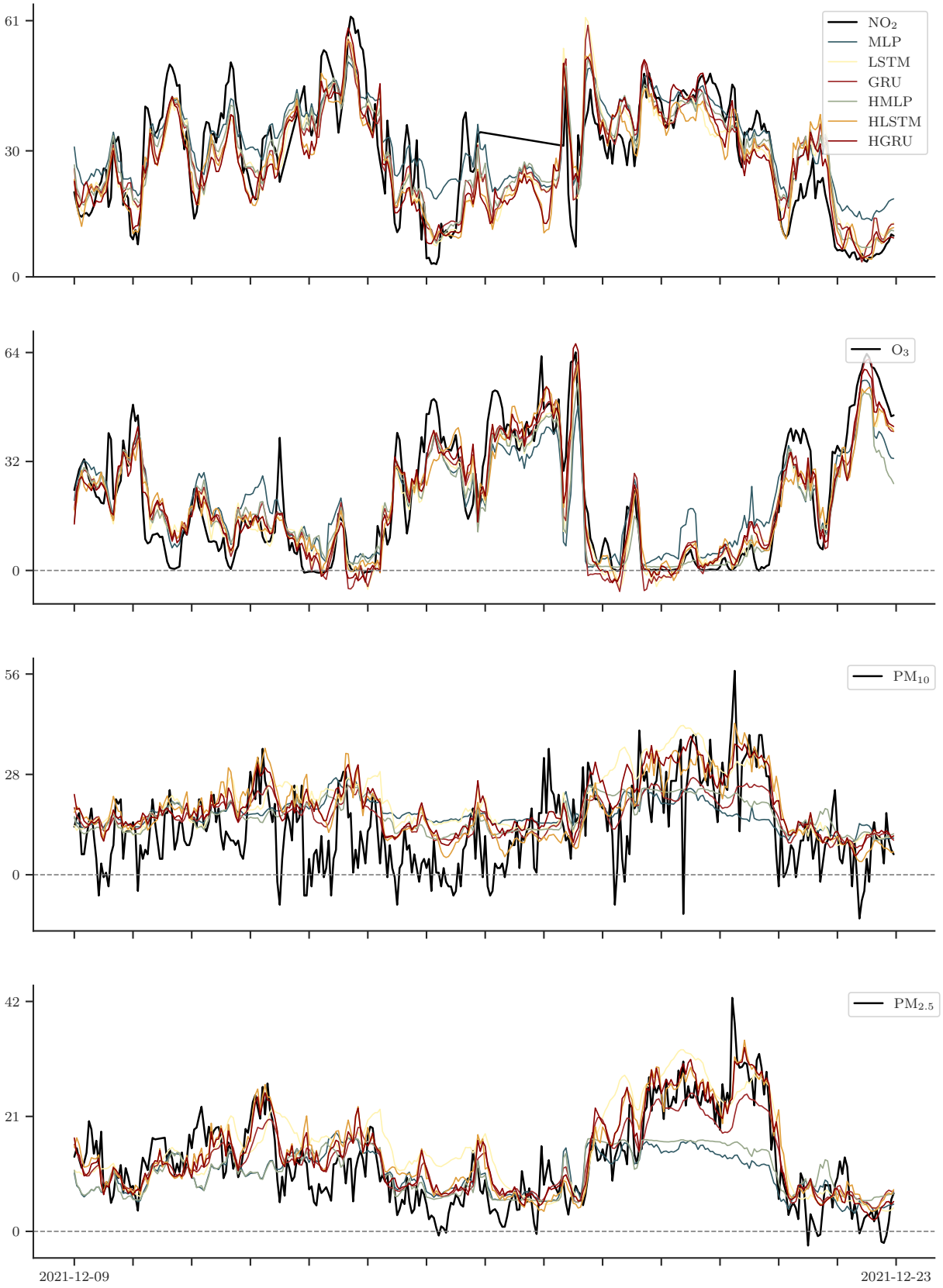


Figure E.2: Forecasts for two weeks (= fourteen 24-hour windows) from the evaluation set, for all models. Black indicates the ground truth and each colour, see the legend, a forecast. Around the middle of the NO_2 ground truth, a linearly interpolated part of the data is visible. Individual forecasts are presented in Figure E.3, E.4, E.5, E.6, E.7, and E.8, and the corresponding numerical metrics are visible in Table 4.1.

