UNIVERSITY OF GLASGOW

High resolution air quality modelling and prediction

by

Yoana Borisova Napier

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

College of Science and Engineering

School of Mathematics and Statistics

April 2022

Declaration of Authorship

I, Yoana Borisova Napier, declare that this thesis titled, 'High resolution air quality modelling and prediction' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

Air pollution is one of the leading world problems. Across the world, many organizations are in charge of researching safe levels of air pollution, which do not affect people's health. This research has resulted in regulations, which in Scotland are set by the Scottish government. However, monitoring air pollution is very expensive which leads to sparsity in the data. This thesis aims to address this issue by investigating the miniature automated sensor (MAS) networks and the emulation of air quality models data. MAS are a cheaper alternative to the current air quality monitoring stations. Therefore, the quality of the measurements from MAS (in a realistic for citizen science application) are assessed using Bland-Altman analysis and compared to the air quality monitoring stations' recordings using linear regression models. It is found that the MAS do not have the required level of accuracy, although their recordings are significantly capturing the pollutants' concentrations' fluctuations.

Alternatively, in order to assess the effect of unobserved meteorological conditions on pollutants' concentrations, simulated data from ADMS-Urban for Scotland is used. Based on single station and multiple station Gaussian Process (GP) models, emulators for the NO_2 annual average are produced and used to identify the meteorological conditions for which the regulations will be breached. Therefore, a variety of measures can be set in motion when such conditions occur to prevent a breach of the regulation. A quasi-Poisson generalised linear model (GLM) is used to emulate the number of NO_2 hourly exceedances in a year over the regulatory limit of 200 μg m^{-3}, thus identifying the meteorological conditions for which the regulations will be breached and for measures preventing the breaches to be placed. To emulate the yearly time series for NO_2 hourly concentrations, a hyperspatial-temporal emulator with a block-design matrix is proposed. In order to improve the computational speed, the emulator is produced for overlapping blocks of data for periods of interest. The results from the emulator identified periods of possible high NO_2 hourly pollutant concentrations and allowed to identify the emissions levels and meteorological conditions, which lead to high hourly NO_2 concentrations. Overall, all proposed emulators have very good out-of-sample performance in predicting the simulated data.

Acknowledgements

First and foremost, I would like to thank Prof. Marian Scott and Prof. Duncan Lee who guided me through every challenge and gave me every opportunity to succeed.

Big thank you to Dr. Francesco Finazzi, Prof. Alessandro Fassó and all the statisticians and mathematicians in the Engineering department in the University of Bergamo for all the advice and the numerous lunches.

I would also like to give special thanks to Dr. Alan Hills and the other members of the Meteorology and Oceanography team in SEPA without whom this work would have never happened.

A huge thank you to Ciara, Dimitra, Jafet, Kate, Laura, Sebastián, Umberto, and all my friends and officemates for all the advice and support through the years.

Special thanks to Prof. John McColl for the wise advise to stay in University of Glasgow, to Prof. Bernard Torsney for the jokes, and to all other members of staff at the University of Glasgow.

Last but not least, I would like to thank my family. This work would not have been possible without their unconditional support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and current legislation

Air pollution is one of the most tangible world problems. Global organizations like the World Health Organization (WHO), through continental organizations such as the European Union (EU) and national agencies like the Scottish Environment Protection Agency (SEPA) are all researching air pollution, its effects on people and their health, and ways to reduce pollutant concentrations. Depending on their level of influence, each of the aforementioned organizations has advised or even imposed legislation in order to ensure that air pollution is within certain boundaries which are chosen to protect people's health. According to the WHO, every year around 4.2 million people die from diseases related to exposure to air pollution. The 2016 WHO report [208] presents stricter guideline values for what constitutes air pollution because WHO research has found that lower air pollution than was previously believed has a lasting effect on people's health. Furthermore, the report states that over 90% of the world population breathe air that does not comply with the WHO guidelines. In 2016, the 194 WHO Member States issued a roadmap "for an enhanced global response to the adverse health effects of air pollution" [208]. Based on this, the United Nations (UN) has set up targets for the expected world air pollution by 2030 [195].

In 2008, the EU Parliament combined most of its air quality laws into a single directive - Directive 2008/50/EC [76]. According to the Clean Air Handbook, "EU citizens have a legal right to clean air" [11]. By the structure of the EU, each individual state has an obligation to provide clean air (as per EU standards) to its citizens or the citizens of that country have the right to sue the country. Although the EU has set up many regulations on air pollution in order to provide its citizens with clean air, "up to one-third of the EU urban populations are exposed to air pollution which exceeds EU limit values"

[11]. In April 2015, ClientEarth, an environmental lawyers group, won a Supreme Court case against the United Kingdom (UK) for breaching the EU limits for nitrogen dioxide (NO_2) [160]. As a response the UK government proposed a plan to comply with the EU regulations but in November 2016, the plan was ruled as "inadequate" by the High Court in London [161]. According to the 2016 report of the Royal College of Physicians [166], every year there are 40,000 deaths that are attributed to ambient (outdoor) air pollution and a lot more people are affected by illnesses due to air pollution. Furthermore, the report claims that improving the air quality could also result in economic benefits of €3.9 billion per year.

The UK has worked on improving the quality of air in the country for more than fifty years. In December 1952, about 4,000 people died as a direct result of the Great Smog in London [23]. To circumvent the repetition of events with such disastrous impact, Parliament passed the Clean Air Act of 1956 and 1968 which "banned emissions of black smoke and decreed residents of urban areas and operators of factories must convert to smokeless fuels" [129]. The Clean Air Act from 1993 banned all black smoke [147]. Furthermore, the Clean Air Act from 1993 imposed regulations on motor fuels, the sulphur content of oil fuels, established smoke control areas, etc [147]. However, in 2018, in a report by the Royal College of Physicians [167], it was stated that 65% of the British public would support a new Clean Air Act.

Similar legislation with regards to air pollution has been developed by many other countries. The most current findings for the air in Europe are presented in the 2020 report "Air quality in Europe" by the European Environment Agency (EEA) [75]. However, countries outside Europe are also regulating their air pollution. For instance, the United States of America (USA) has regulated air pollution through a Clean Air Act (CAA) which is a "comprehensive federal law that regulates air emissions from stationary and mobile sources" [198]. The US Environmental Protection Agency (EPA) supervises both the regulations and research on air pollution in the USA. The UK's equivalent to the EPA is the Department for Environment, Food and Rural Affairs (DEFRA), which is responsible for the UK strategy on air quality. DEFRA has created the Air Quality Strategy for England, Scotland, Wales and Northern Island in 2007 [58], which has been most recently updated in 2020 [62]. The strategy sets up objectives for applying the EU and international regulations on air quality. DEFRA coordinates the work between the devolved administrations of the UK - England, Scotland, Wales and Northern Island.

In Scotland, the Scottish Government are responsible for developing the "domestic policies and initiatives to improve air quality and reduce risks to human health" [188]. The Scottish Government's propositions are summarised in the "Cleaner Air for Scotland (CAFS) - The Road to a Healthier Future" [187]. SEPA is one of the organizations

which monitor compliance with the regulations and report back to CAFS by suggesting management decisions. As Glasgow is the city with the poorest air quality in Scotland [21], specific attention is paid to this city and its transport system.

1.2 Pollutants

Technical advancement, energy consumption and transport systems have been increasing exponentially over the last few centuries and have become the main reasons for air pollution [154]. This is why the focus of most air pollution research is on the anthropogenically (man-made) emitted pollutants. There are two types of pollutants: primary and secondary. Primary pollutants are emitted directly from a source, whereas secondary pollutants are formed in reactions in the atmosphere. For instance, nitrogen oxides (NO_x) are primary pollutants since they are emitted during combustion processes. A reaction between NO_x and carbon monoxide (CO) produces ozone (O_3), a secondary pollutant. As a result of this, research relating air pollution and its effects examines both the overall effect of all pollutants on people's health (such as [48]) and the specific effect that individual pollutants have on people's health (such as [153]). In the following subsections, a summary of the main pollutants monitored is presented as well as the regulations for these pollutants in Scotland.

1.2.1 Nitrogen oxides

Nitrogen oxides are one of the most examined pollutants and specifically, nitrogen dioxide (NO_2). In general, NO_x is a result of the combustion process of oxidation of nitrogen in fuel and air. The most common source of this emission are engines. These types of emissions are close to the ground and often distributed in densely populated areas [73]. Figure 1.1 shows that in the EU, almost 40% of the nitrogen oxides are attributed to road transport and 16% are attributed to energy production and distribution. NO_x pollution has been linked to causing or worsening both respiratory and heart diseases [154].

1.2.2 Ozone

Particular attention must be paid to ozone. The gas is most famously known as being part of the atmosphere and forming the ozone layer which protects humans from ultraviolet (UV) radiation from the sun. However, ground level ozone is dangerous to people's health as it increases the risk of respiratory diseases [186]. As a secondary pollutant,

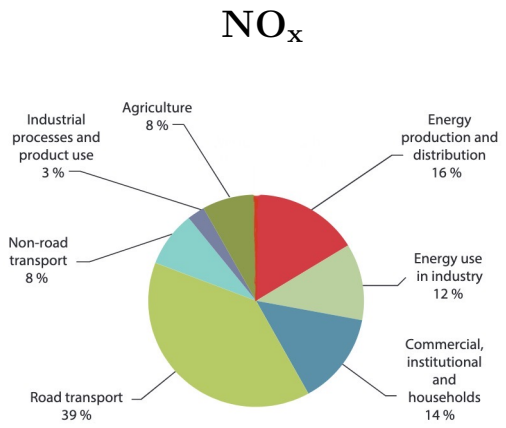

FIGURE 1.1: NO_x emissions in the 28 country members of the EU: share by sector group [74].

ozone's concentration is directly related to the concentration of NO_x and CO, hence, traffic and industrial emissions are considered major contributors to the formation of O_3 [73]. O_3 is formed when oxygen (O_2), carbon compounds known as volatile organic compounds (VOC) and NO_x react in the presence of sunlight [166]. Hence, O_3 levels are highest in the summer when there is more sunlight. The process of forming O_3 may take up to a few days. In this time, the wind tends to carry the compounds away from the urban areas, in which they originated, to the rural areas. Hence, rural areas tend to have higher O_3 concentrations than urban areas [166].

1.2.3 Regulation

Scotland has to comply with the EU and UK wide standards for concentrations of the pollutants. However, Scotland has imposed regulations that are not always the same as the UK ones as can be seen in Table 1.1 where the levels allowed in Scotland and the UK of different pollutants are stipulated. Table 1.1 is a shorter version of the table on the Air Quality in Scotland website [7]. From Table 1.1, it can be seen that Scotland uses the UK wide restrictions for NO_2. On the other hand, the UK has a regulation about O_3 which is not assessed by Scottish local authorities [7]. The differences between the UK and Scotland are a result of the fact that the environmental issues in Scotland are a devolved matter to the Scottish Parliament (and hence the Scottish government) [189].

1.3 Policies to reduce pollutants

The 2018 EU clean air policy [71] outlines examples for measures for reduction of different air pollutants by focusing on reducing the power and heat emissions, the industry

Pollutant	Applies to	Concentration	Measured as	To be achieved by
NO_2	UK	200 μg m^{-3} not to be exceeded more than 18 times a year	1-hour mean	31 December 2005
	UK	40 μg m^{-3}	Annual mean	31 December 2005
O_3	UK excluding Scotland	100 μg m^{-3} not to be exceeded more than 10 times a year	Running 8-hour mean	31 December 2005

TABLE 1.1: National Air Quality Objectives for NO_2 and O_3 [7].

emissions, the agricultural emissions and the transport sector emissions. The EU suggests investments in renewable sources of energy (for instance, solar and wind energy) as well as replacement of many appliances (such as boilers) with newer energy effective devices. The EU provides aid to the Member States to help these changes. Industry emissions are responsible for NO_2 so large industrial installations (for example, power plants) are required to make technical improvements further outlined in the EU Industrial Emissions Directive [77]. In order to reduce the pollution due to agricultural activity, the use of nitrogen fertilisers is limited, new methods for storing manure are implemented, and energy consumption is reduced by using photovoltaic installations or reducing fuel consumption. For both the agricultural and industrial emissions, there are already measures which are proven to reduce the pollution.

However, the EU clean air policy identifies the transport system as the one requiring the largest number of reforms. Transport is credited as a significant contributor for NO_x. The main measures are aimed at technical improvements (promoting cleaner types of transport), behaviour change (car-sharing options) and demand management (urban planning). These changes are key for the implementation of the three Mobility packages of the European Commission [72]. The first package focuses on establishing CO_2 monitoring, reporting of heavy duty vehicles and promoting taxations which are proportional to a distance-based road charging, differentiated according to the environmental performance of all vehicles. The second package establishes a clean vehicles directive (low- and zero-emission vehicles) by introducing new CO_2 emission standards for cars and vans. Such regulations are a step in the right direction but there is evidence that although vehicles pass the standards, there is "considerable error" in the CO_2 emissions [209]. Therefore, alternative engine types should be considered. The most popular alternatives are hydrogen, plug-in hybrids and electrical vehicles. The third mobility package focuses on reducing the CO_2 emissions from heavy duty vehicles as well as the promotion of electric transport. However, all these regulations are non-binding as regulations are dependent on the specifics of the location, where they are implemented.

1.3.1 UK policies

According to an EEA report [73] from 2019, NO_2 is identified as the main pollutant for which monitoring stations indicate that the regulations have been breached in the UK. DEFRA has been addressing this issue and has introduced a UK-wide air quality plan for NO_2 [60]. The plan has identified tackling roadside NO_2 concentrations as key. The government has committed to invest into ultra low emissions vehicles and their support systems, reduction of emissions of public transport (buses and taxis), air quality grants for local authorities, cycling and walking strategies, and improvements in the national railroad network. The aim of the strategy is to make the UK a global leader in air quality and specifies that after leaving the EU, the UK will continue to regulate emissions to deliver improved air quality and healthier environment by supporting cleaner technologies. The rest of this subsection will provide examples of the application of such policies in the UK.

One way of transforming the transport system is the introduction of Low Emission Zones (LEZ). A LEZ was first applied in London in 2008 with further restrictions applied over time. The LEZ covers most of the Greater London Area and it operates 24 hours a day, all year long, without any exceptions. The main idea of the LEZ is to help reduce the pollution caused by older diesel vehicles, i.e. lorries, buses, coaches, large vans and minibuses. Hence, the vehicles that are allowed within the LEZ have to adhere to specific Euro standards [185]. All vehicles have to be registered and pay to enter the LEZ. A fine is applied to vehicles that do not meet the requirements [20]. More recently, in April 2019, an Ultra-LEZ (ULEZ) was introduced in central London. ULEZ has stricter restrictions than LEZ. The ULEZ will be expanded in 2021 which is expected to result in a 30% reduction in NO_x concentrations as well as particle matter (PM) concentrations [200]. A LEZ was introduced in Glasgow in 2018 for local buses only but the zone will be expanded to include all vehicles from 1 June 2023 [91]. Similarly, the Clean Air Zones (CAZ) is a DEFRA initiative which aims at improving the air quality by introducing zones in the city with no emissions. CAZ are similar to ULEZ in terms of their restrictions [1]. The main aim when introducing CAZ is to lower the concentrations of all pollutants, with specific attention being paid to NO_2 and PM. Five cities (Birmingham, Derby, Leeds, Nottingham and Southampton) were expected to start testing the programme in 2020 but were delayed by COVID-19 measures [1].

Another alternative is using different engines from petrol and diesel. Electric vehicles (EV) of transport are most common as they have zero or low emissions [192]. In terms of public transport, York has invested in fully electric public transport which has lead to the UK's first "Clean Air Zone" in 2018 [45]. Furthermore, there are plug-in taxi grant schemes provided by the UK government [94].In terms of private car owners,

EVs are being incentivised by the government using the Plug-in Car Grant as well as help with EV homecharge scheme [93]. Additionally, from 2022, all new homes are required to have charging stations [16]. Another such type of engine are hydrogen engines, which in a sense are very similar to the current petrol and diesel engines as they are all types of internal combustion engines. The main environmental advantage of the hydrogen engines is the fact that their only emission is water [202]. This makes hydrogen vehicles a zero emissions transport. However, there are challenges in producing "eco-friendly" hydrogen with a possible solution described in [202]. A few pilot projects using hydrogen transport have been introduced around the world. For example, Aberdeen has the largest hydrogen bus fleet in Europe [2]. Furthermore, plug-in hybrids (with one internal combustion engine and a battery powered motor [123]) use the electrical engine for short trips within city bounds and the combustion engine for long trips outside of the city bounds. This makes such engines quite suitable for public transport as shown by the 1,700 diesel-electric hybrid buses fleet in London [193].

1.3.2 Scottish policies

As previously stated, the Scottish government can impose different regulation on environmental issues than the rest of the UK [189], thus allowing for the strategies to be adjusted for the different emission sources in Scotland. The Scottish government has stricter regulation than the UK (as seen in Table 1.1). However, the aim is that Scotland will become the first country to introduce in the recommended by WHO standards [208].

The Scottish government started a Clean Air for Scotland (CAFS) initiative in 2015 [68]. CAFS provides a framework for the Scottish government and its partner organizations on how to reduce air pollution. The framework sets transport reforms as the main goal as combustion processes are deemed to be responsible for the largest portion of the air pollution in Scotland. This will be done by "supporting the uptake of low and zero emission fuels and technologies, promoting a modal shift away from the car, through active travel (walking and cycling) and reducing the need to travel" [68]. The strategy was reviewed in 2019 [69] and it was found that for some key pollutants Scotland is complying with both the EU regulation and the stricter WHO regulation but for NO_x, there are still breaches. However, the strategy has a complex structure and has not yet been fully implemented as it requires the co-operation between many governmental and local agencies.

For instance, similarly to the rest of the UK, Scotland has started introducing LEZs into the four major cities (Glasgow, Edinburgh, Aberdeen and Dundee). Glasgow is the first one to introduce the LEZ in 2018 [4]. In Glasgow, the LEZ is implemented gradually

and, currently, is only applicable to local service buses, which are of category Euro 6 [91]. By the end of 2023, the LEZ will be fully implemented with all vehicles having to meet Euro 4 (petrol cars) or Euro 6 (diesel vehicles). It is believed that the LEZ will help to reduce the NO_2 concentrations as well as other combustion related pollutants. Furthermore, there is an added benefit that any public transport used to connect the zones outside the LEZ with the LEZ will comply with the Euro 6 standard ensuring a larger cleaner air benefit even outside the LEZ [91].

Additionally to the LEZ introduction, Glasgow City Council has also began the AVENUES project, which focuses on increasing the pedestrian and cycling space and reducing the street clutter by introducing Intelligent street lighting [90]. Meanwhile, Edinburgh City Council has began organising events as part of the Open Streets movement [22], which limits the vehicle traffic in the city centre but instead offers free cycle hire. Aberdeen City Council has also taken actions to reduce the vehicle traffic in the city centre by improving walking and cycling facilities as well as organising events which are promoting cycling [3]. Furthermore, the city has encouraged clean vehicles using multiple initiatives and created a Grasshopper multi-operator bus ticket to make public transport more attractive.

Although many of the projects undertaken by the City Councils (like the introduction of LEZs) are outlined in the CAFS initiative of the Scottish government, their actual implementation lies with the local City Councils. Therefore, there is a need for agencies to coordinate between the government and the local authorities. One such agency is the aforementioned SEPA, whose main role is to regulate and monitor emissions for specific industrial activities [174]. However, SEPA are also responsible for providing policy and operational advice to the government, the industry and the public, thus working to ensure that the Scottish, UK and EU regulations are observed.