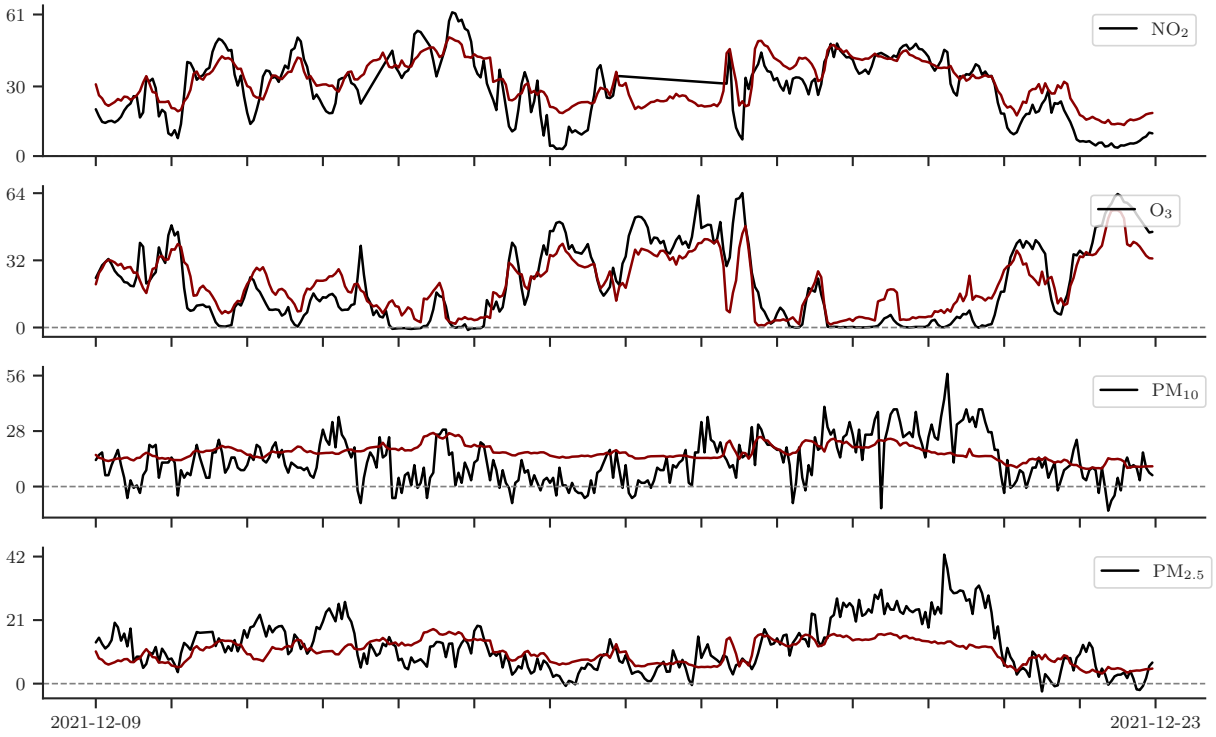


Figure E.3: MLP forecasts for NO_2, O_3, PM_{10}, and $PM_{2.5}$. Black indicates the ground truth and maroon the forecasts. The MLP visibly underperforms with PM, showing a bland line.

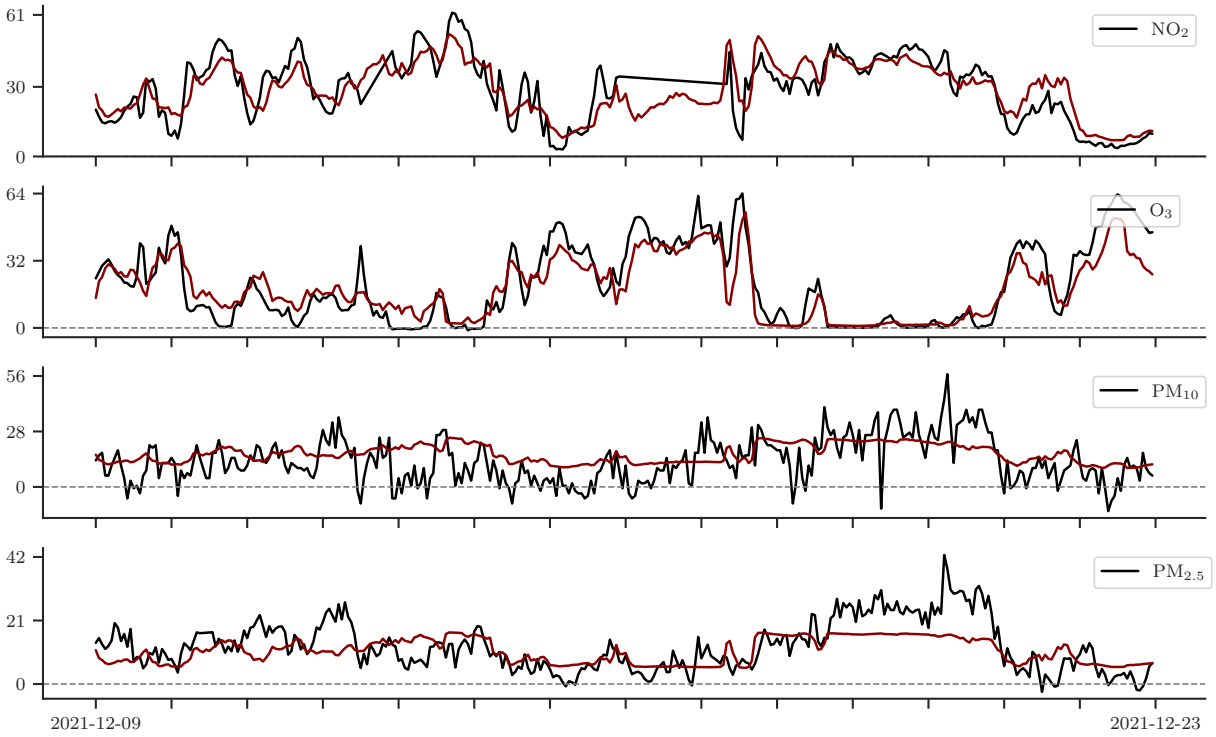


Figure E.4: HMLP forecasts for the pollutants. Black indicates ground truths and maroon forecasts. According to the numbers in Table 4.1, the HMLP performs best, comparatively seen, on $PM_{2.5}$. This is contrary to what is observed in this sample, where the forecast fails to grasp the pattern.

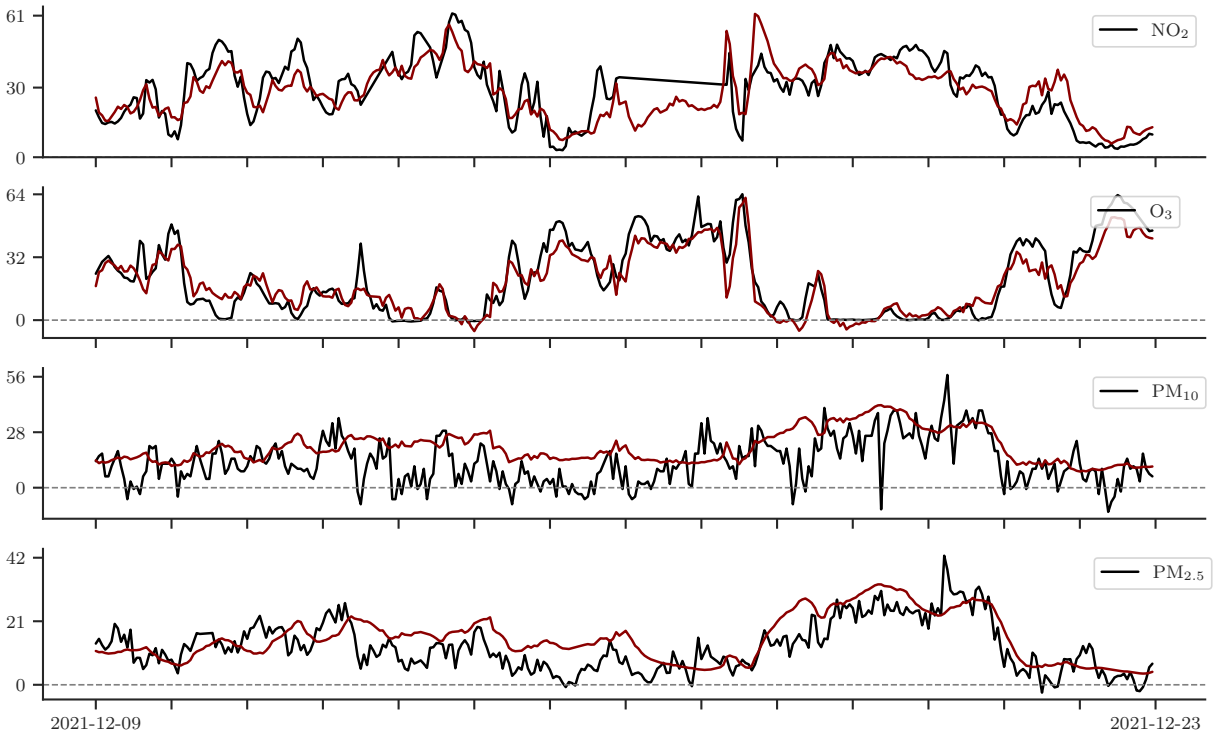


Figure E.5: LSTM forecasts for NO_2, O_3, PM_{10}, and $PM_{2.5}$ taken for two weeks (= fourteen 24-hour windows) from the evaluation set. Black indicates the ground truth and maroon the corresponding forecasts for each. It follows the ground truth to a fair degree with most, and seems to have the most trouble with PM_{10}.