1.4 Air quality measurement and monitoring

In order to establish the level of pollutants in the air, air pollution must be measured. This allows areas with air quality issues to be identified as well as to evaluate the effect of the proposed policies such as the ones presented in Section 1.3. There are different ways of monitoring, and hence, measuring air quality depending on a different number of factors. The following subsections will outline the most common ways to measure and monitor air quality.

1.4.1 Monitoring networks

It is necessary to monitor the air quality on a global scale, given that air pollution is identified as one of the most widespread problems by the WHO [208]. However, there is no global air quality monitoring network and different regions in the world collect data differently. In the USA, the Air Quality System (AQS) was created by the EPA to collect and provide information about the air quality in the USA on different scales [70]. Across the EU, each Member States has their own network for monitoring air quality.

In the UK, monitoring started in the 1950s and 1960s (after the Great Smog, as part of the Clean Air Acts) when the focus was on black smoke and sulphur dioxide (SO_2). In the 1970s, the Automatic Urban and Rural Network (AURN) was developed. The UK air quality data are available online on the National Air Quality Information Archive website `www.airquality.co.uk`. Additionally, Scotland's air quality data can also be accessed from `www.scottishairquality.co.uk`. The AURN network consists of a number of monitors at different locations in both cities and the countryside. Each monitor collects information not only on the concentrations of multiple pollutants but also some meteorological conditions (for instance, ambient temperature, barometric pressure, relative humidity, etc) which are useful for analysis the pollutants' concentrations. The AURN monitoring network consists of different monitors depending on their location. For full details on all types of stations refer to [8] but in this thesis the focus falls on the following types of monitoring stations:

- **Roadside** - the monitoring station is located between 1m of the kerbside of a busy road and the back of the pavement. Typically, this will be within 5m of the road, but could be up to 15m. The main source of pollution at this type of station is local traffic. The air pollution measurements at these stations are used in order to assess the worst case population exposure, evaluate the impacts of vehicle emission controls and determine the impacts of traffic planning/calming schemes.
- **Kerbside** - the monitoring station is located within 1m of the kerbside of a busy road. As with roadside stations, the main source of pollution is local traffic. Besides the three objectives of the roadside station, the kerbside station is also used for identifying vehicle pollution blackspots.
- **Urban Background** - the monitoring station is distanced from sources and, therefore, broadly representative of city-wide background conditions (for instance, urban residential areas). The main source of pollution at this type of station comes from vehicle, commercial and space heating sources. The air pollution measurements at this station are used for trend analysis, urban planning, and traffic and land-use planning.

- **Rural** - the monitoring station is at an open countryside location, in an area of low population density distanced as far as possible from roads, populated and industrial areas. The main source of pollution at this type of station comes from regional long-range transport and urban plume. The air pollution measurements are used for ecosystem impact studies, assessing compliance with critical loads and levels for crops and vegetation, investigating regional and long-range transport and identification of ozone hot spots.

The measurements from the AURN monitoring stations are taken at regular time intervals, usually hourly, for multiple pollutants at a single location. The main advantage of the AURN monitors is that the air pollution measurements taken by them are very accurate. However, the monitors are very expensive to operate. Therefore, not many monitors could be placed. Additionally, it is important to note that not all types of pollutants are recorded at every station but rather pollutants are measured based on the type of monitoring station as described previously. In 2020, there were 150 active monitors in the UK [59] of which 23 were in Scotland. Hence, the AURN network is quite sparse and does not cover in terms of measurements of all pollutants the area of the country. In Scotland, there are additional monitoring stations (identical to the AURN monitors) which together with the AURN ones form a network of 100 stations as shown in Figure 1.2 with the island monitors removed to help with the visibility of the monitors. It is clear that most of the stations are in Central Scotland as the majority of the Scottish population lives in the region. Furthermore, the distance between some of the stations are so small that the stations appear on top of each other. For instance, in Aberdeen, there were 6 active monitors but only 2 are visible on the map. The monitoring stations in Aberdeen and Glasgow will be discussed in further detail in Chapter 4 as the subsequent chapters focus on data from these two locations.

One way to expand the monitoring network is to use lower cost monitors with which members of the public will be able to make air quality observations without any learning time. Lower cost sensors aim at being able to "trace gas measurements to a usable degree of accuracy and precision, and with a stability over time" [116]. For different pollutants the sensors perform differently. Sensors seem to estimate the concentrations for NO and O_3 relatively well but struggle with NO_2 [116]. However, lower cost sensors have not yet reached the level of needed accuracy as it will be demonstrated in Chapter 3. Additionally, the lower cost sensors do not offer such a variety of pollutants to be recorded by a single sensor.

A cheaper alternative to measure the spatial and temporal NO_2 concentrations are diffusion tubes. Diffusion tubes were first introduced in 1976 but since then many improvements have been made. A description on the current way diffusion tubes work is

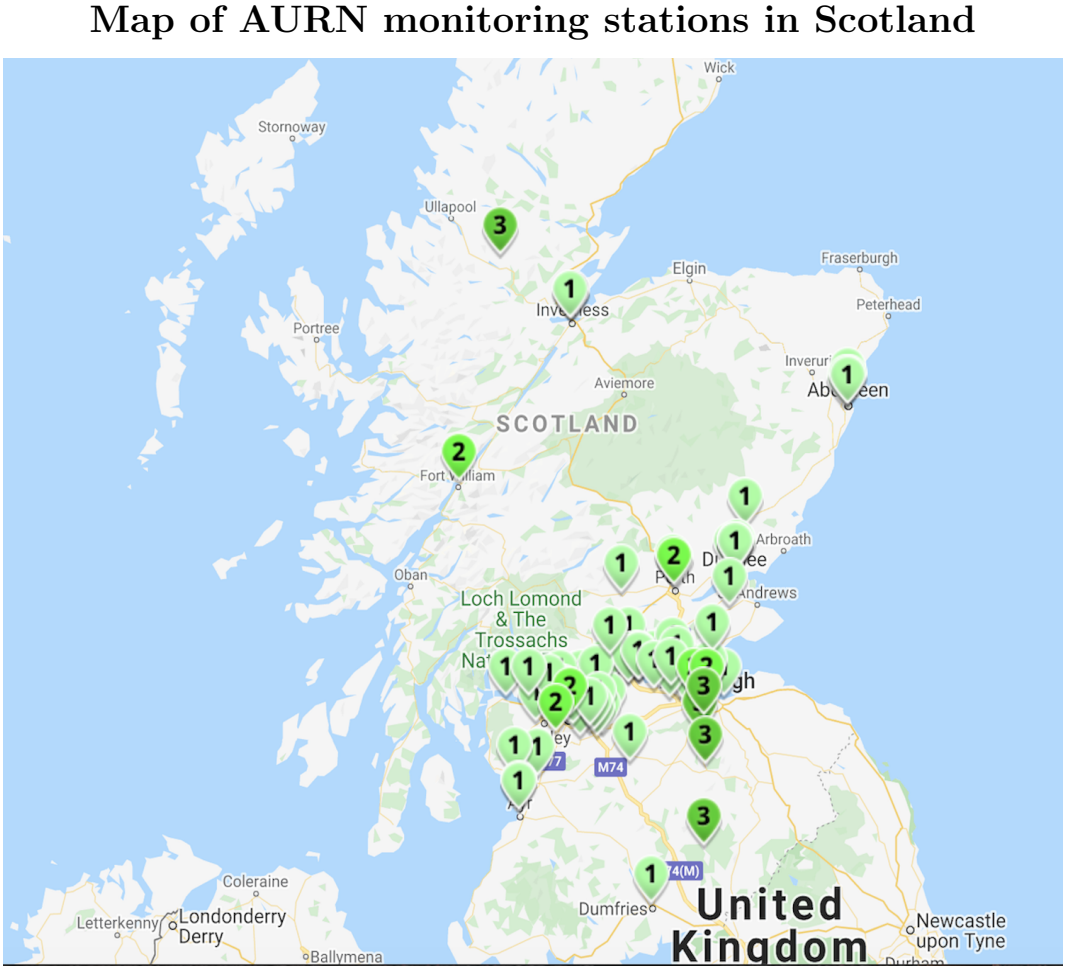

FIGURE 1.2: Map of the AURN monitoring stations across Scotland in 2020 [8]. The numbers on the pinpoints refer to the air pollution levels on a scale 1-10 based on the Daily Air Quality Index (DAQI) [6].

provided in [31]. A major advantage of diffusion tubes is that they are relatively cheap, simple and have relatively small number of errors when collecting data which makes them "sufficiently accurate for assessing exposure and compliance with Air Quality criteria" [31]. The diffusion tubes' measurements are easily biased by the proximity to the source of NO, and therefore the measurements provide an upper limit for the real NO_2 concentrations. Furthermore, there is no standard way of building the tubes. DEFRA uses Palmes-type tubes for outdoor NO_2 measurements and the production of the tubes is described in [184]. An alternative are the Ogawa passive samplers, which can be used to monitor NO_2 in forested areas [54]. Another drawback to using diffusion tubes is that there is variation between the measurements. The diffusion tubes results are impacted by temperature, humidity and wind speed [31], which means that diffusion tubes can give very different readings even though they might be placed close to each other. Hence, DEFRA has produced a manual on the use of diffusion tubes to ensure the readings used in their network are consistent [184].

Another way to measure air pollution is by using satellite data. One such method is aerosol optical depth (AOD). It measures the pollutants concentrations by calculating "the extinction of the solar beam by dust and haze" [197]. However, AOD is more accurate over water rather than land areas [135], thus making it less useful for modelling the air pollution in cities. For more details on AOD refer to [136].

1.4.2 Simulated data from air quality models

Observational data are not always available due to costs even for lower cost sensors. Furthermore, the locations of monitors are not chosen on a grid but rather based on locations of interest such as urban areas. There is a need for a better gridded system and, hence, simulated data from air pollution models are often used. EU Directive (2008/50/EC) [76] stipulates that "where possible modelling techniques should be applied to enable point data to be interpreted in terms of geographical distribution of concentration. This could serve as a basis for calculating the collective exposure of the population living in the area." Hence, the main advantage of this type of simulated data are that simulations under different meteorological conditions (different wind speed, temperatures, etc) or emissions or traffic scenarios can be produced, thus allowing for a better understanding of the underlying processes of air pollution.

The Dutch government developed a number of models under the general name Standard Calculation Method (in Dutch Standaard Reken Methode (SRM)). There are three different versions of SRM. SRM-1, or alternatively referred to as CAR II-model, is used to calculate air quality in urban areas [131]. SRM-2 is developed to calculate the effects along roads and motorways in urban areas [204]. SRM-3 is used to calculate shipping emissions [134].

However, in the UK, DEFRA uses several different models to assess a range of pollutants at different spatial scales, from local to hemispheric in order to comply with different regulations. The models are listed and described in [62] but below a short summary of the most used models is provided. The Pollution Climate Mapping (PCM) model is used for reporting the concentrations of particular pollutants (such as NO_x, NO_2) in the atmosphere based on individual models for each of the pollutants. The Pollution Climate Mapping Model produces background maps of 1×1 km grids of pollutant concentrations in the UK. The Community Multi-scale Air Quality Modelling System (CMAQ) is used to calculate daily air quality forecasts. The Ozone Source Receptor Model (OSRM) is used to advise on the effects of planned or proposed policy on O_3 concentrations to changes in precursor emissions (particularly involving changes in NO_x emissions). The

UK integrated assessment model (UKIAM) is used to investigate cost effective strategies for reducing UK emissions.

Specifically for cities in Scotland, SEPA uses the Atmospheric Dispersion Modelling System (ADMS) developed by the Cambridge Environmental Research Consultants (CERC) [43]. There are several versions of ADMS aimed at different issues. All models take into account the impacts of buildings, complex terrain, coastlines and variations in surface roughness; dry and wet deposition; NO_x chemistry schemes; short term releases (puffs); calculation of fluctuations of concentration on short timescales, odours and condensed plume visibility; and allowance for radioactive decay including γ-ray dose. ADMS 5 [38] focuses on the emissions from existing and proposed industrial installations. The model is used for assessment of simulated air pollution concentrations against air quality regulations, which allows for environmental impact assessments as well as safety and emergency planning. ADMS-Roads [40] focuses on the air pollution problems due to networks of roads that may be in combination with industrial sites. The model is used by local authorities in the UK to review, assess, and develop air pollution action plans and remedial strategies. ADMS-Airport [39] is used for air quality management of airports. ADMS-Screen [41], which models dispersion from a single stack to calculate ground-level concentrations, provides rapid assessments of stack height. The model is used by both governmental and private organisations for quick assessment of the impact of point source emissions. ADMS-Urban [42] is used for modelling air quality in large urban areas, cities and towns. The model is used for assessment of simulated air quality against air quality regulations, developing and testing policy and action plans for air quality improvement such as Clean Air Zones or Low Emission Zones, investigation of air quality management options for the full range of source types including transport sources, provision of detailed street-level air quality forecasts and others. Simulations from ADMS-Urban will be used in Chapters 4, 5, 6, 7 and 8. Therefore, a more detailed review of the model will be provided in Subsection 4.1.3.

1.5 Approaches to modelling air quality measurements

When modelling air quality measurements over a network of stations, the approaches vary. The data can be looked at as a time series at a single location, or the spatial dependencies between measurements at different locations can be compared as well as a spatio-temporal approach which allows the comparison of the times series for air pollutants at multiple locations as well as interpolation across unmonitored locations.

Additionally, the approaches can be applied to single or multiple pollutants. In the following subsections, different modelling approaches for the two types of data (monitoring or simulated) will be discussed both in the frequentist and Bayesian paradigms.

1.5.1 Modelling of monitoring network data

Data from monitoring networks are sparse in design due to the high price of the monitors as previously discussed. Often, the time series from a single station are modelled using ARIMA time series analysis as in [172], [107] and [24]. This allows one to check for periods of high pollution and the autocorrelation between the time steps. These time series data are often regressed against health data to assess the effect of exposures to high pollution levels as in [114], [203], and [210]. However, modelling a single time series does not allow for one to account for the differences between emissions at different locations. Therefore, it is of use to perform spatial analysis using measurements from multiple monitoring stations. Kriging is a common approach to do so as seen in [12] and [151]. In [114] a Bayesian spatial regression model is used to compare the pollution concentrations across four cities (Glasgow, Edinburgh, Aberdeen and Dundee) in Scotland. Alternatively, [115] proposed a Bayesian geostatistical approach allowing for the preferential sampling (placing air quality monitors at locations, where the highest air pollutions concentrations are expected, for more information see [64]) applied to monitoring data from London, UK. Temporal and spatial analysis can be combined in different spatio-temporal approaches to estimate pollutant concentrations over time and space. Similarly, [32] presented a spatio-temporal model for the hourly O_3 concentrations in Harris County, Texas, USA. An alternative frequentist space-time data modelling using hierarchical space-time models is the linear coregionalization model (LCM) first introduced in [164]. The model utilises the spatial relations between the covariates at different locations when missing data are present. One such model within the frequentist paradigm estimated using the Expectation-maximization (EM) algorithm is proposed in [80] and expanded in [81] and applied to NO_2 monitoring data from Northern Italy. A Bayesian approach was presented in [207], which was then adopted for air pollution data independently by [191] and [176], variations of which continue to be applied in [108].

Alternatively, it is of great interest to model data from networks of lower cost monitors. The modelling of lower cost sensors data is usually split into two types - one is assessing the quality of the lower cost measurements, and one is fusing of the lower cost sensors data with measurements from the official monitoring networks. The fusion approaches will be discussed in a subsection below. Assessing the quality of lower cost measurements is often done in comparison to a higher cost monitor as higher cost monitors are assumed to take measurements of the pollutants' concentrations without

error. The monitors have to be compared in both laboratory and outdoor conditions. Reproducibility under field conditions is investigated in [128] using least squares regression and adjusting for correlation in the observations. Agreement between lower cost sensors for multiple pollutants under laboratory conditions is analysed using numerical summaries in [117]. A multiple linear regression is used to compare the measurements from the lower cost monitors to a reference AURN monitor for NO_2, O_3 and $PM_{2.5}$ in [118]. A low cost monitor for O_3 is compared in both laboratory and field conditions in [145] using numerical summaries.

1.5.2 Modelling of simulated data from air quality models

Similarly to the lower cost sensors data, the simulated data compliment the monitored data by lowering the cost of measuring air pollution and allowing for different environmental scenarios to be tested. However, the simulation models need to be compared to the monitoring networks to ensure that the simulations are realistic. The predicted simulated values for O_3 and $PM_{2.5}$ from the CMAQ model are compared to Pacific 2001 measurement data in [181]. Air pollution diffusion simulation was compared to the general air pollution in Taiwan (winter and spring) in [119] using spatial risk analysis and it was found that the simulated data are consistent with the observed data. The results from ADMS-Urban simulation for CO are compared to the air quality monitoring network in [159]. Pearson's product correlation coefficient, normalised mean squared error and fractional bias were used to compare the two types of measurements. It was found that the simulations match "to a greater extent" the observed values, although ADMS-Urban is found to have a tendency to underestimate the CO concentrations, therefore, the simulations require a correction. However, simulated data are more commonly used as a source of additional information in fusion models, which are presented in the following subsection.

1.5.3 Data fusion

Data fusion has been increasingly used for air pollution prediction as it allows combining data from multiple monitoring networks with each other as well as combining monitoring and simulated data. Wald noted in [201] that there is a lack of unified definitions in the field of data fusion. However, in this thesis, data fusion is as defined in [205] as "a process dealing with association, correlation and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats as well as their significance". Using fusion data for global assessment of air pollution is very practical given that there

is no unified world air quality monitoring network. However, in recent years, the fusion data technology has been used to create gridded ambient air pollution estimates that allow for estimating even personal exposure. For instance, [83] fuses ground monitoring data with location data from smartphone applications.

Ground monitoring data of high quality are sparse, but it is often fused with other types of monitoring or simulated data for more information. In [175], ground monitoring data are fused with remote sensing satellite data and chemical transport models to provide global estimation of exposures to ambient air pollution using a data integration model. A geostatistical fusion model is proposed in [146] for the fusing of monitors and diffusion tubes data on NO_2. [152] proposes a Bayesian hierarchical spatio-temporal model for fusing PM_{10} data with ADMS-Urban simulations to assess the short-term exposure to PM_{10} pollution in London. An innovative approach is taken in [155], where data from hourly and annual output of ADMS-Airport are combined with NO_2 pollution measurements from lower cost monitors at London Heathrow Airport to investigate the effects of adding a third runaway. In [100], multiple sensors data are fused with simulated data from Weather Research and Forecasting chemistry and CO_2 models to investigate the effects on CO_2 and $PM_{2.5}$ concentrations of a cold front in Eastern China.

However, often fusing two types of air pollution data is challenging as it requires combining data on different spatial scales. For instance, ground monitoring network (whether or not lower cost) are point referenced, whereas satellite data (such AOD) are areal data. The proposed frequentist spatio-temporal model in [81] can also be used for fusion data, where ground monitoring data for $PM_{2.5}$ is merged with remote sensing data from a satellite. High-frequency hourly AOD data were evaluated against the temporal and spatial variations of simulated AOD data in [211]. Using a Bayesian space-time downscaler approach, [170] fuses O_3 monitoring data with simulation data from CMAQ, whereas [127] applies Bayesian space-time modelling by combing $PM_{2.5}$ monitoring data with CMAQ results. In [25] a Bayesian spatio-temporal downscaler model is used for the fusing of O_3 monitoring data with simulations from CMAQ grid cells.