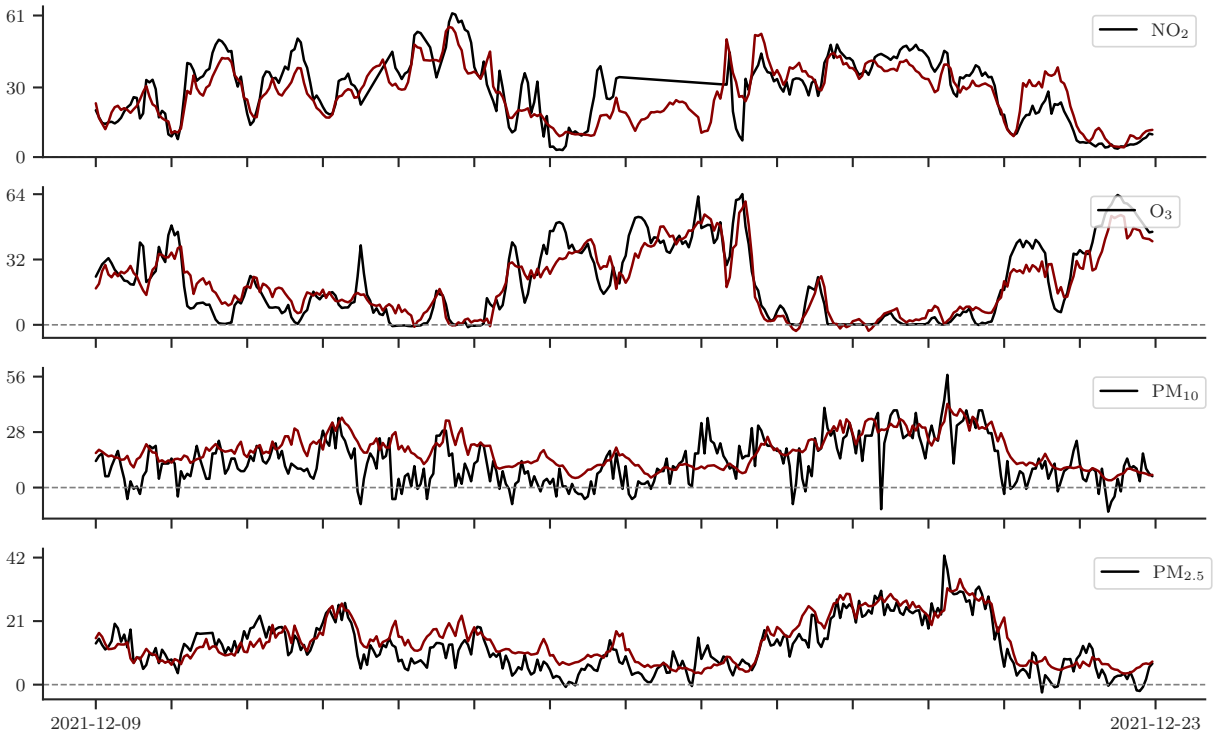


Figure E.6: HLSTM forecasts for NO_2, O_3, PM_{10}, and $PM_{2.5}$ taken for two weeks (= fourteen 24-hour windows) from the evaluation set. Black indicates the ground truth and maroon the corresponding forecasts for each. The HLSTM's runner-up performance, see Table 4.1, shows. Of all the RNNs, it used the least amount of parameters (72,244). It performed best with NO_2.

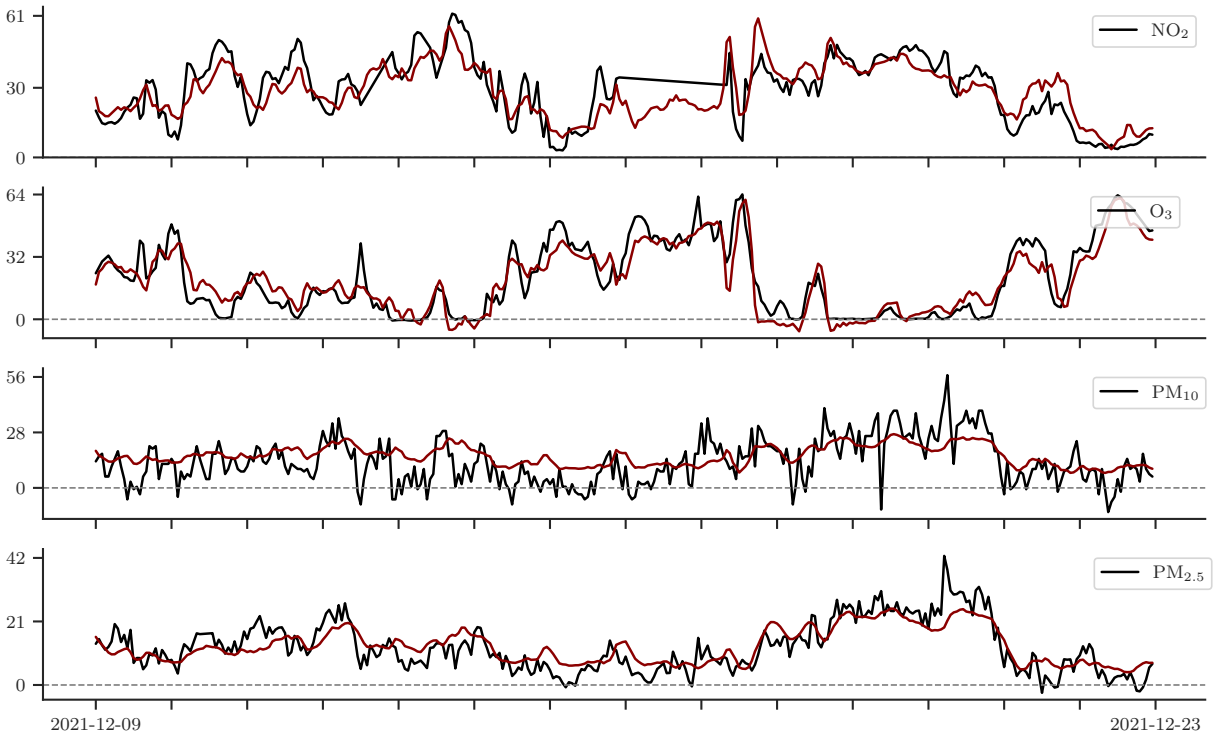


Figure E.7: GRU forecasts for NO_2, O_3, PM_{10}, and $PM_{2.5}$ taken for two weeks (= fourteen 24-hour windows) from the evaluation set. Black indicates the ground truth and maroon the corresponding forecasts for each. It performs generally well. Figure 4.1 showed how all models have a slight negative bias. Here, although not obvious, the NO_2 prediction resembles that notion.

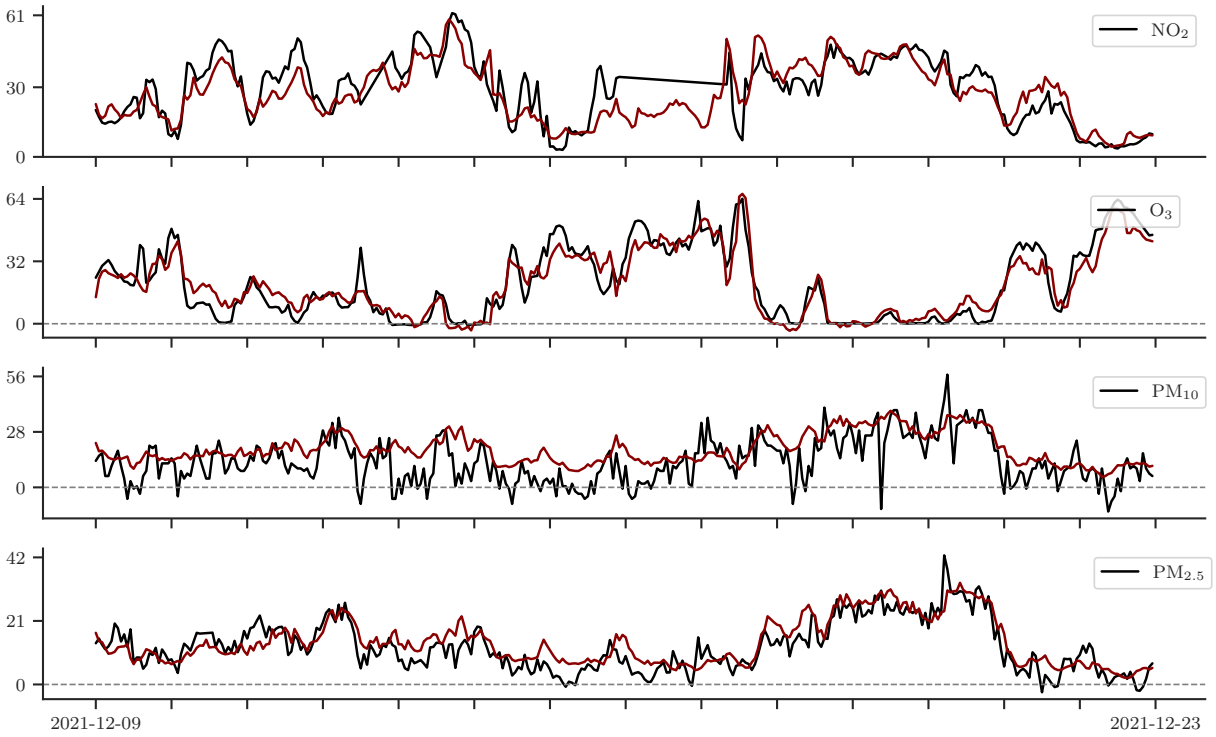


Figure E.8: HGRU forecasts for NO_2, O_3, PM_{10}, and $PM_{2.5}$ taken for two weeks (= fourteen 24-hour windows) from the evaluation set. Black indicates the ground truth and maroon the corresponding forecasts for each. Another HGRU forecast is depicted in Figure 4.2. Of the six considered models, the HGRU achieved the lowest average RMSE and sMAPE, see Table 4.1.