1.6 Emulation

In Subsection 1.4.2, simulated data were introduced as an alternative to monitoring data as simulated data are easier and cheaper to gather from more locations. However, producing simulated data is time-consuming due to the costly computations that need to be performed. One solution to the problem is to create an *emulator*. This approach requires a number of key runs of the simulation model based on which a surrogate statistical model is built [169]. This model is called an emulator and is used to predict

the simulation output at untried inputs [141]. The following subsections introduce the general theory of emulation followed by some applications of emulators on different types of computer simulations in both the frequentist and Bayesian paradigms. Air pollution emulators will be specifically reviewed.

1.6.1 Background

In order to create an emulator, there are two key features which need to be chosen. The first one is to select a sampling design to choose the simulation scenarios to be run using the simulator model. In [125], Latin Hypercube sampling is proposed as it selects values for inputs. Furthermore, [102] compared several sampling techniques and recommends the use of Latin Hypercube sampling as it is easy to use, it is flexible for multiple settings and has "reliable results". The approach was extended in [132] by allowing for maximin distance designs [109], which ensures that points are not clustered together, and then again in [180], where a partial stratified scheme is introduced. Alternatively, a Sobol sequence sampling scheme [182], which generates uniform quasi-random sequences for multiple parameters in the hyperspace [173], can be used as in [168].

The second feature of building an emulator is predicting the simulator model's output at untried inputs. All emulators (both frequentist and Bayesian) treat the response as a stochastic process and, therefore, assume a Gaussian Process distribution for the response. This allows for "an analytically very tractable form of stochastic process" [140]. The use of Gaussian Processes is discussed in more detail in [158]. A review of the most recent advances and challenges in emulation is presented in [17].

One of the first univariate emulators is proposed in [169], where a statistical model based on kriging is used. Although the proposed model is frequentist, the paper discusses how the model can be adapted to the Bayesian paradigm. Another frequentist univariate emulator based on universal kriging with a choice of different covariance and mean functions is presented in [165], whereas [19] suggests a mixture of frequentists and Bayesian techniques to account for the uncertainty in the inputs. [52] presents a fully Bayesian emulator for computer code output. This model is then extended to handle calibration using a small number of real observations in addition to the simulation model runs in [111], which is continued in [112] and [139]. An emulator that takes both qualitative and quantitative factors is proposed in [156]. The emulator in [156] can be applied in both the frequentists and Bayesian paradigms. The diagnostics for emulator models are discussed in [18]. In [85], an R Shiny application to visualise simple cases of Bayesian emulators is presented.

However, as a simulator can be used to produce multivariate dynamic output such as functional data or multiple time series, there is a need for multivariate emulators. [86] proposed a Bayesian multivariate emulator, which models a small number of responses at the same time. [82] extends this to a frequentist multivariate emulator estimated using the EM algorithm to handle missing data. [143] proposes a Bayesian lightweight multivariate emulator based on the Bayesian lightweight emulator proposed in [163]. However, the lightweight multivariate emulator in [143] is not as accurate as the GP emulator.

Both univariate and multivariate emulators have correlated errors within the design space. Therefore, it is crucial to choose a correlation function for the errors. In the univariate cases, only the distances in the input space are explored. For example, [169] uses products of one-dimensional correlations, whereas [52] suggests either a non-negative linear or cubic correlation function. Moreoever, [138] uses a diagonal matrix of smoothing parameters based on [169]. Alternatively, [19] proposes the use of the product of exponential correlation functions and this method is also applied in [82]. Different correlation structures for univariate models with qualitative covariates are presented and compared in [156]. In the multivariate cases, there is not only correlation to be modelled between the inputs but between the outputs as well. Hence, two correlation structures have to be estimated. There is a choice between separable and non-separable functions to be made, though separable functions are preferred because they are "easier to implement and interpret" [143] as seen in [49], [19] and [143]. A comparison between separable and non-separable functions is presented in [86].

1.6.2 Examples of application fields

Computer simulations are used in almost every field and therefore, emulators are used in many fields. For instance, [26] applies a univariate frequentist emulator on global gridded crop simulations for maize, rice, soybean and wheat yields and assesses the performance of the emulators using in- and out-of-sample validation. [139] test their proposed Bayesian emulator on both tailored simulated data and the results from an oil-field simulator. A volcanic hazard simulator is emulated in [19] by applying their proposed mixed frequentist-Bayesian emulator.

Similarly, multivariate emulators are applied in various settings. In [86], a multivariate Bayesian model is applied to a climate model, whereas [49] apply their multivariate Bayesian model to simulated output for ecosystem carbon. A Bayesian emulator is used in [99] for mimicking binary, multivariate and correlated within individual data from an ecological simulation for COVID-19, whereas [144] compare their lightweight

multivariate Bayesian emulator to a GP emulator using a humanitarian relief simulator. A comparison of different emulation approaches is given in [28] using an urban dispersion model.

In air pollution modelling, simulators are often used because they allow one to explore the effect of unobserved meteorological conditions on pollutants' behaviour. Hence, [82] is useful in providing conditions for which the NO_2 annual averages will be breached for the monitoring stations in Aberdeen. Smaller studies similar to this one but for small areas or road intersections are presented in [183] and [29]. A Bayesian approach is applied to create an emulator for the ADMS-Urban model using dimensionality reduction is presented in [120]. A Bayesian approach is also applied to the Computational Fluid Dynamics modelling of urban flows in [97], which explores the wind effects in cities. Emulators were used for a short-term study in Beijing to assess the sensitivity of $PM_{2.5}$ concentrations to different types of emissions [13].

1.7 Thesis overview

The main aim of this thesis is to address the issues in the sparsity of air pollution data by exploring the use of miniature automated sensor networks and emulation of physical models. The performance of miniature automated sensors will be appraised to assess whether more sensors can be placed at multiple locations and extend the monitoring network. In order to evaluate what effect unobserved emissions levels and meteorological conditions will have on pollution concentrations, new ADMS-Urban simulations (chosen using Latin Hypercube sampling) will be used to create emulators (a combination of already established techniques and developing new data driven methods) for both the NO_2 annual average and hourly concentrations. This will allow SEPA and other governmental agencies to examine conditions which lead to potential breaches of the regulations.

The remainder of the thesis is organised as follows. Chapter 2 offers an overview of the existing statistical methodologies with specific reference to their application in this thesis. Chapter 3 presents a new data comparative study on the consistency of lower cost miniature automated sensors with each other and with a reference sensor in measuring NO_2 and O_3 concentrations in Edinburgh. Chapter 4 details the exploratory analysis for the AURN recordings for NO_2 in 2012 in Aberdeen and NO_2 in 2015 in Glasgow. Those two years are taken as the baseline conditions, and 98 and 100 ADMS-Urban scenarios (based on a Latin Hypercube design) are produced for Aberdeen and Glasgow respectively. Exploratory analysis for the ADMS-Urban simulations are also included. Chapter 5 presents a frequentist univariate emulator (from the **DiceKriging** package

in **R**) for the NO_2 annual average based on the ADMS-Urban simulation scenarios for each of the monitoring stations in both Aberdeen and Glasgow with hyperspatial correlation structure between the scenarios in the Latin Hypercube space. Chapter 6 provides a novel extension of the work to a Bayesian multivariate emulator which models the NO_2 annual average based on the ADMS-Urban simulation scenarios for all the monitoring stations in Aberdeen and Glasgow by using a hyperspatial correlation between the scenarios in the Latin Hypercube space and free-form scaling matrix for the locations of the monitoring stations. A simulation study comparing the frequentist emulator from Chapter 5 with the Bayesian emulator from Chapter 6 is also included. Chapter 7 models and emulates the number of hourly breaches for the simulated NO_2 hourly concentrations for a year at a single monitoring station using Poisson generalised linear models. Chapter 8 proposes a new hyperspatial-temporal emulator (within the Latin Hypercube design space) for the simulated hourly NO_2 time series for a year at a single monitoring station. Due to a lack of sufficient data, the emulators in Chapters 7 and 8 are only applied to the ADMS-Urban simulations for Central Station in Glasgow. Chapter 9 provides an overall concluding discussion and possible future work.

Chapter 2

Statistical methods for modelling air pollution

Chapter 2 presents the main statistical methods used for analysing temporal, spatial and spatio-temporal data, and emulation of air quality models in this thesis. Section 2.1 presents the general background for time series modelling. Section 2.2 reviews different types of regression modelling. Section 2.3 gives the background on spatial and spatial-temporal data modelling. Lastly, Section 2.4 introduces the general background for the emulation of computer models.

2.1 Time series

A time series is a data set where the observations are ordered in time, such as second by second, minute by minute, hourly, daily, weekly or yearly measurements of the same quantity. In this thesis, a time series is defined as the set $\{Y_1, \ldots, Y_T\}$, where each Y_t $(t \in T)$ is a random variable with T time steps in total. The observations of the time series are defined as the set $\{y_1, \ldots, y_T\}$. The different air pollution data analysed in Chapters 3, 4 and 8 are a time series of observations, and thus here are outlined key quantities and concepts related to time series that will be used in the following chapters. Some main concepts of time series are outlined below from [44], which can also be referred to for further details.

2.1.1 Stationarity

Stationarity is one of the most important features of a time series as it determines future modelling choices. There are two types of stationarity. A time series process

$\{Y_t|t \in T\}$ is **strictly stationary** if the joint distribution $f(Y_{t_1}, \ldots, Y_{t_k})$ is identical to the joint distribution $f(Y_{t_{1+\tau}}, \ldots, Y_{t_{k+\tau}})$ for all sets $\{t_1, \ldots, t_k\}$ and lags (separation values) τ. However, strict stationarity is very restrictive, so more often, weak stationarity is assessed. A time series process is **weakly stationary** if:

- the mean function is constant and finite: $\mu_t = \mathbb{E}[Y_t] = \mu < \infty$;
- the variance function is constant and finite: $\sigma_t^2 = \mathrm{Var}[Y_t] = \sigma^2 < \infty$; and
- the autocovariance and autocorrelation functions only depend on the lag:

$$\gamma_{t,t+\tau} = \mathrm{Cov}[Y_t, Y_{t+\tau}] = \gamma_\tau; \tag{2.1}$$

$$\rho_{t,t+\tau} = \mathrm{Corr}[Y_t, Y_{t+\tau}] = \rho_\tau. \tag{2.2}$$

It is important to determine whether a time series is stationary, because typically one would first remove any non-stationarity from the data, for example via a temporally varying mean model, and then model the remaining variation with a stationary process. This remaining variation is typically correlated in time, and a common class of models for representing this correlation are autoregressive processes. These are defined below, but first details on how to assess whether a time series contains correlation are presented.

2.1.2 Identifying correlation

Correlation

Correlation measures the strength of the linear relationship between two random variables. Consider first random variables X and Y, then their covariance is defined to be:

$$\gamma_{X,Y} = \mathbb{E}[(X - \mu_\mathrm{X})(Y - \mu_\mathrm{Y})] = \mathbb{E}[XY] - \mu_\mathrm{X}\mu_\mathrm{Y}\,, \tag{2.3}$$

where $\mathbb{E}[\cdot]$ is the expectation operator, and $\mu_\mathrm{X} = \mathbb{E}[X]$ and $\mu_\mathrm{Y} = \mathbb{E}[Y]$. However, the covariance is not bounded and so the correlation $\rho_{X,Y}$ is given by:

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_\mathrm{X})(Y - \mu_\mathrm{Y})]}{\sigma_\mathrm{X}\sigma_\mathrm{Y}}\,, \tag{2.4}$$

and ranges between $[-1, 1]$. Here a correlation of 0 corresponds to no linear relationship, while a correlation close to 1 or -1 represents a strong linear relationship. In the above equation σ_X and σ_Y are the standard deviations of X and Y. Then given two data

sets each with n observations (x_i, y_i) for $i = 1, \ldots, n$, Pearson's correlation coefficient is given by:

$$\widehat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\,, \tag{2.5}$$

where $(\bar{x}, \bar{y})$ are the sample means of the two variables. For further details see [47].

Autocorrelation function

Following the definitions of the covariance and correlation between random variables (X, Y), the autocovariance function (ACVF) between random variables (Y_s, Y_t) is defined for all $s, t \in T$ as:

$$\gamma_{s,t} = \mathrm{Cov}[Y_s, Y_t] = \mathbb{E}[Y_s Y_t] - \mathbb{E}[Y_s]\mathbb{E}[Y_t]\,, \tag{2.6}$$

where $\gamma_{t,t} = \mathrm{Cov}[Y_t, Y_t] = \mathrm{Var}[Y_t] = \sigma_t^2$ denotes the variance. From here, the autocorrelation function (ACF) is defined as:

$$\rho_{s,t} = \mathrm{Corr}[Y_s, Y_t] = \frac{\mathrm{Cov}[Y_s, Y_t]}{\sqrt{\mathrm{Var}[Y_s]\mathrm{Var}[Y_t]}} = \frac{\gamma_{s,t}}{\sigma_s \sigma_t}\,, \tag{2.7}$$

with $\rho_{t,t} = \mathrm{Corr}[Y_t, Y_t] = 1$. In order to estimate the ACVF and ACF for a real data set, it is assumed that the dependence structure in the data does not change over time. Hence, it is assumed that for any time points (s, t), temporal shift r and an increment vector τ:

$$\gamma_{s,t} = \mathrm{Cov}[Y_s, Y_t] = \mathrm{Cov}[Y_{s+r}, Y_{t+r}] = \gamma_{s+r,t+r}\,. \tag{2.8}$$

Under this assumption, the only factor affecting the covariance is the lag or distance $\tau = ||s - t||$ between the observations. Therefore, the only set of autocovariances to be estimated are:

$$\gamma_\tau = \mathrm{Corr}[Y_t, Y_{t+\tau}]\,, \qquad \tau = 0, 1, 2, \ldots\,. \tag{2.9}$$

Given an observed time series $(y_1, \ldots, y_n)$, the sample ACVF is:

$$\widehat{\gamma}_\tau = \frac{1}{n}\sum_{t=1}^{n-\tau}(y_t - \bar{y})(y_{t+\tau} - \bar{y})\,, \qquad \tau = 0, 1, \ldots\,, \tag{2.10}$$

where $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$. Hence, the sample ACF is:

$$\widehat{\rho}_\tau = \frac{\sum_{t=1}^{n-\tau}(y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = \frac{\widehat{\gamma}_\tau}{\widehat{\gamma}_0}\,. \tag{2.11}$$

The ACF is a key function to assess the extent to which a time series is correlated in time. This is usually done by producing a correlogram, which is a plot of lag τ against correlation $\widehat{\rho}_\tau$. If a time series is independent then the $\widehat{\rho}_\tau$ values for all $\tau > 0$ will be close to zero, where as correlation will be present if the values of $\widehat{\rho}_\tau$ for small τ are not close to zero. To assess the significance of the correlation, one can produce 95% confidence intervals for the correlation at each lag τ under the assumption of independence, which are estimated as $\pm 1.96/\sqrt{n}$. Hence, any $\widehat{\rho}_\tau$ values that lie inside the confidence interval are not statistically different from zero.

Partial autocorrelation function

The partial autocorrelation function (PACF) is denoted by π_τ and represents the excess correlation in the time series that has not been accounted for by the $\tau - 1$ smaller lags. The sample PACF at lag τ is equal to the estimated lag τ coefficient $\widehat{\alpha}_\tau$, obtained when an Autoregressive process of order τ (AR(τ), see below) model is fitted to the data set. The sample PACF is definite iteratively as follows:

$$\pi_1 = \text{Corr}(y_{t+1}, y_t) = \rho(1), \tag{2.12}$$

$$\pi_\tau = \text{Corr}(y_{t+\tau} - \widehat{y}_{t+\tau}, y_t - \widehat{y}_t) \quad \text{for} \quad \tau \leq 2. \tag{2.13}$$

The PACF could be plotted against the lags to determine an appropriate AR(p) process for the given data set. As with the ACF, a 95% confidence interval of $\pm 1.96/\sqrt{n}$ is defined for no correlation, and for an AR(p), then the order p is chosen to be equal to the lag τ at which the last significantly different from zero value is present.

It has to be noted that for the ACF and PACF plots, the more lags that are plotted, some of the lags will appear significant by chance. As the significance level is set to 95%, it is expected on a plot with 20 lags, one of the lags to be significant by chance. Multiple testing (such as Bonferroni [66] or Tukey's honest significant difference [194]) could be applied but as the ACF and PACF plots in this thesis are used only for initial impressions, this has not be done.

2.1.3 Autoregressive and Moving average processes

An Autoregressive process of order p, AR(p) is defined as:

$$Y_t = \zeta_1 Y_{t-1} + \cdots + \zeta_p Y_{t-p} + Z_t\,, \tag{2.14}$$

where Y_t ($t \in T$) is regressed on its past values $Y_{t-1}, ..., Y_{t-p}$, the parameters $\zeta_1, ..., \zeta_p$ are the autoregressive coefficients and Z_t is a normally distributed purely random process with mean 0 and variance σ_z^2. In order to choose the order, the partial autocorrelations on the PACF plot lose significance after p lags, whereas on the ACF plot the autocorrelation decreases exponentially over time.

The Moving Average process of order q, MA(q) is defined as:

$$Y_t = \lambda_0 Z_t + \lambda_1 Z_{t-1} + \cdots + \lambda_q Z_{t-q}\,, \tag{2.15}$$

where each Z_{t-i} is an independent purely random process with mean 0 and variance σ_z^2 and $\lambda_0, ..., \lambda_q$ are the lag coefficients with $\lambda_0 = 1$. In order to choose the order q, the autocorrelations on the ACF plot lose significance after q lags and the partial autocorrelations on the PACF plot decrease exponentially over time.

The two processes can be combined into an Autoregressive Moving Average process of order (p, q) (ARMA(p, q)) defined as:

$$\begin{aligned} Y_t &= \zeta_1 Y_{t-1} + \cdots + \zeta_p Y_{t-p} + Z_t + \lambda_1 Z_{t-1} + \cdots + \lambda_q Z_{t-q} \\ &= \sum_{y=1}^{p} \zeta_j Y_{t-j} + \sum_{y=1}^{q} \lambda_j Z_{t-j} + Z_t\,, \end{aligned} \tag{2.16}$$

which combines the AR and MA processes.

2.1.4 Adjusted confidence intervals

The time series data examined in this thesis are taken at regular time intervals. In this subsection, an explanation of how to adjust for autocorrelation in confidence intervals for the means and medians of time series, respectively, is presented.

Confidence intervals for the mean

Given data $\mathbf{y} = (y_1, ..., y_n)$, a confidence interval is useful as it provides an interval of plausible values for a specific parameter. For instance, for the population mean μ, the standard confidence interval has the form:

$$\widehat{\mu} \pm t\left(1 - \frac{\alpha}{2}, \mathrm{n} - 1\right) \times \frac{\widehat{\sigma}}{\sqrt{n}}\,, \tag{2.17}$$

where, if the data are normally distributed:

- $\widehat{\mu}$ is the estimated sample mean, calculated as $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$;
- α is the significance level and in this thesis, confidence intervals with $\alpha = 0.05$ will be produced;
- n is the sample size;
- $t\left(1-\frac{\alpha}{2}, n-1\right)$ is the critical value for the Student t-distribution with $n-1$ degrees of freedom at $1-\frac{\alpha}{2}$ significance level; and
- $\widehat{\sigma}$ is the estimated standard deviation of the sample estimated as $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2}$.