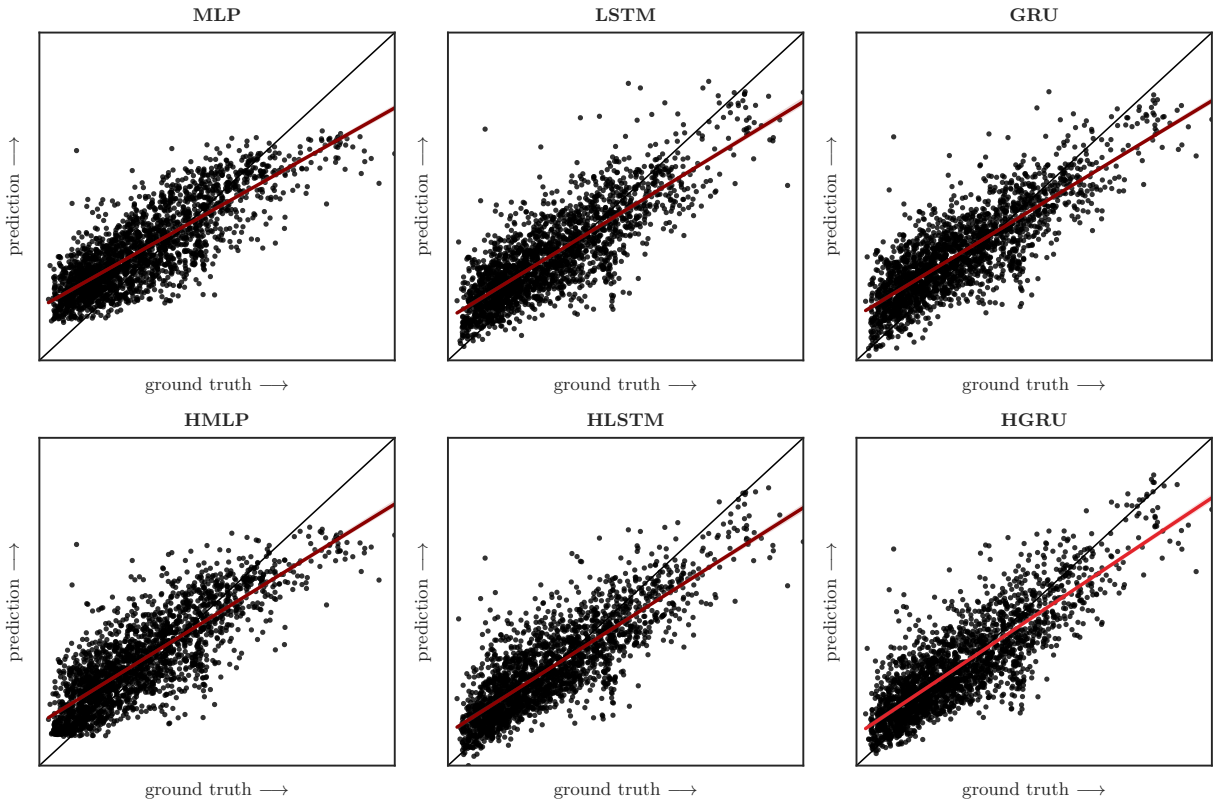


Figure E.9: NO_2's forecasts versus actual observations scattered. Precisely correct predictions are along the diagonal. The maroon trend line gives a visual indication of performance. As indicated by the red line, the HGRU performs optimally.

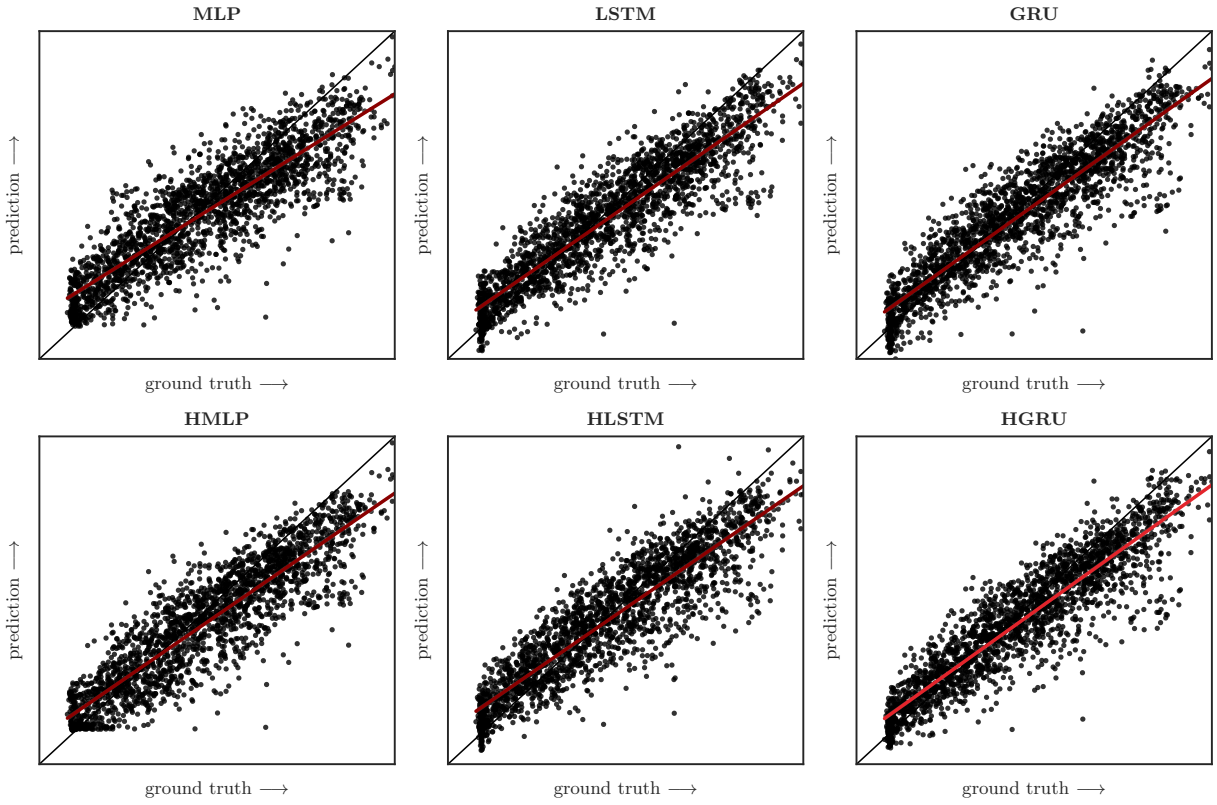


Figure E.10: O_3's forecasts versus actual observations scattered. Precisely correct predictions are along the diagonal. The maroon trend line gives a visual indication of performance. As indicated by the red line, the HGRU performs optimally.

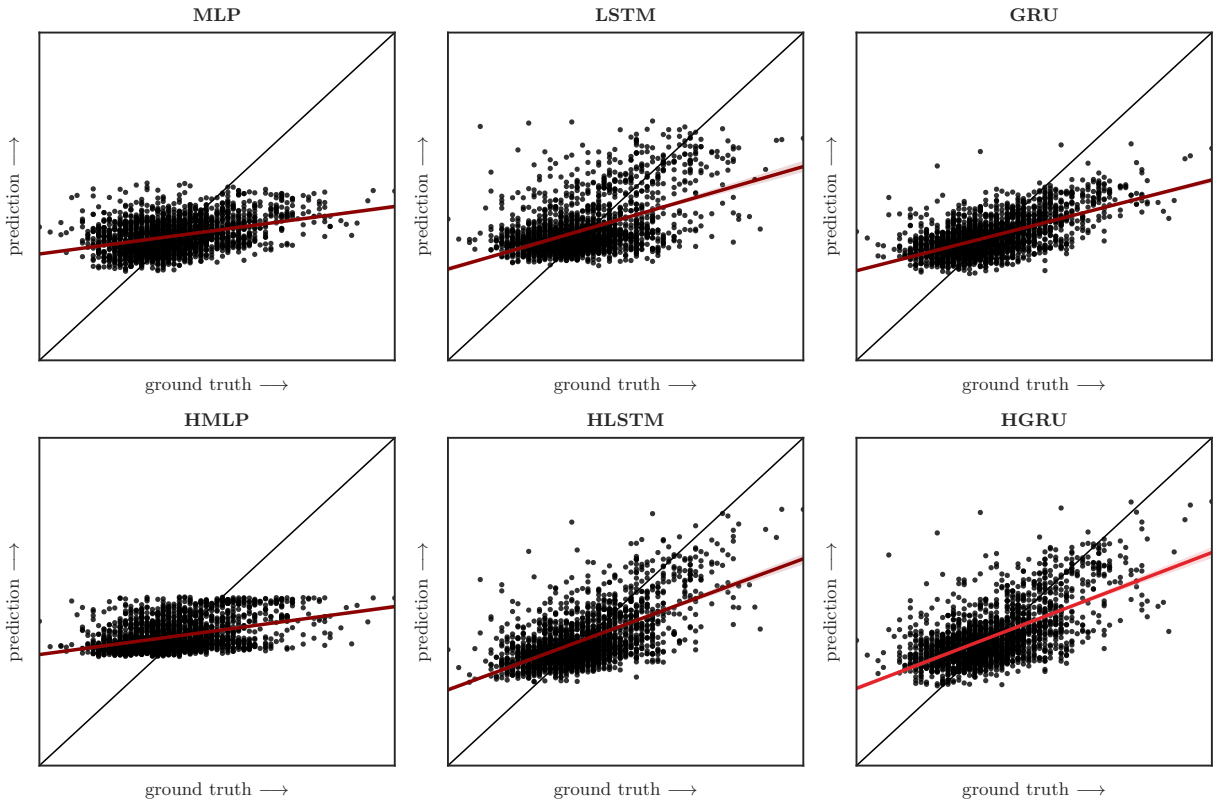


Figure E.11: PM_{10}'s forecasts versus actual observations scattered. The maroon trend line gives a visual indication of performance. As indicated by the red line, the HGRU performs optimally. PM_{10} predictions are off the most of the four pollutants, see Table 4.1 and Figure E.9, E.10, E.12.

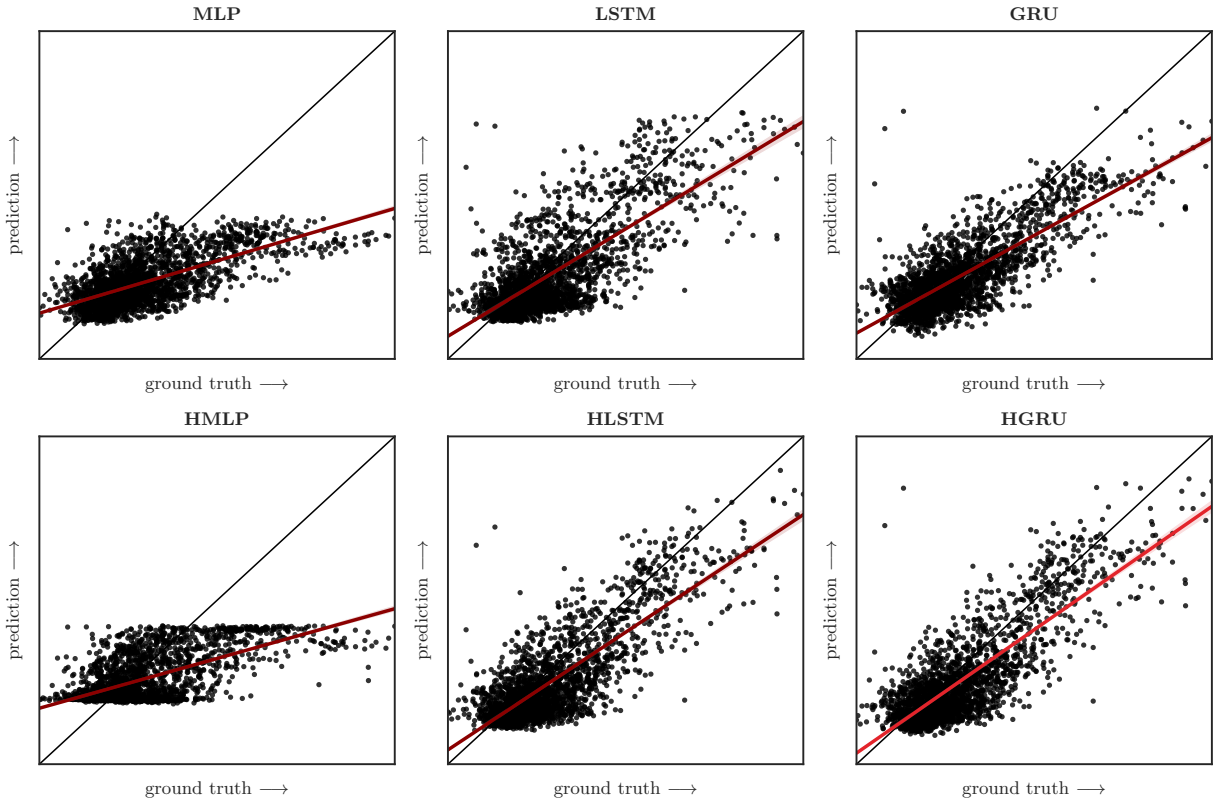


Figure E.12: $PM_{2.5}$'s forecasts versus actual observations scattered. Precisely correct predictions are along the diagonal. The maroon trend line gives a visual indication of performance. As indicated by the red line, the HGRU performs optimally. The contrast between models is most evident here.