However, the standard confidence interval assumes that the observations are independent of each other, but as previously discussed, this is not the case when time series data are used. Therefore, the confidence intervals have to be adjusted to allow for autocorrelation (refer to Subsection 2.1.2). One option is to use bootstrapping. This technique does not make any distributional assumptions. To produce a 95% bootstrap confidence interval, the following steps are taken:

1. Resample (with replacement) the data set 1000 times.
2. The mean is computed for each new sample.
3. The resampled means form the empirical distribution of the mean. Therefore, the 2.5^{th} and 97.5^{th} quantiles from this distribution form the 95% bootstrap confidence interval.

Another alternative is to adjust the standard error $\widehat{\sigma}$ to account for the autocorrelation using the following variance result [44]:

$$\text{Var}(\widehat{\mu}) = \frac{\widehat{\sigma}^2}{n}\left[1 + 2\sum_{\tau=1}^{n-1}\left(1-\frac{\tau}{n}\right)\widehat{\rho}(\tau)\right], \tag{2.18}$$

where:

- $\widehat{\sigma}^2$ is the estimated variance of the time series defined earlier;
- n is the sample size;
- τ is the separation value (lag); and
- $\widehat{\rho}(\tau)$ is the estimated autocorrelation (see Subsection 2.1.2) at lag τ.

Then the confidence interval is estimated as:

$$\widehat{\mu} \pm t\left(1-\frac{\alpha}{2}, n-1\right) \times \sqrt{\mathrm{Var}(\widehat{\mu})}\,. \tag{2.19}$$

In the specific case of modelling data with an Autoregressive process of order 1 (AR(1), for more information see Subsection 2.1.3) with AR(1) parameter ζ, the formula for the estimated variance above simplifies to $\mathrm{Var}(\widehat{\mu}) = \frac{\sigma^2}{n}\frac{1+\zeta}{1-\zeta}$ [44].

Confidence intervals for the median

The median is useful in the cases of skewed data because the mean is adversely affected by outliers. For example, in the case of right-skewed data (which is often the case with air pollution data), the mean is no longer at the centre of the data and is less representative for the average value of the distribution. In that sense, the median offers a more robust estimate than the mean. One way is to provide a crude 95% interval for the median using the 2.5^{th} and the 97.5^{th} quantiles of the sample as suggested by [87]. A confidence interval for the median can be computed by bootstrapping as follows:

1. Resample (with replacement) the data set 1000 times.
2. The median is computed for each new sample.
3. The resampled medians form the empirical distribution of the median. Therefore, the 2.5^{th} and 97.5^{th} quantiles from this distribution form the 95% bootstrap confidence interval.

Another alternative is to calculate the $100(1-\alpha)\%$ confidence interval, presented in [30]. The confidence interval is calculated by first calculating the pair of numbers (r, s):

$$r = \frac{n}{2} - \left(\mathrm{N}_{1-\frac{\alpha}{2}} \times \frac{\sqrt{n}}{2}\right), \tag{2.20}$$

$$s = 1 + \frac{n}{2} + \left(\mathrm{N}_{1-\frac{\alpha}{2}} \times \frac{\sqrt{n}}{2}\right), \tag{2.21}$$

where r and s are rounded to the nearest integers and N is the normal cumulative distribution probability for a confidence level $1-\frac{\alpha}{2}$. Then order the n observations from smallest to largest, and the r^{th} and s^{th} values in the ordered sample form the $100(1-\alpha)\%$ confidence interval for the median.

2.2 Regression modelling

This section presents different types of regression modelling which is applied in a variety of scenarios in Chapters 3 through 8. In this thesis, regression modelling is used for both exploratory analysis modelling and de-trending time series data. This section presents linear regression and its extension, generalised linear models, including estimating the model's parameters, assessing the fit and model comparison.

2.2.1 Linear regression

Linear regression is used for exploratory analysis and modelling in Chapters 3 through 8 as well as for de-trending time series data. Linear regression is described as in [79].

Ordinary Least Squares

Linear regression is a model fitting technique which aims to minimise the sum of squared errors as presented in [79]. Let $\mathbf{y} = (\mathrm{y}_1, ..., \mathrm{y}_n)$ be the response vector with n entries and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be the $n \times p$ matrix of covariates with $\mathbf{x}_i = (\mathrm{x}_{i1}, ..., \mathrm{x}_{ip})$, for $i = 1, \ldots, n$. Then the regression model has the general form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \tag{2.22}$$

where:

- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is the parameter vector with length p; and
- $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbb{I})$ is the error vector $(n \times 1)$ with $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$ where $\mathbb{I}$ denotes the identity matrix so that ϵ_i and ϵ_j are independent for all $i \neq j$.

The coefficients are calculated by minimising the sum of squares using ordinary least squares (OLS), that is minimising:

$$\mathbf{C}_{\mathrm{OLS}}^{-1} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\,, \tag{2.23}$$

which leads to:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\,. \tag{2.24}$$

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is estimated as:

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}. \tag{2.25}$$

Four assumptions are considered when fitting a linear regression:

1. The observations are independent of each other.
2. The errors are normally distributed.
3. All non-random structure of the data is captured by the deterministic part of the model.
4. The errors have constant variance, σ^2.

To check these assumptions, two plots will be used. A residuals vs. fitted values plot will be used to check whether the errors have constant variance, as well as checking for patterns (due to the model not capturing all the non-random structures of the data) and outliers. Additionally, a normal quantile-quantile (qq) plot of the residuals will be used to assess the normality by checking that the points lie on an approximately straight line.

A further check for the normality of the residuals is performed using a Shapiro-Wilk test [177]. For the set of the residuals $\epsilon_1, \ldots, \epsilon_n$ from a linear model, the test statistic W is:

$$\text{W} = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.26}$$

where a_i is a constant generated from the means, covariances and variances (of the residuals set) from a normally distributed sample. A hypothesis test is performed where the null hypothesis is that the residuals are normally distributed. If the p-value is smaller than a significance level α (in this thesis $\alpha = 0.05$), the null hypothesis is rejected and there is evidence that the residuals are not normally distributed.

Generalised Least Squares

When the assumption that the errors are independent is broken (which is the case in time series modelling where the observations are correlated), a generalised least squares (GLS) fit can be applied. The description of GLS is taken from [79]. In the OLS case, it is assumed that $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$, whereas in the GLS case, it is assumed that $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{\Sigma}$ where σ^2 is unknown but $\boldsymbol{\Sigma}$ is the known correlation matrix. Applying a Choleski decomposition for $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}^\top$, where $\mathbf{T}$ is a $n \times n$ triangular matrix, the linear model can be rewritten as:

$$\mathbf{T}^{-1}\mathbf{y} = \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{T}^{-1}\boldsymbol{\epsilon}\,. \tag{2.27}$$

Therefore, the sum of squares objective function is now:

$$\begin{aligned}\mathbf{C}_{\text{GLS}}^{-1} &= (\mathbf{T}^{-1}\mathbf{y} - \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{T}^{-1}\mathbf{y} - \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{T}^{-\top}\mathbf{T}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\,,\end{aligned} \tag{2.28}$$

and is minimised by:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{y}\,. \tag{2.29}$$

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is estimated as:

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\,, \tag{2.30}$$

which results in larger variances in comparison to the OLS fit.

For time series data, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots of the residuals from the OLS fit can be used to assess $\boldsymbol{\Sigma}$. For more information on the ACF and PACF plots, refer to Subsection 2.1.2. The plots are examined to determine whether an autoregressive (AR) or moving average (MA) error structure (see Subsection 2.1.3) is more appropriate and of what order, or a combination of the two. It has to be noted that the ACF and PACF plots of the GLS models do not reflect the change in the correlation structure of the residuals as only the variances of $\widehat{\boldsymbol{\beta}}$ are changed.

2.2.2 Smooth functions

Locally weighted regression smoother

When plotting a time series, there are often considerable fluctuations, which make the general trend in the time series hard to establish. Therefore, a smoothing technique can be applied. One such common exploratory approach is locally weighted regression smoother (lowess), first proposed in [46]. For $\mathbf{y} = [y_1, \ldots, y_n]^{\top}$,

$$y_i = f(x_i) + \epsilon_i\,, \tag{2.31}$$

where:

- $f(x_i)$ is a regression function based on the data $(x_i, y_i); i = 1, \ldots, n$; and
- ϵ_i is the error term.

For each observation x_i, a window (neighbourhood) surrounding the observation is obtained by identifying the k nearest neighbouring observations to this point. This area is written as $N(x_i)$. The distance between x_i and the furthest away point within the neighbourhood is defined as:

$$\Delta(x_i) = \max_{N(x_i)} |x_j - x_i| \,, \tag{2.32}$$

where $j = 1, \ldots, k$. Within a neighbourhood, weights are then assigned to each observation using a tri-cube weight function:

$$w(x_j - x_i, \Delta(x_i)) = W\left(\frac{|x_j - x_i|}{\Delta(x_i)}\right) , \tag{2.33}$$

where:

$$W(u) = \begin{cases} (1 - u^3)^3 & \text{for } 0 \leq u < 1 \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{2.34}$$

The weights are used to produce the locally weighted regression smoother. The smoothing parameter (called **span**) determines the quantity of the data which contributes to the estimate at each point. The span specifies the number of k nearest neighbours to the target point x_i and is usually set to $\frac{2}{3}$ of the data as recommended by [206]. This means that $\frac{2}{3}$ of the data are used for the fit at each target point.

B-spline

In time series modelling, some covariates exhibit complex non-linear relationships with the response and therefore, a b-spline approach can be implemented to estimate this non-linear relationship. The work here is based on the material in [56]. B-splines are a special case of basis functions. Consider the simple non-linear relationship $f(\cdot)$:

$$y_i = f(x_i) + \epsilon_i \,, \tag{2.35}$$

where x_i is the i^{th} observed covariate $\mathbf{x} = [x_1, \ldots, x_n]^\top$. A curve estimate is produced by fitting the regression:

$$y_i = \beta_0 B_0(x_i) + \beta_1 B_1(x_i) + \beta_1 B_1(x_i) + \cdots + \beta_p B_p(x_i) + \epsilon_i \,, \tag{2.36}$$

where the B_js ($j = 0, 1, \ldots, p$) are called **basis functions** and the β_js are called **basis coefficients**. Hence,

$$f(\mathbf{x}) = \sum_{j=0}^{p} \beta_j B_j(x) \,. \tag{2.37}$$

In the b-spline case, each basis function is only non-zero in the interval between a small number of adjacent knots (knots are the points at which joins of the spline occur), thus resulting in a sparse design matrix, i.e. making the b-splines more computationally efficient. Let $t = (t_0, t_1, \ldots, t_{p+d+1})$ be the knot vector for $t_0 \leq t_1 \leq \cdots \leq t_{p+d+1}$, where d is the degree of the polynomial, then b-splines are defined recursively as:

$$B_{j,d}(x) = \frac{x - t_j}{t_{j+d} - t_j} B_{j,d-1}(x) + \frac{t_{j+d+1} - x}{t_{j+d+1} - t_{j+1}} B_{j+1,d-1}(x) \,, \tag{2.38}$$

and

$$B_{j,0}(x) = \begin{cases} 1 & t_j \leq x < t_{j+1} \,, \\ 0 & \text{otherwise.} \end{cases} \tag{2.39}$$

2.2.3 Likelihood estimation

In every type of statistical modelling used in this thesis, there is a vector of n random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ which follows a probability density function $p(\mathbf{Y}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector containing parameters of interest. The **likelihood function** allows the estimation of plausible values for $\boldsymbol{\theta}$ based on the observed data $\mathbf{y} = (y_1, \ldots, y_n)$. Assuming that all observations are independent, the likelihood function is:

$$\mathrm{L}(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i; \boldsymbol{\theta}) = p(y_1; \boldsymbol{\theta}) p(y_2; \boldsymbol{\theta}) \cdots p(y_n; \boldsymbol{\theta}) \,. \tag{2.40}$$

Generally, the likelihood function is the probability or probability density of obtaining the observed data with particular values for $\boldsymbol{\theta}$, hence, $\mathrm{L}(\mathbf{y}; \boldsymbol{\theta})$ should be regarded as a function of $\boldsymbol{\theta}$.

The most popular strategy for using the likelihood function to provide appropriate estimates is using a **maximum likelihood estimator** (MLE). The MLE is defined as the values of $\boldsymbol{\theta}$, which maximise the likelihood function. These values are denoted as $\widehat{\boldsymbol{\theta}}$. The easiest way of estimating $\widehat{\boldsymbol{\theta}}$, in simple situations, is to use the log-likelihood function $l(\boldsymbol{\theta}) = \log(\mathrm{L}(\mathbf{y}; \boldsymbol{\theta}))$. Taking the first derivative and setting it equal to zero will identify $\widehat{\boldsymbol{\theta}}$ in a simple situation. To make sure that this value is the maximum, the second derivative must be negative. In a multivariate case, a Hessian matrix (matrix of the second derivatives with respect to all the parameters) must be positive definite. However, there are cases where there is no explicit solution to the MLE. Therefore, a numerical method should be used, for instance, the Newton-Raphson method. Nevertheless, the Newton-Raphson method is complicated to use when the data set is incomplete (has missing values) [126]. In Subsection 2.2.4, a quasi-Newtonian approach (the BFGS algorithm) is presented for numerical optimisation.

2.2.4 Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm

In Chapters 5, 6 and 8, the parameters are estimated using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [137]. BFGS is a quasi-Newton algorithm, where x_0 is a starting point with convergence tolerance $\epsilon > 0$, inverse Hessian approximation H_0 (a square matrix of the second-order derivates of a function is called the Hessian, for more details see [95]), α_0 is a step length from a line search in order to ensure sufficient decrease in the likelihood, likelihood value f_0 and gradient ∇f_0.

For step k:

1. Set the search direction $p_k = -H_k \nabla f_k$.
2. Define $x_{k+1} = x_k + \alpha_k p_k$.
3. Compute $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$.
4. Calculate $H_{k+1} = \left(\mathbb{I}_n - \frac{1}{y_k^\top s_k} s_k y_k^\top\right) H_k \left(\mathbb{I}_n - \frac{1}{y_k^\top s_k} y_k s_k^\top\right) + \frac{1}{y_k^\top s_k} s_k s_k^\top$.
5. Repeat the steps until $||\nabla f_k|| < \epsilon$.

The L-BFGS-B is an extension of the BFGS algorithm, which is applied with a limited-memory version for large number of variables and there are constraints on the parameters used for estimating the likelihood [137].

2.2.5 Model comparison

When multiple linear models are fitted using the same response but different combinations of explanatory variables, the models in this thesis are compared using the coefficient of determination $R^2_{adj.}$, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as recommended by [79]. The coefficient of determination $R^2_{adj.}$ is defined as:

$$R^2_{adj.} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,, \tag{2.41}$$

where $\widehat{y}_i$ is the fitted value for the i^{th} response y_i and $\bar{y}$ is the mean value for the responses. The model with the highest $R^2_{adj.}$ is the preferred one. Linear models under the OLS fit can be compared using the $R^2_{adj.}$. However, information criteria can be used to compare different types of correlated and uncorrelated error models. The two information criterion penalise the likelihood in different ways. They are defined as follows:

$$\text{AIC} = 2\text{p} - 2\ln(\text{L}(\mathbf{y};\widehat{\boldsymbol{\theta}}))\,, \quad \text{and} \quad \text{BIC} = \ln(n)\text{p} - 2\ln(\text{L}(\mathbf{y};\widehat{\boldsymbol{\theta}}))\,, \tag{2.42}$$

where p is the number of parameters in the model, $\text{L}(\mathbf{y};\widehat{\boldsymbol{\theta}})$ is the maximised likelihood and n is the number of observations. The model with the smallest AIC/BIC is considered to be the best one. Overall, BIC tends to choose models with less covariates than AIC because the penalty term for BIC is larger. Therefore, it is beneficial whether the two criteria would agree on the same model but in case when different models are chosen as best, BIC would be the preferred criterion as it selects the model with less covariates. Degrees of freedom will also be provided to ease the comparison of models.

2.2.6 K-fold cross validation

Alternatively, different models can be compared based on their prediction power using the Root Mean Squared Prediction Error (RMSPE) based on a k-fold cross validation as described in [105]. A k-fold cross validation requires that the data are split in to k groups (folds). The tested model is fitted on k-1 folds, while the remaining fold is used as a validation set and the model predicts its values. The RMSPE is calculated as:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}\,, \tag{2.43}$$

where in fold k:

- y_i is the true value of the i^{th} observation ($i = 1, \ldots, n$); and
- $\widehat{y}_i$ is the predicted value for y_i.

The RMSPE estimates the prediction error, on average, of an observation. Hence, when comparing models, the model with the smallest RMSPE has better prediction power than the other models. Furthermore, [87] states that a 95% bootstrap confidence interval for RMSPE can be produced as it makes no distributional assumptions. This is done in the following way:

1. Resample (with replacement) the set of paired actual and predicted points 1000 times.
2. The RMSPE is computed for each new sample.
3. The resampled RMSPE form the empirical distribution of the RMSPE. Therefore, the 2.5^{th} and 97.5^{th} quantiles from this distribution form the 95% bootstrap confidence interval.

2.2.7 Generalised linear models

This subsection introduces generalised linear models (GLMs) as an extension of regression modelling. In Chapter 7, the number of occurrences of NO_2 hourly concentrations above 200 μg m^{-3} are modelled using a GLM approach. This subsection is based on the material in [65], [78] and [122].

A generalised linear model is an extension of the linear regression model (Subsection 2.2.1) for a response vector $\mathbf{y} = [y_1, \ldots, y_n]^\top$. For $i = 1, \ldots, n$, it is defined as:

$$y_i \sim f(y_i; \theta_i, \phi)\,, \tag{2.44}$$

with $\mathbb{E}(\mathbf{y}) = \boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^\top$ and ϕ is a dispersion parameter, if required. Then a link function $g(\cdot)$ is defined as:

$$g(\theta_i) = \mathbf{x}_i\boldsymbol{\beta} = \eta_i\,, \tag{2.45}$$

where:

- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is the design matrix ($n \times p$) with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ being its i^{th} row; and

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the vector of parameters to be estimated.

The difference from linear regression is that $\mathbf{y}$ does not have to follow a normal distribution but any distribution $f(\cdot)$ that belongs to the exponential family. Let the probability density function (pdf) of a random variable y depend on the canonical parameter θ and the dispersion parameter ϕ. Then, the pdf belongs to the exponential family if it can be written as:

$$f(y; \theta, \phi) = \exp\left[a(y)b(\theta) + c(\theta) + d(y, \phi)\right] . \tag{2.46}$$

The log-likelihood for exponential family distributions is then:

$$l(y; \theta, \phi) = a(y)b(\theta) + c(\theta) + d(y, \phi) , \tag{2.47}$$

where $a(y)$ is called the canonical form. The mean of $a(y)$ is estimated by solving $\int_{y \in R_y} \frac{df(y;\theta,\phi)}{d\theta} dy = 0$, from where:

$$\mathbb{E}(a(y)) = -\frac{c'(\theta)}{b'(\theta)} . \tag{2.48}$$

Similarly, the variance of $a(y)$ is estimated as $\int_{y \in R_y} \frac{d^2 f(y;\theta,\phi)}{d\theta^2} dy = 0$ from where:

$$\mathrm{Var}(a(y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} . \tag{2.49}$$

The score function U is defined as the first derivate of the log-likelihood with respect to θ:

$$U(y; \theta) = \frac{dl(y; \theta)}{d\theta} = a(y)b'(\theta) + c'(\theta) . \tag{2.50}$$

From $\mathbb{E}(a(y))$, $\mathbb{E}(U) = 0$ and from $\mathrm{Var}(a(y))$, $\mathrm{Var}(U) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta)$, which is also known as the information $\mathcal{I}$ $(p \times p)$.

Using the score statistics results, the parameters $\boldsymbol{\beta}$ (in model 2.44 - 2.45) can be estimated using an iteratively reweighted least squares (IRWLS) algorithm. Let $\mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X}$, where $\mathbf{W}$ is $n \times n$ diagonal matrix with elements $w_{ii} = \frac{1}{\mathrm{Var}(y_i)(g'(\mu_i))^2}$ and $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{X}^\top \mathbf{W} \mathbf{z}$ where $z_i = (y_i - \mu_i) g'(\mu_i)$. Then, the IRWLS algorithm is as follows:

1. Set an initial value for $\widehat{\boldsymbol{\beta}}^{(0)}$.
2. For iteration $m = 1, 2, \ldots$: $\boldsymbol{\eta}^{(m-1)} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(m-1)}$ and $z_i^{(m-1)} = (y_i - \mu_i^{(m-1)})g'(\mu_i^{(m-1)})$.
3. Calculate $\widehat{\boldsymbol{\beta}}^{(m)} = [\mathbf{X}^\top \mathbf{W}^{(m-1)}\mathbf{X}]^{-1}\mathbf{X}^\top \mathbf{W}^{(m-1)}(\boldsymbol{\eta}^{(m-1)} + \mathbf{z}^{(m-1)})$.
4. Repeat steps 2 and 3 until the difference of $\widehat{\boldsymbol{\beta}}^{(m)} - \widehat{\boldsymbol{\beta}}^{(m-1)}$ is below a threshold and hence, convergence has been reached.

It has to be noted that the dispersion parameter ϕ has no effect on the estimates of $\widehat{\boldsymbol{\beta}}$. However, if ϕ is required, it affects the estimate of the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$:

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\widehat{\phi}\,. \tag{2.51}$$

Model testing for a GLM can be based on the deviance statistic, where H_0 is the simpler model (with less covariates) and the alternative, H_1, is the more complex model (with more covariates). The deviance is defined by:

$$D = 2\left[\text{l}\left(\widehat{\boldsymbol{\beta}}_{\text{max}}; \mathbf{y}\right) - \text{l}\left(\widehat{\boldsymbol{\beta}}; \mathbf{y}\right)\right], \tag{2.52}$$

where $\text{l}\left(\widehat{\boldsymbol{\beta}}_{\text{max}}; \mathbf{y}\right)$ is the maximised log-likelihood for the full model with all covariates and $\text{l}\left(\widehat{\boldsymbol{\beta}}\right.$ is the maximised log-likelihood for the model of interest. The deviance between a few models can be compared to a $\chi^2(p-q)$, where p is the number of parameters under H_1 and q is the number of parameters under H_0 (with $q < p < n$). Alternatively, GLMs can also be compared based on AIC or BIC as described in Subsection 2.2.1.

For the diagnostic plots of GLMs, deviance residuals must be calculated depending on the distribution of the response vector $\mathbf{y}$. The i^{th} deviance residual is defined as:

$$d_i = \text{sign}(y_i - \widehat{y}_i)\sqrt{2\left(y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) - (y_i - \widehat{y}_i)\right)}, \tag{2.53}$$

where $\text{sign}(\cdot)$ is an indicator variable for the sign of the expression. Standard plots for all GLMs are the qq-plot of the deviance residuals to check whether they are normally distributed. The deviance residuals are plotted against the fitted values as well as the covariates to check for any patterns indicating unexplained variability in the data. Lastly, the observed vs. fitted values are plotted to check whether the fitted values from the models are similar to the observed values.

2.3 Spatial and spatial-temporal modelling

In Chapters 4 through 8, the time series data are collected at multiple locations. Therefore, it is of interest to model spatio-temporal data. This section first introduces the key features of spatial modelling, which is then extended to spatio-temporal modelling.

2.3.1 Geostatistics

Geostatistics is a branch of spatial statistics which is used to analyse data collected at a set of n point locations. The air pollution data presented in Chapters 4 through 8 are gathered at n different locations across Aberdeen and Glasgow, or within the Latin Hypercube design space. Therefore, in this subsection key quantities and concepts related to geostatistics which will be used in the aforementioned chapters are presented based on the description in [63], which can be referred to for further details on geostatistics.

In general, univariate geostatistical data are defined as $(\mathbf{s}_i, y_i)$ for $i = 1, \ldots, n$ with $\mathbf{s}_i$ specifying the spatial location (usually in two dimensions) and y_i being the the measurement for some response variable. A geostatistical process, on the other hand, is defined as:

$$\mathbf{Y} = \{Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\}, \tag{2.54}$$

where Y is the random variable at location $\mathbf{s}_i$. Naturally, geostatistical data will have positive correlation, because the nearer two observations are in space typically, the more similar are the observations at these locations. The covariance function between two locations $\mathbf{s}$ and $\mathbf{p}$ is:

$$\mathrm{C}_{Y(\mathbf{s},\mathbf{p})} = \mathrm{Cov}[Y(\mathbf{s}), Y(\mathbf{p})] = \mathbb{E}\left[(Y(\mathbf{s}) - \mu_Y(\mathbf{s}))(Y(\mathbf{p}) - \mu_Y(\mathbf{p}))\right], \tag{2.55}$$

where $\mu_Y(\mathbf{s}) = \mathbb{E}\left[Y(\mathbf{s})\right]$ is the mean function. For a covariance function to be valid, it has to be non-negative definite.

Stationarity and isotropy

By definition, a geostatistical process $\mathbf{Y}$ is stationary if it has the same characteristics at any location in space, for instance, constant mean, constant variance, etc. As with time series, there are two types of stationarity - strict and weak. A geostatistical process $\mathbf{Y}(\mathbf{s})$ is defined as strictly stationary if:

$$f(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)) =_d f(Y(\mathbf{s}_1 + \boldsymbol{\tau}), \ldots, Y(\mathbf{s}_n + \boldsymbol{\tau})), \tag{2.56}$$

for any displacement vector $\boldsymbol{\tau}$ and any set of locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, and $=_d$ meaning equal in distribution. If $\mathbf{Y}$ is strictly stationary, then $\mathbf{Y}$ has the same distribution for all locations $\mathbf{s}$, from which it follows that the mean is constant $\mathbb{E}(Y(\mathbf{s})) = \mu_Y(\mathbf{s}) = \mu_Y$ as well as the variance $\mathrm{Var}(Y(\mathbf{s})) = \sigma^2_Y(\mathbf{s}) = \sigma^2_Y$. The bivariate distribution does not depend on spatial location, that is $f(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) =_d f(Y(\mathbf{0}), Y(\mathbf{h}))$ from which it follows that the covariance function between two points depends only on the distance and direction between them, but not the two locations $\mathrm{Cov}[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})] = C_Y(\mathbf{s}, \mathbf{s}+\mathbf{h}) = C_Y(\mathbf{h})$.

On the other hand, a geostatistical process $\mathbf{Y}$ is defined as weakly stationary if the mean is constant without depending on locations, $\mathbb{E}[Y(\mathbf{s})] = \mu_Y(\mathbf{s}) = \mu_Y$, and the covariance function only depends on the lag, $\mathrm{Cov}[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})] = C_Y(\mathbf{s}, \mathbf{s}+\mathbf{h}) = C_Y(\mathbf{h})$, with both of these being finite. This in turn means that $\mathrm{Var}(Y(\mathbf{s})) = \sigma^2(\mathbf{s}) = \sigma^2$.

A further simplification of stationary geostatistical processes is isotropy. If a geostatistical process is isotropic, the covariance function is directionally invariant and therefore, not the direction but only the size of the distance (lag) between two points determines the covariance. The covariance is simplified to $C_Y(\mathbf{h}) = C_Y(||\mathbf{h}||) = C_Y(\mathrm{h})$, where $||\mathbf{h}|| = \mathrm{h}$ is the Euclidean distance (for details on Euclidean distance refer to [14]) of the lag $\mathbf{h}$.

Variogram

In geostatistics, variograms are commonly used instead of covariance functions to represent correlation. A semi-variogram of a geostatistical process $\mathbf{Y}(\mathbf{s})$ is a function $\gamma_Y(\mathbf{s}, \mathbf{p})$, which measures the variance of the difference in the process at two spatial locations $\mathbf{s}$ and $\mathbf{p}$:

$$\gamma_Y(\mathbf{s}, \mathbf{p}) = \frac{1}{2}\mathrm{Var}[Y(\mathbf{s}) - Y(\mathbf{p})]\,. \tag{2.57}$$

Since $\gamma_Y(\mathbf{s}, \mathbf{p})$ is called the semi-variogram, $2\gamma_Y(\mathbf{s}, \mathbf{p})$ is called the variogram. A variogram is related to intrinsic stationarity. A geostatistical process $\mathbf{Y}$ is intrinsically stationary when the difference of the geostatistical processes is weakly stationary. Hence, the mean $\mathbb{E}[Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})] = 0$ for all locations $\mathbf{s}$ and lag vectors $\mathbf{h}$, and the semi-variogram $\gamma_Y(\mathbf{s}, \mathbf{s}+\mathbf{h}) = \frac{1}{2}\mathrm{Var}[Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})] = \gamma_Y(\mathbf{h})$ depends only on the displacement vector $\mathbf{h}$ for all locations $\mathbf{s}$. If the intrinsically stationary process is also isotropic, there is a further simplification to the semi-variogram:

$$\gamma_Y(\mathbf{s}, \mathbf{s}+\mathbf{h}) = \frac{1}{2}\mathrm{Var}[Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})] = \gamma_Y(\mathrm{h})\,, \tag{2.58}$$

where $\mathrm{h} = ||\mathbf{h}||$ and, hence, only the distance between the two points is important. The semi-variogram for a weakly stationary and isotropic process can be plotted (see below for the estimator) to check for the presence of spatial correlation in the data as seen in Figure 2.1, which shows the expected shape under positive spatial correlation.

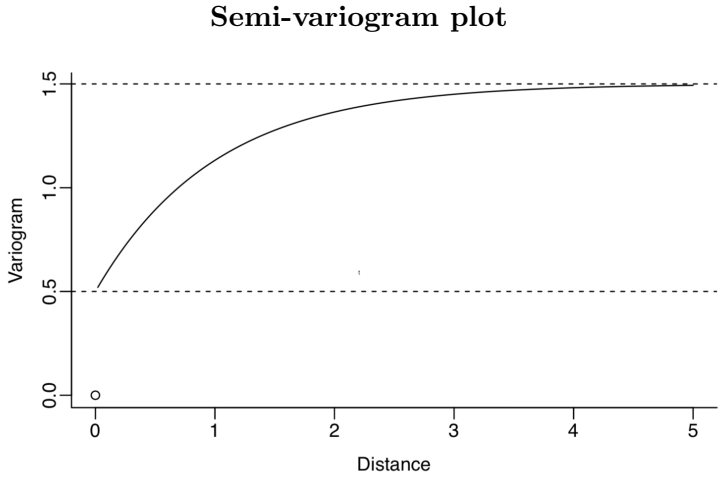

FIGURE 2.1: A general semi-variogram plot [113].

The semi-variogram is discontinuous at zero, where the horizontal axis reflects the increase in the Euclidian distance between two points. The **sill** (total variation) of the semi-variogram is the limiting value of the semi-variogram as the Euclidian distance increases, and measures the total variation in the data. The **range** is the minimal Euclidian distance at which the observations are uncorrelated (independent). However, as the points are in hyperspace, the Euclidian distance between is not measured in a specific unit. It also has to be noted that the range could go to infinity. The **nugget** is the limiting value of the semi-variogram as the distance tends to zero. In essence the nugget is the measurement error, hence the non-spatial variation in the data. Lastly, the **partial sill** is the amount of spatial variation, which is the difference between the sill and the nugget.

In order to check for the presence of spatial autocorrelation in a data set $\mathbf{Y} = \{y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)\}$, one has to estimate the empirical semi-variogram. Let $N(\mathrm{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| = \mathrm{h}\}$ be the set of pairs of spatial locations at a distance h, and let $|N\mathrm{h}|$ denote the number of points in this set. Then, the empirical semi-variogram at a distance h, $\widehat{\gamma}_Y(\mathrm{h})$ is:

$$\widehat{\gamma}_Y(\mathrm{h}) = \frac{1}{2|N\mathrm{h}|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathrm{h})} [y(\mathbf{s}_i) - y(\mathbf{s}_j)]^2 . \tag{2.59}$$

In some cases, however, there may not be enough points to average in order to calculate a "good" estimate of the true semi-variogram. Instead, a binned estimator can be used.

Let, the space of distances be partitioned into K bins, $I_K = (\mathrm{h}_{k-1}, \mathrm{h}_K]$ for $k = 1, \ldots, K$ where $0 = \mathrm{h}_0 < \mathrm{h}_1 < \ldots < \mathrm{h}_K$. The midpoint of each interval is defined to be $\mathrm{h}_k^m = (\mathrm{h}_{k-1} + \mathrm{h}_k)/2$. Then the pairs of the distances in each interval are calculated as $N(\mathrm{h}_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : ||\mathbf{s}_i - \mathbf{s}_j|| \in I_K\}$. Therefore, the **binned empirical semi-variogram** is defined as:

$$\widehat{\gamma}_Y(\mathrm{h}_k^m) = \frac{1}{2|N(\mathrm{h}_k)|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathrm{h}_k)} [y(\mathbf{s}_i) - y(\mathbf{s}_j)]^2 \,. \tag{2.60}$$

To ensure that there is statistically significant spatial correlation present, a Monte Carlo envelop is added to the binned empirical semi-variogram by adding the lower and upper limits for the set of semi-variograms likely under independence. The Monte Carlo envelop is calculated by repeating two steps for a large number, $j = 1, \ldots, J$, of iterations. Firstly, a pseudo data set $\mathbf{y}^{(j)} = \{y^{(j)}(\mathbf{s}_1), \ldots, y^{(j)}(\mathbf{s}_n)\}$ is created by randomly permuting the n data points to the spatial locations $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Then the semi-variogram for each distance is computed $(\widehat{\gamma}_Y^{(j)}(\mathrm{h}_1^n), \ldots, \widehat{\gamma}_Y^{(j)}(\mathrm{h}_K^n))$. For each distance, h_i^n, a 95% envelope is computed from the 2.5^{th} and the 97.5^{th} percentiles of the set $\{\widehat{\gamma}_Y^{(j)}(h_i^n)\}_{j=1}^J$. The envelope presents the range of plausible semi-variograms that could be produced if there was no spatial correlation in the data. Hence, if the semi-variogram from the real data lies outside the envelope at some point, there is evidence of spatial correlation. Finally, it is important to note that the assessment of the spatial dependence assumes isotropy.

Covariance models

For the semi-variogram and the covariance functions, there are many parametric models in geostatistics. Here, a few models are presented with the assumption that the geostatistical process $\mathbf{Y}$ is stationary and isotropic. In these models, distance will be written as $||\mathbf{h}|| = \mathrm{h}$, the nugget will be $\lambda^2 > 0$, the partial sill will be $\sigma^2 > 0$ and the range will be $\phi > 0$. One of the most common ways to model spatial correlation is using a stationary and isotropic Matérn covariance function. The Matérn function is defined for two locations $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$, where $\mathbf{h} = ||\mathbf{s}_i - \mathbf{s}_j||$, as:

$$\mathrm{C}_Y(\mathrm{h}) = \begin{cases} \sigma^2 + \lambda^2 & \mathrm{h} = 0\,, \\ \sigma^2 \frac{\left(\frac{\mathrm{h}}{\phi}\right)^\nu}{2^\nu \Gamma(\nu)} K_\nu\left(\frac{\mathrm{h}}{\phi}\right) & \mathrm{h} > 0\,, \end{cases} \tag{2.61}$$

where:

- $\nu > 0$ is a smoothness parameter;

- $\Gamma(\cdot)$ is the gamma function; and
- $K_\nu(\cdot)$ is a modified Bessel function of the second kind.

The semi-variogram for the Matérn function is, therefore:

$$\gamma_Y(\mathrm{h}) = \begin{cases} 0 & \mathrm{h} = 0\,, \\ \lambda^2 + \sigma^2 \left(1 - \frac{\left(\frac{\mathrm{h}}{\phi}\right)^\nu}{2^\nu \Gamma(\nu)}\right) & \mathrm{h} > 0\,. \end{cases} \tag{2.62}$$

There are two common simplified models depending on the smoothness parameter, which will be discussed in further detail here. Firstly, if $\nu = \frac{1}{2}$, the function is known as the exponential covariance function:

$$\mathrm{C}_Y(\mathrm{h}) = \begin{cases} \sigma^2 + \lambda^2 & \mathrm{h} = 0\,, \\ \sigma^2 \exp\left(-\frac{\mathrm{h}}{\phi}\right) & \mathrm{h} > 0\,, \end{cases} \tag{2.63}$$

and the respective semi-variogram is:

$$\gamma_Y(\mathrm{h}) = \begin{cases} 0 & \mathrm{h} = 0\,, \\ \lambda^2 + \sigma^2 \left(1 - \exp\left(-\frac{\mathrm{h}}{\phi}\right)\right) & \mathrm{h} > 0\,. \end{cases} \tag{2.64}$$

However, the exponential covariance function is quite rough so the second common version of the Matérn function is very smooth, the Gaussian covariance function, where $\nu \to \infty$:

$$\mathrm{C}_Y(\mathrm{h}) = \begin{cases} \sigma^2 + \lambda^2 & \mathrm{h} = 0\,, \\ \sigma^2 \exp\left(-\left(\frac{\mathrm{h}}{\phi}\right)^2\right) & \mathrm{h} > 0\,, \end{cases} \tag{2.65}$$

and the respective semi-variogram is:

$$\gamma_Y(\mathrm{h}) = \begin{cases} 0 & \mathrm{h} = 0\,, \\ \lambda^2 + \sigma^2 \left(1 - \exp\left(-\left(\frac{\mathrm{h}}{\phi}\right)^2\right)\right) & \mathrm{h} > 0\,. \end{cases} \tag{2.66}$$

The choice of the smoothness parameter ν adds estimation burden so to achieve a sensible level of smoothness, $\nu = \frac{3}{2}$ is recommended by [63].

Kriging

Commonly, the main goal in geostatistics is to predict the measurements of a process at a new location $\mathbf{s}_0$. The most famous method for predicting geostatistical process is Kriging [121]. To perform Kriging for a random two-dimensional vector $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, a property of the multivariate Gaussian distribution is used. Therefore, in this thesis, kriging and Gaussian Processes (GP) are used interchangeably. Suppose that:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathrm{N}\left[\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix} \right],$$

then the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$ is:

$$\mathbf{X}_1|\mathbf{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}}^{-1}\boldsymbol{\Sigma_{21}}). \tag{2.67}$$

The joint geostatistical process at n data locations $\mathbf{Y}$ is a stationary process with mean $\boldsymbol{\mu}_Y$ and a covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is defined by a stationary and isotropic covariance function $\mathrm{Cov}_Y(h, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\sigma^2, \lambda^2, \phi\}$ is the set of parameters to be estimated based on a function of one's choice from Subsection 2.3.1. For the joint geostatistical process $\mathbf{Y}$ and a prediction location $Y(\mathbf{s}_0)$, the property is re-written as:

$$\mathbf{Y}^* = \begin{pmatrix} Y(\mathbf{s}_0) \\ \mathbf{Y} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_Y \\ \mu_Y \mathbf{1} \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}_Y(\mathbf{0}, \boldsymbol{\theta}) & \mathbf{C}_Y(\mathbf{s}_0, \boldsymbol{\theta})^\top \\ \mathbf{C}_Y(\mathbf{s}_0, \boldsymbol{\theta}) & \boldsymbol{\Sigma}(\boldsymbol{\theta}) \end{pmatrix} \right), \tag{2.68}$$

where:

- $\mathrm{Cov}_Y(\mathbf{0}, \boldsymbol{\theta}) = \mathrm{Var}(Y(\mathbf{s}_0))$ is a scalar for the variance at the new point $Y(\mathbf{s}_0)$; and
- $\mathbf{C}_Y(\mathbf{s}_0, \boldsymbol{\theta}) = (\mathrm{Cov}_Y(Y(\mathbf{s}_0), Y(\mathbf{s}_1)), \ldots, \mathrm{Cov}_Y(Y(\mathbf{s}_0), Y(\mathbf{s}_n)))$ is a vector $(n \times 1)$ containing the covariances between the new point $Y(\mathbf{s}_0)$ with each of the locations in the set $\mathbf{Y}(\mathbf{s})$.