Table E.1: Mean (M) and standard deviation (SD) of each model's RMSE and sMAPE based on the testing set, relevant for the paired t-tests conducted and shown in Table E.2 and E.3.

		MLP	HMLP	LSTM	HLSTM	GRU	HGRU
RMSE	M	6.709	6.348	6.040	5.633	5.743	5.468
	SD	5.668	5.559	5.250	4.935	5.004	4.906
sMAPE	M	49.044	46.274	46.321	44.981	45.474	44.519
	SD	46.160	45.344	46.515	45.850	46.969	46.453

Table E.2: Results of paired t-tests assessing statistical significance between RMSE scores for the six models. (Description and definition of RMSE is detailed in Section 3.5, its results in Table 4.1, and its mean and standard deviation in Table E.1). The sample size, degrees of freedom, or n, is equal to 8927 (hours) for all. All p-values are below the significance level of $\alpha = .05$, leading to rejection of the null hypothesis in all cases. Notably, the combination of MLP and HGRU, the benchmark model and best performing model (see Section 4), yields the smallest p-value, suggesting more pronounced evidence to reject the null hypothesis of no difference in performance (based on RMSE) between the pair compared to the others—which is consistent with the expectations.

		MLP	HMLP	LSTM	HLSTM	GRU	HGRU
MLP	$t(n) =$		13.270	14.223	19.762	21.838	22.863
	$p =$		$8.35e{-40}$	$2.09e{-45}$	$4.32e{-85}$	$5.17e{-103}$	$1.91e{-112}$
HMLP	$t(n) =$	−13.270		5.985	13.036	13.617	16.483
	$p =$	$8.35e{-40}$		$2.25e{-9}$	$1.73e{-38}$	$8.38e{-42}$	$3.84e{-60}$
LSTM	$t(n) =$	−14.223	−5.985		11.392	9.235	16.035
	$p =$	$2.09e{-45}$	$2.25e{-9}$		$7.41e{-30}$	$3.18e{-20}$	$4.63e{-57}$
HLSTM	$t(n) =$	−19.762	−13.036	−11.392		−4.177	5.922
	$p =$	$4.32e{-85}$	$1.73e{-38}$	$7.41e{-30}$		$2.98e{-5}$	$3.30e{-9}$
GRU	$t(n) =$	−21.838	−13.617	−9.235	4.177		9.389
	$p =$	$5.17e{-103}$	$8.38e{-42}$	$3.18e{-20}$	$2.98e{-5}$		$7.60e{-21}$
HGRU	$t(n) =$	−22.863	−16.483	−16.035	−5.922	−9.389	
	$p =$	$1.91e{-112}$	$3.84e{-60}$	$4.63e{-57}$	$3.30e{-9}$	$7.60e{-21}$	

Table E.3: Results of paired t-tests assessing statistical significance between sMAPE scores for the six models. (sMAPE is defined as (3.9), results are in Table 4.1, and its mean and std can be found in Table E.1). The degrees of freedom n is equal to 8927 (hours) for all. All p-values, except for one, are below the significance level of $\alpha = .05$. The combination of HMLP and LSTM sMAPE scores did not yield sufficient evidence to reject the null hypothesis. Similar to the paired t-test results for RMSE in Table E.2, the lowest p-value is observed for the combination of the benchmark and best performing model, the MLP and HGRU.

		MLP	HMLP	LSTM	HLSTM	GRU	HGRU
MLP	$t(n) =$		12.775	11.888	17.919	13.313	19.845
	$p =$		$4.81e{-37}$	$2.41e{-32}$	$1.47e{-70}$	$4.71e{-40}$	$8.79e{-86}$
HMLP	$t(n) =$	−12.775		1.011	8.666	3.302	10.619
	$p =$	$4.81e{-37}$		$3.12e{-1}$	$5.28e{-18}$	$9.65e{-4}$	$3.49e{-26}$
LSTM	$t(n) =$	−11.888	−1.011		9.721	3.126	12.346
	$p =$	$2.41e{-32}$	$3.12e{-1}$		$3.18e{-22}$	$1.78e{-3}$	$9.86e{-35}$
HLSTM	$t(n) =$	−17.919	−8.666	−9.721		−5.854	2.855
	$p =$	$1.47e{-70}$	$5.28e{-18}$	$3.18e{-22}$		$4.97e{-9}$	$4.32e{-3}$
GRU	$t(n) =$	−13.313	−3.302	−3.126	5.854		9.298
	$p =$	$4.71e{-40}$	$9.65e{-4}$	$1.78e{-3}$	$4.97e{-9}$		$1.77e{-20}$
HGRU	$t(n) =$	−19.845	−10.619	−12.346	−2.855	−9.298	
	$p =$	$8.79e{-86}$	$3.49e{-26}$	$9.86e{-35}$	$4.32e{-3}$	$1.77e{-20}$	