Then using the property of the multivariate Gaussian distribution gives a predictor:

$$\mathbb{E}[Y(\mathbf{s}_0)|\mathbf{Y}] = \mu_Y + \mathbf{C}_Y(\mathbf{s}_0, \boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{Y} - \mu_Y \mathbf{1}), \tag{2.69}$$

with variance:

$$\text{Var}[\text{Y}(\mathbf{s}_0)|\mathbf{Y}] = \text{Cov}_Y(\mathbf{0}, \boldsymbol{\theta}) - \mathbf{C}_Y(\mathbf{s}_0, \boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{C}_Y(\mathbf{s}_0, \boldsymbol{\theta})\,. \tag{2.70}$$

However, $(\mu_Y, \boldsymbol{\theta})$ are unknown, hence, the predictor is:

$$\text{Y}(\mathbf{s}_0)|\mathbf{Y} \sim \text{N}\left(\widehat{\mathbb{E}[\text{Y}(\mathbf{s}_0)|\mathbf{Y}]}, \widehat{\text{Var}[\text{Y}(\mathbf{s}_0)|\mathbf{Y}]}\right)\,, \tag{2.71}$$

where the universal Kriging predictor is:

- $\widehat{\mathbb{E}[\text{Y}(\mathbf{s}_0)|\mathbf{Y}]} = \widehat{\mu}_Y + \mathbf{C}_Y(\mathbf{s}_0, \widehat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}(\mathbf{Y} - \widehat{\mu}_Y \mathbf{1})$; and
- $\widehat{\text{Var}[\text{Y}(\mathbf{s}_0)|\mathbf{Y}]} = \text{Cov}_Y(\mathbf{0}, \widehat{\boldsymbol{\theta}}) - \mathbf{C}_Y(\mathbf{s}_0, \widehat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{C}_Y(\mathbf{s}_0, \widehat{\boldsymbol{\theta}})$.

This yields a $100\alpha\%$ prediction interval of the form:

$$\widehat{\mathbb{E}[\mathbf{Y}(\mathbf{s}_0)|\mathbf{Y}]} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\widehat{\text{Var}[\mathbf{Y}(\mathbf{s}_0)|\mathbf{Y}]}}\,, \tag{2.72}$$

where $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the inverse cumulative distribution function of the normal distribution for a chosen significance level α.

$\widehat{\boldsymbol{\theta}} = (\sigma^2, \lambda^2, \phi)$ is estimated using MLE. Let $\mathbf{Y} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\mathbf{X}$ ($n \times p$ where p is the number of covariates) is the matrix of covariates and $\boldsymbol{\beta}$ ($p \times 1$) is a vector of parameters. The parameters λ^2 and ϕ do not have a closed form solution for $\widehat{\lambda}^2$ and $\widehat{\phi}$, and numerical optimisation methods should be used to estimate them. For more information on the estimation, refer to [63] and [50]. However, assuming an exponential autocovariance model $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2 \text{exp}\left(-\frac{\mathbf{D}}{\phi}\right) + \lambda^2 \mathcal{I}$ with $\mathbf{D}$ ($n \times n$) being a distance matrix between the spatial locations in the set, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ have closed form solutions as follows:

$$\widehat{\boldsymbol{\beta}}(\nu^2, \phi) = (\mathbf{X}^\top \mathbf{V}(\widehat{\nu}^2, \widehat{\phi})^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}(\widehat{\nu}^2, \widehat{\phi})^{-1} \mathbf{Y}\,, \tag{2.73}$$

$$\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}, \nu^2, \phi) = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top \mathbf{V}(\widehat{\nu}^2, \widehat{\phi})^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,, \tag{2.74}$$

where $\mathbf{V}(\widehat{\nu}^2, \widehat{\phi}) = \text{exp}\left(-\frac{\mathbf{D}}{\phi}\right) + \nu^2 \mathcal{I}$ is a variance matrix ($n \times n$) and $\nu^2 = \frac{\lambda^2}{\sigma^2}$ is the noise to signal ratio.

2.3.2 Spatio-Temporal data modelling

The air pollution data presented in Chapter 8 are not just spatial data, as the observations were taken at exact locations over T regular time intervals. Therefore, spatio-temporal modelling is required. In this subsection, main quantities and concepts related to spatio-temporal data are presented as described in [51] and [179], which can be referred to for further information.

A spatio-temporal process is defined by extending a geostatistical process $\mathbf{Y}(\mathbf{s})$ to $\mathbf{Y}(\mathbf{s}, t)$ where $t \in \mathbb{N}$ is equally spaced (discretised) time steps. In general, the observed data are $\mathbf{Y}_{n\times T} = \{Y(\mathbf{s}_i, t_j)\}$ (where $i = 1, \ldots, n$ and $j = 1, \ldots, T$) for the set of spatial locations $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and time points $t = 1, 2, \ldots, T$. As in the previous subsections, spatial and temporal lags will be denoted as $\mathbf{h}$ and τ respectively. The mean of a spatio-temporal process $\mathbf{Y}(\mathbf{s}, t)$ is:

$$\mu_Y(\mathbf{s}, t) = \mathbb{E}[\mathbf{Y}(\mathbf{s}, t)]\,. \tag{2.75}$$

The covariance function for two locations $\mathbf{s}$ and $\mathbf{p}$ and two time points t and v is:

$$\mathrm{C}_Y(\mathbf{s}, \mathbf{p}, t, v) = \mathrm{Cov}(Y(\mathbf{s}, t), Y(\mathbf{p}, v)) = \mathbb{E}[(Y(\mathbf{s}, t) - \mu_Y(\mathbf{s}, t))(Y(\mathbf{p}, v) - \mu_Y(\mathbf{p}, v))]\,. \tag{2.76}$$

Naturally, the variance of a spatio-temporal process $\mathbf{Y}(\mathbf{s}, t)$ is defined as a special case of the covariance when $\mathbf{s} = \mathbf{p}$ and $t = v$:

$$\begin{aligned} \mathrm{Var}(Y(\mathbf{s}, t)) &= \mathrm{C}_Y(Y(\mathbf{s}, t), Y(\mathbf{s}, t)) \\ &= \mathrm{Cov}[Y(\mathbf{s}, t), Y(\mathbf{s}, t)] \\ &= \mathbb{E}[(Y(\mathbf{s}, t) - \mu_Y(\mathbf{s}, t))^2] \\ &= \sigma^2_Y(\mathbf{s}, t)\,. \end{aligned} \tag{2.77}$$

Stationarity/isotropy

Second-order stationarity of a spatio-temporal process $\mathbf{Y}(\mathbf{s}, t)$ is defined for any locations $\mathbf{s}$ and $\mathbf{p}$, and times t and v as occurring if:

- $\mathbb{E}[\mathbf{Y}(\mathbf{s}, t)] = \mu_Y(\mathbf{s}, t) = \mu_Y$; and
- $\mathrm{C}_Y(\mathbf{s}, \mathbf{p}, t, v) = \mathrm{Cov}[\mathbf{Y}(\mathbf{s}, t), \mathbf{Y}(\mathbf{p}, v)] = \mathrm{Cov}[\mathbf{Y}(\mathbf{s}-\mathbf{p}, t-v), \mathbf{Y}(\mathbf{0}, 0)] = \mathrm{C}(\mathbf{s}-\mathbf{p}, t-v)$.

From the covariance function, the variogram function is defined for the spatial distance lag $\mathbf{h}$ and time lag τ as:

$$\gamma(\mathbf{h}, \tau) = \text{Var}[Y(\mathbf{s} + \mathbf{h}, t + \tau) - Y(\mathbf{s}, t)]\,. \tag{2.78}$$

For a spatio-temporal process $\mathbf{Y}(\mathbf{s}, t)$ with second order stationarity, it holds that:

$$\gamma(\mathbf{h}, \tau) = \text{C}(\mathbf{0}, 0) - \text{C}(\mathbf{h}, \tau)\,. \tag{2.79}$$

Empirically, the variogram for the spatio-temporal process $\mathbf{Y}(\mathbf{s}, t)$ is estimated by using pairs of points at distances $\mathbf{h}$ from each other and time lag τ:

$$\widehat{\gamma}(\mathbf{h}, \tau) = \frac{1}{2|N(\mathbf{h}, \tau)|} \sum_{N(\mathbf{h},\tau)} \{[Y(\mathbf{s}_i, t_i) - Y(\mathbf{s}_j, t_j)]^2\}\,, \tag{2.80}$$

where $N(\mathbf{h}, \tau) = \{[(\mathbf{s}_i, t_i), (\mathbf{s}_j, t_j)]^2 : ||\mathbf{s}_i - \mathbf{s}_j|| = \mathbf{h} \text{ and } |t_i - t_j| = \tau\}$.

Separability

In the spatial setting, for the process $\mathbf{Y}$, the covariance function is defined to be symmetric when $\text{C}_Y(\mathbf{s} - \mathbf{p}) = \text{C}_Y(\mathbf{p} - \mathbf{s})$. However, for the covariance function of a spatio-temporal process $\mathbf{Y}$, the symmetry does not hold by definition. The full symmetry requires that:

- $\text{C}_Y(\mathbf{s} - \mathbf{p}, t - v) = \text{C}_Y(\mathbf{p} - \mathbf{s}, t - v)$; and
- $\text{C}_Y(\mathbf{s} - \mathbf{p}, t - v) = \text{C}_Y(\mathbf{s} - \mathbf{p}, v - t)$.

From the set of fully symmetric covariance functions, there is a simplification called **separable** when:

$$\text{C}_Y(\mathbf{s} - \mathbf{p}, t - v) = \text{C}_Y^*(\mathbf{s} - \mathbf{p})\widetilde{\text{C}}_Y(t - v)\,. \tag{2.81}$$

From here, it can be seen that the spatio-temporal covariance can be separated into two parts - spatial (C_Y^*) and temporal ($\widetilde{\text{C}}_Y$). These two parts could be taken to be any model described in Subsections 2.3.1 and 2.1.2, respectively. For instance, a combination of a spatial Gaussian covariance function with an AR(1) process:

$$\begin{aligned} \text{C}_Y(\mathbf{s} - \mathbf{p}, t - v) &= \{\text{C}_Y^*(||\mathbf{s} - \mathbf{p}|| = \mathbf{h})\} \left\{\widetilde{\text{C}}_Y(t - v = \tau)\right\} \\ &= \left\{\sigma^2 \text{exp}\left(-\left(\frac{\mathbf{h}}{\phi}\right)^2\right)\right\} \left\{\sigma_Z^2 \frac{\zeta^\tau}{1 - \zeta^2}\right\}\,. \end{aligned} \tag{2.82}$$

Prediction

In spatio-temporal modelling, in a similar way to spatial modelling, it is of key interest to be able to predict the measurements of a spatio-temporal process at a new space-time location. Kriging can be used again as described in Subsection 2.3.1. The method only requires a covariance function to be defined. Therefore, for the spatio-temporal process $\mathbf{Y}$ with a set of n locations $\mathbf{s}_i$ and T time intervals t_j, for the unknown location $\mathbf{s}_0$ and a new time point t_0, the universal Kriging predictor for the measurement $Y(\mathbf{s}_0, t_0)$ is again:

- $\mathbb{E}[\widehat{Y(\mathbf{s}_0, t_0)}|\mathbf{Y}] = \widehat{\mu}_Y + \mathbf{C}_Y(\mathbf{s}_0, t_0, \widehat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{Y} - \widehat{\mu}_Y \mathbf{1})$; and
- $\mathrm{Var}[\widehat{Y(\mathbf{s}_0, t_0)}|\mathbf{Y}] = \mathrm{Cov}(\mathbf{0}, 0, \widehat{\boldsymbol{\theta}}) - \mathbf{C}_Y(\mathbf{s}_0, t_0, \widehat{\boldsymbol{\theta}})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{C}_Y(\mathbf{s}_0, t_0, \widehat{\boldsymbol{\theta}})$;

where:

- $\mathbf{Cov}_Y(\mathbf{s}_0, t_0, \widehat{\boldsymbol{\theta}}) = \mathrm{Var}(Y(\mathbf{s}_0, t_0, \widehat{\boldsymbol{\theta}}))$ is a scalar for the variance at the new point $Y(\mathbf{s}_0, t_0)$; and
- $\mathbf{C}_Y(\mathbf{s}_0, t_0, \widehat{\boldsymbol{\theta}}) = [\mathrm{Cov}_Y(Y(\mathbf{s}_0, t_0, Y(\mathbf{s}_1, t_1))), \ldots, \mathrm{Cov}_Y(Y(\mathbf{s}_0, t_0, Y(\mathbf{s}_n, t_T)))]$ is a vector $((n \times T) \times 1)$ of the covariances between the new point $Y(\mathbf{s}_0, t_0)$ and the data set $\mathbf{Y}$.

A $100 - \alpha\%$ prediction interval could be estimated as:

$$\mathbb{E}[\widehat{\mathbf{Y}(\mathbf{s}_0, t_0)}|\mathbf{Y}] \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\mathrm{Var}[\widehat{\mathbf{Y}(\mathbf{s}_0, t_0)}|\mathbf{Y}]}\,. \tag{2.83}$$

2.4 Emulation of computer simulations

This section focuses on the key features of emulation of computer models. As discussed in Section 1.6, an emulator is a statistical model which predicts the output of computer models for untried inputs based on a set of runs of the computer model. Therefore, it is crucial to choose on which inputs the computer models will be run. In the literature review, Latin Hypercube sampling is one of the most common ways of selecting the inputs for which the computer model will be used. This section introduces Latin Hypercube sampling and the general form of emulation.

2.4.1 Latin Hypercube

A Latin Hypercube (LHC) design is used to choose the inputs for the ADMS-Urban simulation scenarios modelled in Chapters 4 through 8. In this section, background on Latin Hypercube sampling is provided. The Latin Hypercube design was first introduced in [125] as a type of stratified sampling for the inputs of simulation models. The sampling is done in a way which guarantees that the ranges of each of the inputs is fully explored. Therefore, the set of inputs to be run from the simulation model provide a very close approximation to the real variability in the simulation scenarios. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ be the input space for p inputs to be explored. Then, the range of each $\mathbf{x}_i$ $(i = 1, \ldots, p)$ is divided into N strata each with equal marginal probability $\frac{1}{N}$ and one sample is taken per stratum. Hence, a sample $\mathbf{x}_{ij}$ for $j = 1, \ldots, N$ is obtained. The components of the samples for each $\mathbf{x}_i$ are matched at random. Therefore, the set of input values for which the simulation model is to be run is chosen. A computationally inexpensive extension to LHC was proposed in [132], where a distancing criterion (maximin) is applied to ensure the design points are spaced out by calculating the distances between the possible scenarios in the LHC design space and choosing the largest distances.

2.4.2 Emulation

In Chapters 5 through 8, different emulators are built for the different outputs of the ADMS-Urban runs. This subsection provides a general background to emulators, whereas each of Chapters 5, 6, 7 and 8 will provide specific information to the emulator applied in them based on the type of output modelled in the specific chapter.

As previously discussed, a frequentist emulator based on kriging was proposed in [169] and this description is summarised below. Let $\mathbf{y}$ be a vector $(n \times 1)$ of the deterministic output from a simulation model. Then:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\,, \tag{2.84}$$

where:

- $\mathbf{X}$ is a design matrix $(n \times p)$ with an intercept term and $p - 1$ inputs for the simulation model. Each row of $\mathbf{X}$ contains the input values for one run of the simulation model as chosen using the LHC design;
- $\boldsymbol{\beta}$ is a vector $(p \times 1)$ of the fixed parameters to be estimated; and

- $\mathbf{z}$ is an error vector $(n \times 1)$, which follows a normal distribution $\mathbf{z} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $n \times n$ variance-covariance matrix.

Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be the i^{th} and j^{th} input scenarios (i.e. the i^{th} and j^{th} rows of $\mathbf{X}$), then the covariance is estimated as:

$$\boldsymbol{\Sigma}_{ij} = \sigma^2 \text{Corr}(\mathbf{x}_i, \mathbf{x}_j)\,, \tag{2.85}$$

where σ^2 is the overall variance and $\text{Corr}(\mathbf{x}_i, \mathbf{x}_j)$ is a correlation function. Different correlation functions are discussed in further detail in Chapters 5 through 8 depending on the data being modelled. The correlation parameters are estimated using the BFGS algorithm described in Subsection 2.2.4.

Let $\mathbf{R}$ be the correlation matrix $(n \times n)$ for LHC input space such that $\boldsymbol{\Sigma} = \sigma^2 \mathbf{R}$. Then, the fixed effect parameters are estimated using the GLS fit for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{R}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{R}^{-1} \mathbf{y}\,, \tag{2.86}$$

and

$$\widehat{\sigma^2} = \frac{1}{n} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^\top \mathbf{R}^{-1} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)\,. \tag{2.87}$$

Predictions are calculated using a multivariate normal distribution in a similar fashion as described for universal kriging in Subsection 2.3.1. Let $\mathbf{y}_0$ be the output from the simulation model for a set of untested inputs $\mathbf{x}_0$. Then:

$$\widehat{\mathbf{y}_0} = \mathbf{x}_0 \widehat{\boldsymbol{\beta}} + \mathbf{r}^\top \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,, \tag{2.88}$$

where $\mathbf{r}$ is the vector $(n \times 1)$ with the estimated correlations $\text{Corr}(\mathbf{x}_0, \mathbf{x}_i)$ between $\mathbf{x}_0$ and each row of $\mathbf{X}$. The variance of the new observation $\mathbf{y}_0$ is:

$$\text{Var}(\mathbf{y}_0) = \widehat{\sigma^2} - \mathbf{r}^\top \mathbf{R}^{-1} \mathbf{r}. \tag{2.89}$$

However, in more recent years, a wider definition for emulation has been used. Every statistical model used for prediction of untested output from a computer code has been referred to as an emulator as seen in [162]. This wider definition is applied to the model in Chapter 7.

Chapter 3

Using miniature automated sensors to measure air pollution

In this chapter, a sensor package containing electrochemical sensors (ALPHASENSE B2) for NO_2 and O_3, as well as for temperature and humidity with low-energy wireless communication, hardware and battery energy supply "AirSpeck" [15] is assessed with no prior assumptions, which reflects a realistic setting for citizen science applications without any training on how to use the sensor. Therefore, it is investigated how well sensors of lower cost such as the ALPHASENSE B2 can be used to supplement high quality sensors. In order to directly compare the miniature automated sensor (MAS) package with reference monitor observations under ambient conditions, three replicated AirSpeck sensor packages were placed next to the reference sensor at the Edinburgh St. Leonards AURN site for two weeks. The aim of the experiment is to evaluate the accuracy of the AirSpecs in two respects (i) how consistent are the three MAS outputs with each other (i.e. how reproducible are the results); and (ii) how well do the MAS outputs correlate with the reference sensor and, hence, robustly measure air pollution. The rest of this chapter is organised as follows: Section 3.1 presents the study data. Section 3.2 presents a comparison of the performance of the three MAS to each other, whereas Section 3.3 compares the performance of the MAS to the AURN sensor. Concluding remarks and areas for future work are given in Section 3.4. The work in this chapter is part of a pilot project from the EPSRC funded SECURE network EP/M008347/1 (http://www.gla.ac.uk/research/az/secure/).

3.1 Data and study design

3.1.1 Data collection and study region

Three 'AirSpeck' MAS packages [15] were installed by the AirSpeck developers on the fence at the Edinburgh St. Leonards AURN monitoring stations. A picture of two of the sensors at the station is provided in Figure 3.1. St. Leonards is an urban background location within a small park area on the south side of the city. The nearest main road, Pleasance, is approximately 35 metres away and is a busy main road running into the city centre [5]. A map of Edinburgh with the station marked is provided in Figure 3.2. The site is classified for air quality monitoring purposes as an urban background site, which should be representative of general ambient concentrations in the urban environment. The site provided mains power and a secure location for setting up the sensor packages in close proximity to a reference monitor.

'AirSpeck' MAS packages

FIGURE 3.1: Two of the 'AirSpeck' MAS packages at St. Leonards AURN monitoring station.

The reference monitor for NO_2 at this site is a Teledyne API200A chemiluminescence analyser and for O_3, it is a Thermo 49i UV absorbance analyser. Operation and data ratification of the reference instruments is covered by specified procedures that ensure compliance to measurements and metadata objectives specified in the EU Air Quality Directive 2008/50/EC [76]. The reference analyser data are reported as hourly averaged concentrations in μg m^{-3}. Therefore, hourly data are aggregated before being provided from minute MAS measurements to be comparable to the AURN monitor measurements using the method described in [15]. The MAS measurements used in this chapter capture the period from 15:00 on 18/07/2017 to 12:00 on 07/08/2017. In the AirSpeck, each electrochemical sensor measured two voltages (in millivolts (mV)) to represent NO_2 and O_3 pollutant concentrations at two electrodes: the auxiliary electrode (AE) and the working electrode (WE). For the rest of the chapter, the MAS voltages are used but will be referred to as 'NO_2' and 'O_3'. Application notes from the sensor manufacturer

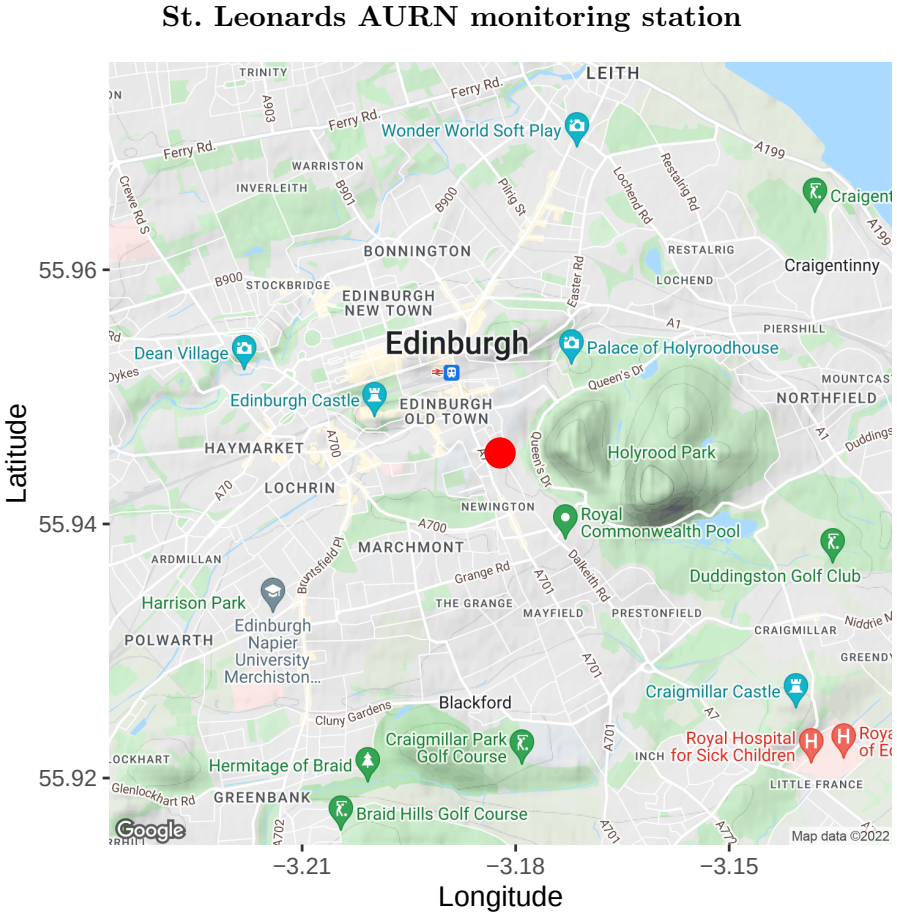

FIGURE 3.2: Map of Edinburgh with a red dot to signify the location of the St. Leonards AURN monitoring station [92].

states that as gas concentrations for both NO_2 and O_3 are typically 20 to 200 parts per billion (ppb) at the roadside, so good design of the sensor, housing and electronics plus intelligent data analysis are all required for air quality measurements [10].

The three AirSpecks are referred to as Sensor 1 (or S1), Sensor 2 (or S2) and Sensor 3 (or S3) for the rest of the chapter. All sensors recorded values for either 'NO_2' or 'O_3' from both WE and AE, respectively. Additionally, the difference for 'O_3 - NO_2' is also modelled as the ALPHASENSE website states that "the difference (in voltage) between the two sensors gives the O_3 concentration" [10]. Furthermore, the three AirSpecks recorded the temperature (°C) and the relative humidity (RH, %), which were used as covariates when modelling concentrations. Due to the noise in the temperature and RH, the averaged values across the three MAS units are used when modelling.

3.1.2 MAS data

The time series plots of the AE and WE values and for NO_2, O_3 and O_3 - NO_2 MAS are displayed in Figure 3.3. The raw hourly measurements show considerable fluctuations,

therefore, lowess (locally weighted smoothing, see Subsection 2.2.2) curves are added to each plot to check the general trends in the data, e.g. long term drift in the voltage. Sensor 3 has failed to record continuously as there are 150/478 data points missing (31.4%) which result in visible gaps in the time series (Figure 3.3 a)). Sensor 1 failed to record 14/478 (2.9%) observations, while Sensor 2 had the highest data capture with only 2/478 (0.4%) missing observations. Although there are missing data, the data were not interpolated as interpolation would interfere with the clarity of the data (i.e. focus solely on the recorded measurements). The smoothed curves show an almost constant concentration throughout the period, with Sensor 3 measuring a slight decrease at the end of the period on all plots, which may indicate degradation of the sensors or components of that AirSpeck unit.

Additionally, the histograms of the AE and WE are examined in Figure 3.4. The AE voltages for NO_2 show a clear overlap between Sensors 1 and 2, whilst Sensor 3 is mostly separate, whereas the NO_2 WE voltages seem to overlap for all three sensors. For both the AE and WE voltages of O_3, it appears that sensors 1 and 3 overlap each other quite well with Sensor 2 being mostly separate from the other two. The biggest difference between the AE and WE voltages is for O_3-NO_2, where the distributions for each sensor are mostly separate, whilst for the WE voltage there is a very good overlap between the three measurements.

Pollutant	**Sensor**	**Mean Difference**		**Correlation**	
	Voltage	**AE**	**WE**	**AE**	**WE**
NO_2	**S1 vs. S2**	15.13	137.47	0.11 (0.03, 0.19)	0.89 (0.87, 0.90)
	S1 vs. S3	170.15	12.71	-0.04 (-0.15, 0.07)	0.49 (0.40, 0.57)
	S2 vs. S3	167.43	161.62	-0.14 (-0.24, -0.03)	0.51 (0.42, 0.58)
O_3	**S1 vs. S2**	112.09	-180.57	0.06 (-0.03, 0.14)	0.60 (0.54, 0.65)
	S1 vs. S3	38.91	55.54	0.22 (0.11, 0.32)	0.50 (0.41, 0.58)
	S2 vs. S3	40.58	282.94	-0.07 (-0.18, 0.04)	0.36 (0.26, 0.45)
O_3 - NO_2	**S1 vs. S2**	80.27	–70.10	0.23 (0.14, 0.31)	0.78 (0.74, 0.81)
	S1 vs. S3	-131.25	42.84	0.81 (0.77, 0.85)	0.87 (0.83, 0.89)
	S2 vs. S3	-208.01	121.32	0.20 (0.09, 0.30)	0.75 (0.70, 0.80)

TABLE 3.1: Mean differences and Pearson's correlation coefficients (and their corresponding 95% confidence intervals) between the hourly measurements (in mV) from the three MAS.

Overall, there is no pattern about the mean differences between the sensors regardless of whether AE or WE voltages are examined as it can be seen in Table 3.1. However, looking at the correlations, a pattern emerges. The WE hourly measurements have much

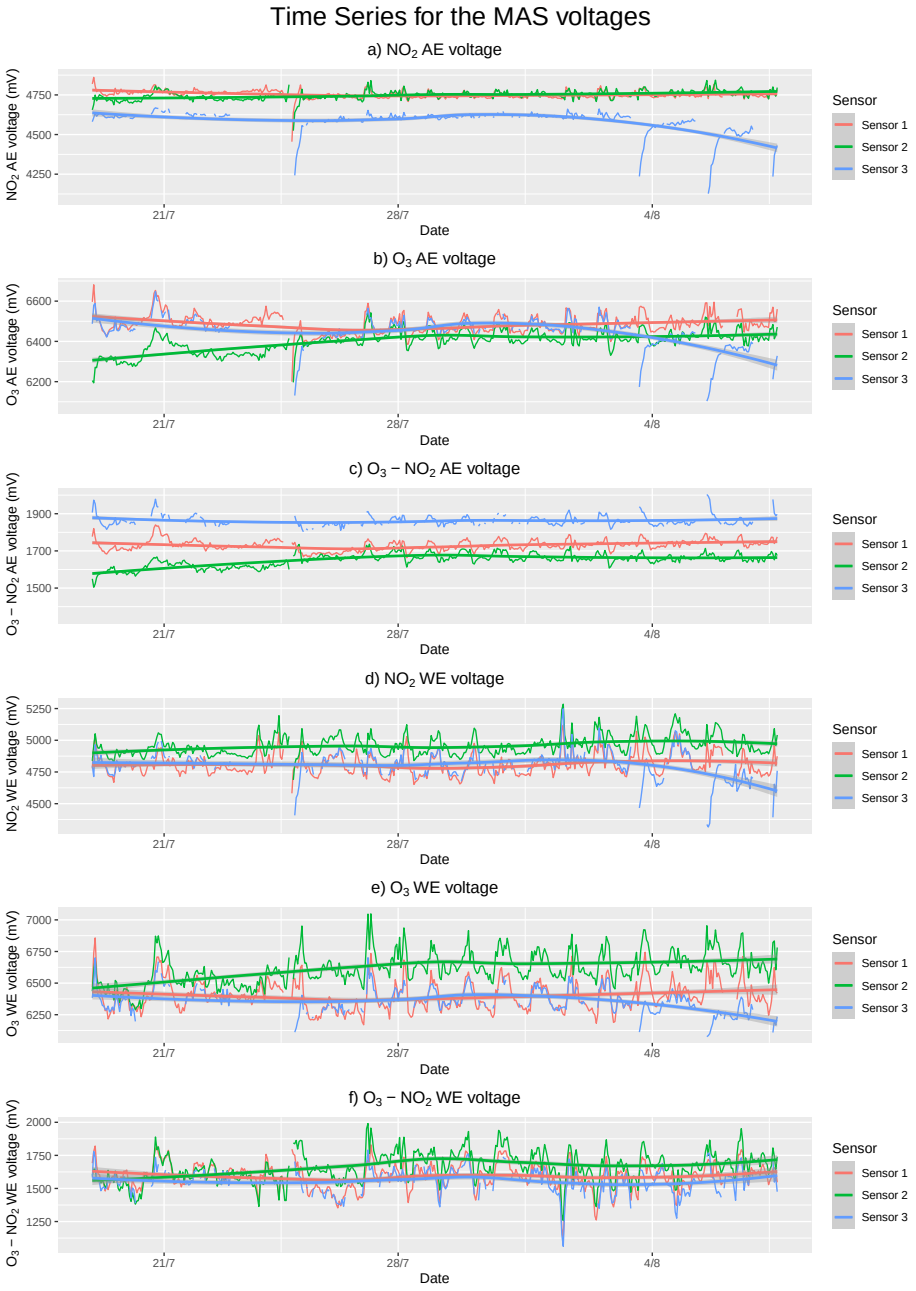

FIGURE 3.3: Time series of all hourly measurements (in mV) taken by the MAS at hourly intervals with a lowess smoothing line.

stronger positive linear relationships between the sensors' measurements, whereas the AE hourly measurements seem to have weak relationships between each other with only the O_3 - NO_2 measurements between Sensors 1 and 3 but even in this case, the WE correlation is stronger. Similarly, in Figure 3.4, it appears that there is more overlap of the hourly voltage measurements for WE than for AE.

Figure 3.3 a) shows that the NO_2 AE voltages recorded by Sensors 1 and 2 are very similar to each other and follow approximately (to the eye) the same high resolution variation and the same underlying trend. The units operated consistently with each

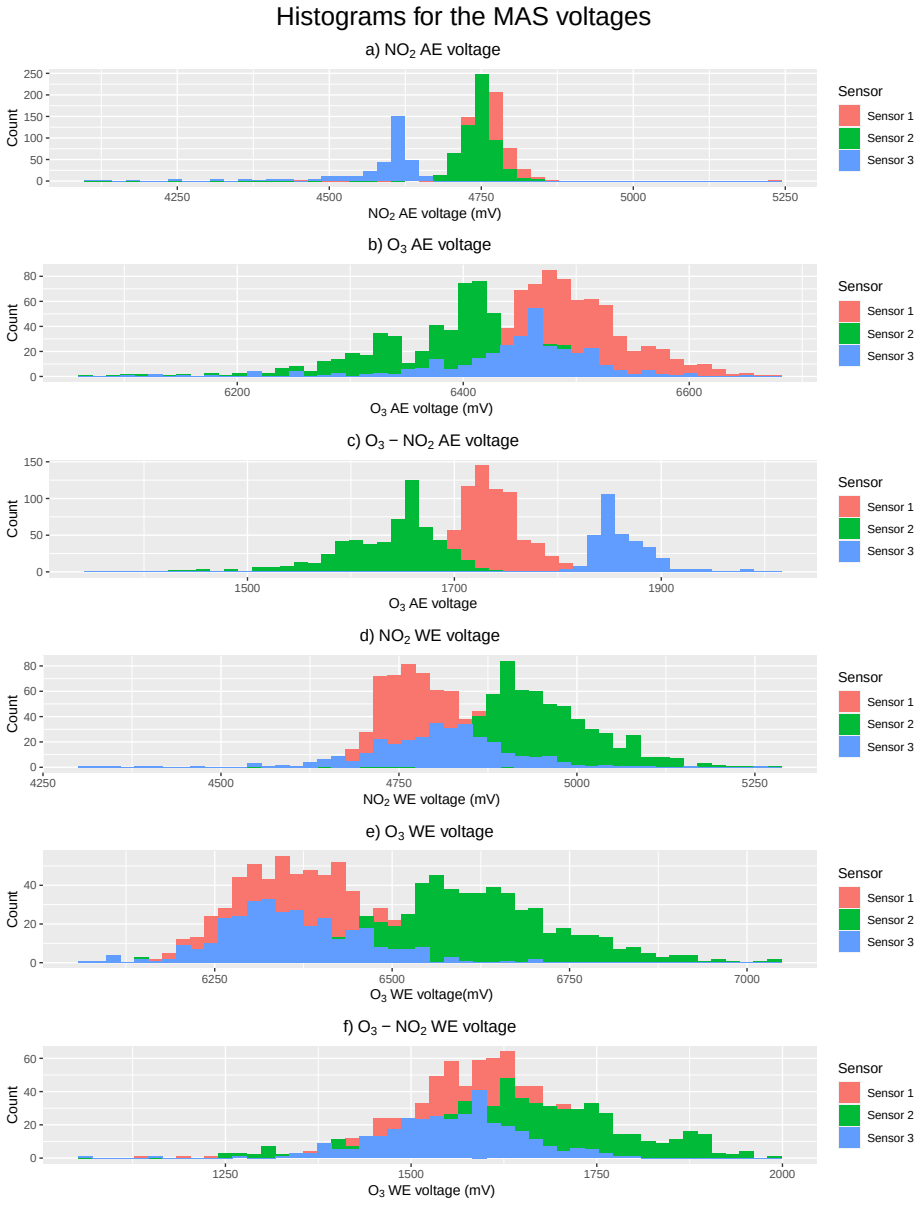

FIGURE 3.4: Histograms of all hourly measurements (in mV) taken by the MAS at hourly intervals.

other with an average difference of 15.13 millivolts (mV) as shown in Table 3.1. However, in Table 3.1, the Pearson's correlation coefficient (referred to as correlation or r for the rest of the chapter) between the hourly data is 0.11, which indicates a weak agreement between the two NO_2 AE voltages. Sensor 3 consistently has lower AE voltages with the average difference $\sim$ 170 mV to both Sensor 1 and Sensor 2. From Table 3.1, the correlation of Sensor 3 to Sensor 1 is -0.04 and its 95% confidence interval contains zero, and to Sensor 2 is -0.14, both of which are close to zero and indicate again poor association. The histogram in Figure 3.4 a) shows that the hourly measurements from Sensors 1 and 2 almost perfectly overlap each other, whereas Sensor 3 has taken lower hourly measurements.

The three AirSpeck units have recorded very different values of O_3 AE voltage as shown in Figure 3.3 b), with the sensors variability overlapping or diverging from each other over the measurement period. This issue is further highlighted by the smoothed curves. In the beginning of the period, Sensor 2 has measured lower values than the other two sensors. In Table 3.1, it is seen that on average, Sensor 1 and Sensor 2 differ by 112.09 mV and the correlation is 0.06 with zero in the 95% confidence interval, which shows that there is no association between the hourly measurements from the two sensors. The hourly measurements from Sensor 1 and Sensor 3 differ by 38.91 mV and have a correlation of 0.22, which again indicates a very weak relationship between the hourly measurements as seen in Table 3.1. Finally, in Table 3.1, on average, Sensor 2 and Sensor 3 differ by 40.58 mV and have a correlation coefficient of -0.07 with zero in the 95% confidence interval again indicating no relationship between the hourly measurements. The histograms of the O_3 AE voltages in Figure 3.4 b) shows that there is quite a large spread in the hourly measurements for all sensors with Sensors 1 and 2 being almost separate from each other, while Sensor 3 spans across the bins for the other two sensors.

In Figure 3.3 c), the hourly measurements of O_3 - NO_2 AE voltage are quite different from each other and there is no overlap of the points. The lowess smooth lines are almost straight for each sensor creating the impression of almost parallel lines. This is further highlighted by histograms in Figure 3.4 c), where the hourly measurements from all the Sensors are quite different from each other. From Table 3.1, Sensor 1 and Sensor 2 differ on average by 80.27 mV and have a correlation of 0.23. Interestingly, Sensor 1 and Sensor 3 have an average difference of 131.25 mV but have strong correlation of 0.81. Finally, Sensor 2 and Sensor 3 on average differ by 208 mV and have weak correlation of 0.20.

The NO_2 WE hourly measurements Sensor 1 and Sensor 3 in Figure 3.3 d) are well-calibrated to each other (as opposed to AE hourly measurements for Sensor 1 and Sensor 2) with an average difference of 12.71 mV (Table 3.1). The smoothed curves for the two sensors track each other well and for Sensor 1 and Sensor 3 the correlation is r=0.49 as shown in Table 3.1, which is similar to the correlation between Sensor 2 and Sensor 3 (r=0.51). Sensor 1 and Sensor 2 have similar fluctuations and correlation of 0.89 - the highest correlation between any two sensors. The offset between Sensor 1 and Sensor 3 to Sensor 2 was 137.47 mV and 161.62 mV respectively. These differences are noticed in Figure 3.4 d), where the Sensors 1 and 3 almost overlap each other but Sensor 2 has taken higher hourly measurements.

The O_3 WE time series in Figure 3.3 e) show that the hourly measurements from all three sensors appear well-calibrated with each other, with just Sensor 3 measuring a small decrease in the end of the period. In Table 3.1, Sensor 1 and Sensor 2 differ from

each other by 180.57 mV on average and have a correlation of 0.60, while Sensor 1 and Sensor 3 differ on average by 55.54 mV and have a correlation of 0.50. Finally, Sensor 2 and Sensor 3 differ on average by 282.94 mV and have a correlation of 0.36. This is all confirmed by the histograms of the hourly measurements in Figure 3.4 e), where the hourly measurements by Sensor 1 and 3 are almost overlapping each other, whereas the hourly measurements from Sensor 3 are higher.

On the other hand, in Figure 3.3 f), the O_3 - NO_2 WE voltage measurements from the three sensors appear well-calibrated to each other and track each other quite well. Furthermore, the histograms of the hourly measurements in Figure 3.4 f) overlap each other quite well. All sensors have strong correlations between each other varying from 0.75 to 0.87 as shown in Table 3.1. On average, Sensor 1 and Sensor 2 differ by 70.10 mV, Sensor 1 and Sensor 3 differ by 42.84 and Sensor 2 and Sensor 3 differ by 121.32 mV (Table 3.1).

Overall, based on the initial comparisons for all NO_2, O_3 and NO_2 - O_3, it appears that the AE voltages are not as reliable as the WE voltages. The WE voltages appear more in agreement with each other. This is likely the results of different ways the voltages are produced as described in [10].

3.1.3 Reference data

The hourly interval reference data were obtained from the AURN reference monitor (ratified data downloaded on 29/01/2018 from the Scottish Air Quality website from https://uk-air.defra.gov.uk/). The NO_2 and O_3 data are concentrations in μg m^{-3}, and, therefore, on a different scale compared to the changes in voltage which form the raw output of the MAS.

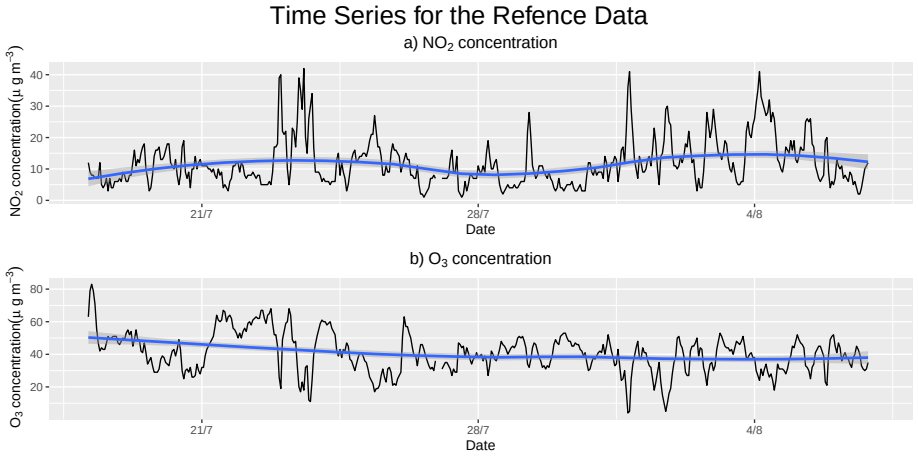

FIGURE 3.5: Time series of the NO_2 and O_3 concentrations (in μg m^{-3}) recorded by the reference AURN monitor at hourly intervals with a lowess smoothing line.

Overall, the hourly measurements from the reference sensor in Figure 3.5 are also noisy as were the MAS measurements. There are no visible gaps in the plots in Figure 3.5 and that is because there are only 6 missing observations (1.26%) for each pollutant, occurring at the same hours. However, the smoothed curves do not show much trend in the data as the lines are almost horizontal. For O_3, there appears to be a slight decreasing trend.

The histogram for the reference NO_2 shows the hourly measurements are slightly right skewed as seen in Figure 3.6 a). This is quite different to the histograms of all MAS measurements in Figure 3.4, where the measurements appeared symmetric. However, the histogram for reference O_3 is more symmetrical and the data look normally distributed as seen in Figure 3.6 b) and, hence, more similar to the histograms for the MAS measurements for O_3.

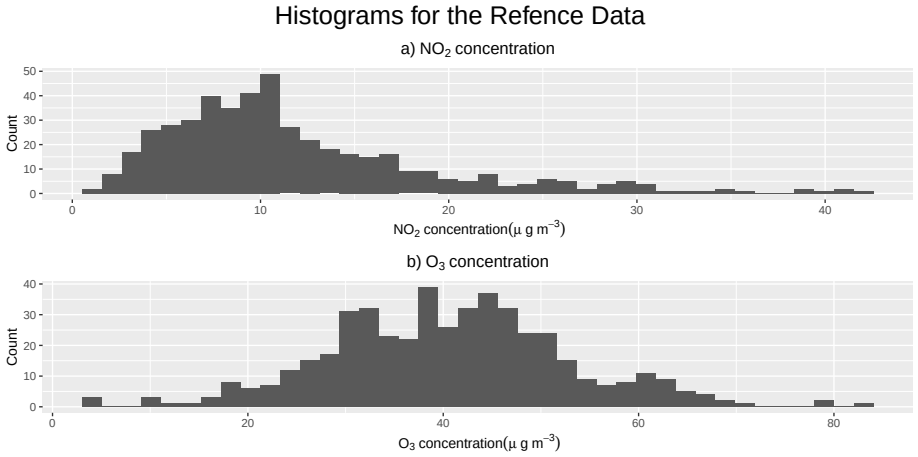

FIGURE 3.6: Histograms of the NO_2 and O_3 concentrations (in μg m^{-3}) recorded by the reference AURN monitor at hourly intervals.

To visualise the relationships between the reference sensor measurements and the MAS' measurements Figure 3.7 provides a scatterplot comparing the MAS measurements to the AURN ones. No filtering was performed on the MAS measurements to keep the clarity of the data. It is clear that there is little or no correlation between the reference sensor hourly measurements with all AE hourly measurements from the MAS for both pollutants in Figure 3.7 a), b), c), d), e), f), g), h) and i), with correlations ranging from -0.35 to 0.14. However, there are moderate linear relationships between the reference sensors and the NO_2 WE hourly measurements in Figure 3.7 j), k) and l) with correlations ranging from 0.38 to 0.63. There is weak to moderate linear relationships between the reference sensor and the MAS measurements for O_3 WE in Figure 3.7 m), n) and o) as indicated by the correlation coefficients ranging from 0.02 to 0.37. There are weak to moderate linear relationships between the reference sensor and the MAS measurements for O_3 - NO_2 WE in Figure 3.7 p), q) and r) with correlation coefficient ranging from 0.28 to 0.56. Overall, it appears that the WE MAS measurements are in agreement with

themselves and the reference sensor measurements, whereas the AE MAS measurements are not.

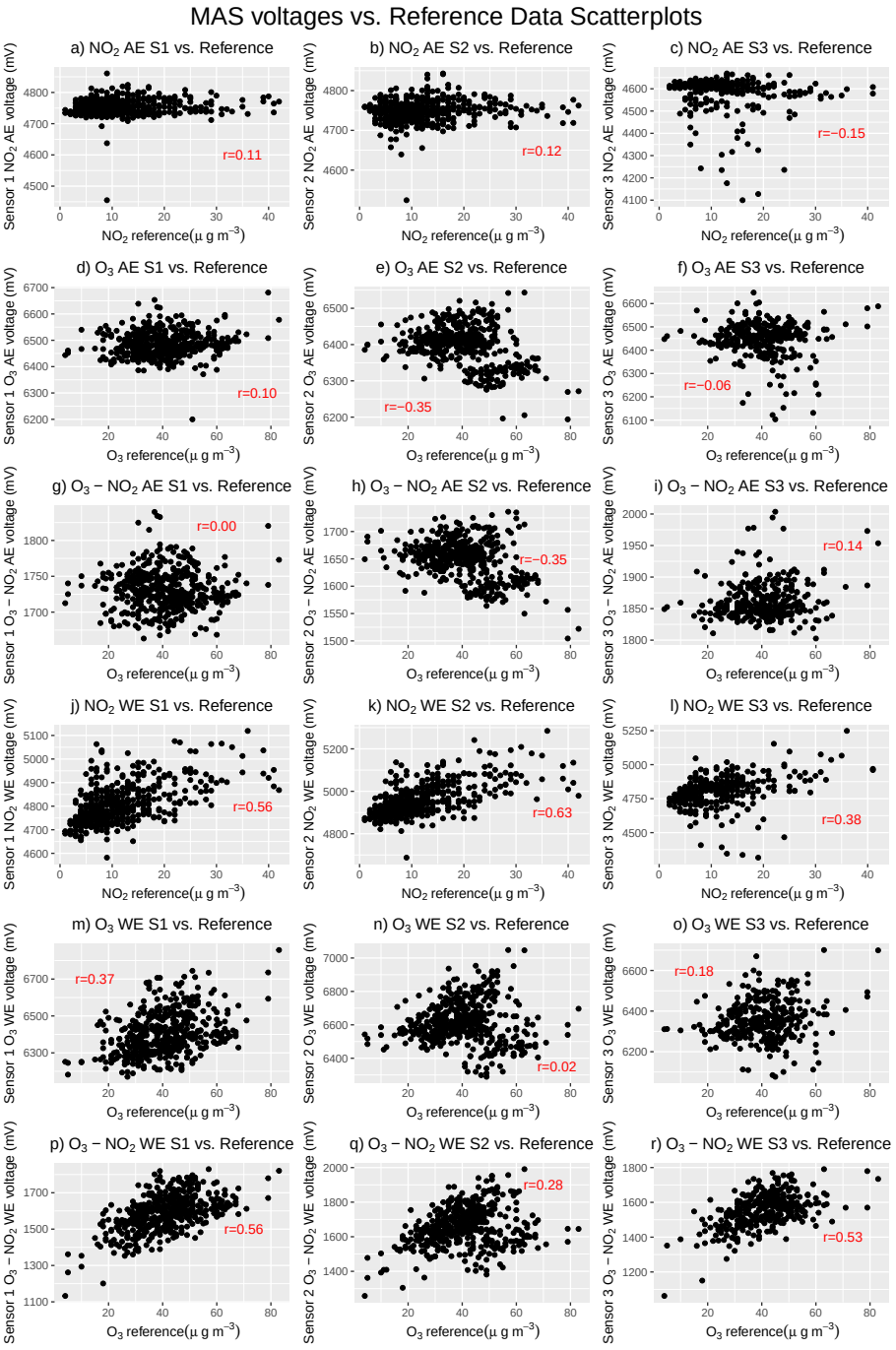

FIGURE 3.7: Scatterplots comparing the NO_2 and O_3 concentrations (in μg m^{-3}) recorded by the AURN monitor at hourly intervals with the MAS AE and WE hourly measurements (in mV). The correlations for each pairing are also provided in red.

3.2 Bland-Altman analysis to compare MAS to each other

This section presents comparisons of the MAS outputs with each other using Bland-Altman plots. A scatterplot of the hourly measurements will also be presented with the correlation coefficient from Table 3.1 referred to again. The section is split into three parts: Subsection 3.2.1 introduces Bland-Altman plots as a method for comparing the agreement between two data sets. Subsection 3.2.2 applies Bland-Altman analysis to the AE voltages for both pollutants, whereas Subsection 3.2.3 applies Bland-Altman analysis to the WE voltages again for both pollutants.

3.2.1 Bland-Altman analysis

In 1986, Bland and Altman argued that using correlation coefficients to compare two measurements on the same variable from different methods is misleading. They suggested "an alternative approach, based on graphical techniques and simple calculations" [27]. Bland and Altman suggest a plot where the mean values for the two experiments are plotted against the difference in the measurements. For two sets of measurements $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, and $i = 1, \ldots, n$,

$$\text{Mean}_i = \frac{x_i + y_i}{2}\,, \qquad \text{and} \qquad \text{Difference}_i = x_i - y_i\,, \tag{3.1}$$

are estimated and plotted against each other. On the Bland-Altman plot, the limits of agreement are also plotted. For the limits of agreement the 2.5^{th} and 97.5^{th} percentiles of the differences were used to create an interval, without relying on any distributional assumptions, within which 95% of the observed differences lie.

3.2.2 AE voltages

NO_2

From Figure 3.8 a), it is clear that the hourly measurements from Sensors 1 and 2 are in agreement with each other as zero lies between the 2.5^{th} and 97.5^{th} percentiles. The majority of the points form a cloud close to the zero difference line. Furthermore, both the mean and median values are quite close to zero. The scatterplot in Figure 3.8 b) highlights that the NO_2 AE hourly measurements from Sensors and 1 and 2 are quite similar as the majority of the points lie on or close to the equivalence line. However, the correlation coefficient would only indicate a weak linear relationship.

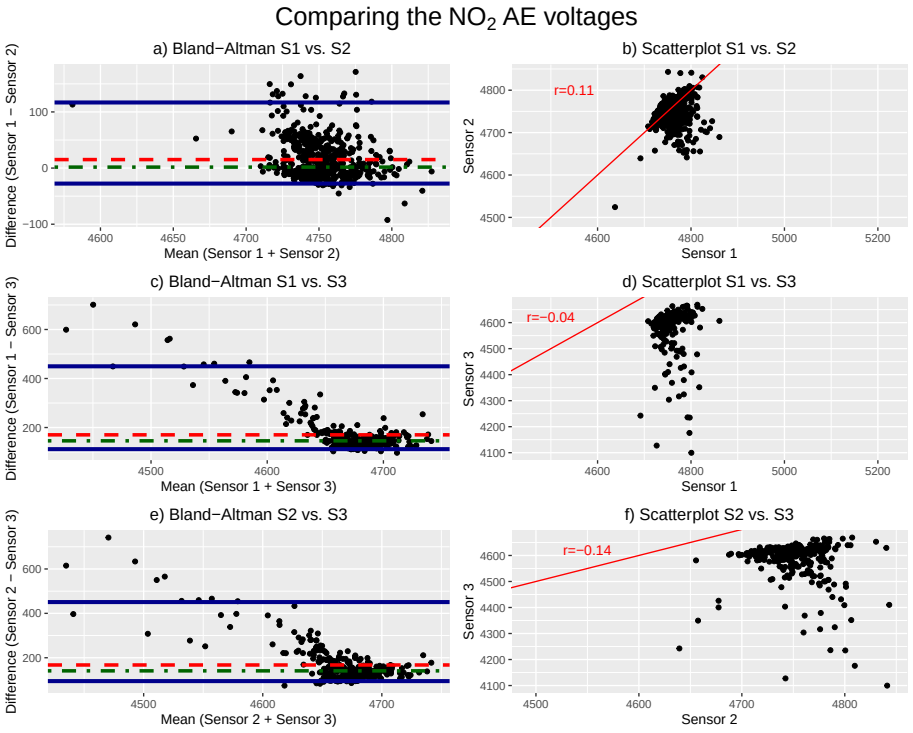

FIGURE 3.8: On the left, there are Bland-Altman plots to assess the existence of agreement between hourly measurements of the NO_2 AE voltage hourly measurements (mV) from the MAS. The dashed/dotted green line is the mean difference, the dashed red line is the median difference and the blue solid lines are the 2.5^{th} and 97.5^{th} percentiles. On the right, there are scatterplots between the NO_2 AE voltages with the equivalence line in red and the correlation between the hourly measurements. The correlations for each pairing are also provided.

The plots in Figure 3.8 c) and e) look analogous to each other and they indicate that the NO_2 AE hourly measurements from Sensor 3 are not in agreement with the other two sensors. In both cases, the values between the 2.5^{th} and 97.5^{th} percentiles are positive. The mean and median values are close to each other, in both cases, slightly lower than 200 mV. The scatterplots in Figure 3.8 d) and f) are also similar to each other as on both the points lie below the equivalence line. Both correlations indicate a weak negative linear relationship. Overall, the Bland-Altman analysis confirms the expectations from the exploratory analysis from Figures 3.3 a) and 3.4 a).

O_3

Figure 3.9 a) shows that the O_3 AE hourly measurements from Sensors 1 and 2 are not in agreement with each other as the values between the 2.5^{th} and 97.5^{th} percentiles are entirely positive. The mean and median values are close to each other around 100 mV. Furthermore, the correlation coefficient ($r = 0.06$) suggests that there is a very weak positive linear relationship and all but one point are below the equivalence line, which suggests that Sensor 1 has taken larger measurements than Sensor 2.

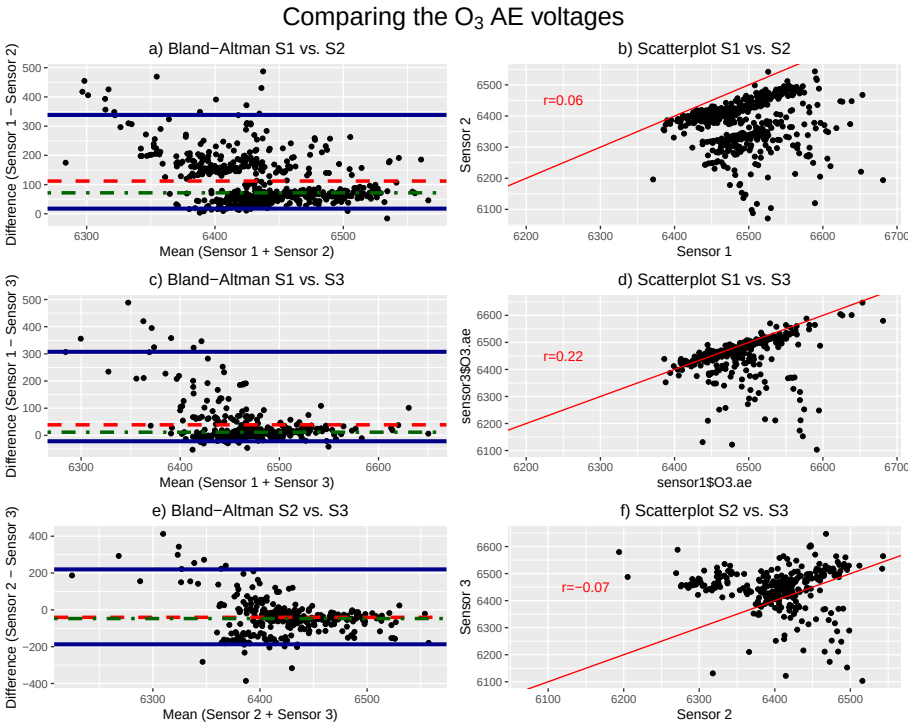

FIGURE 3.9: On the left, there are Bland-Altman plots to assess the existence of agreement between hourly measurements of the O_3 AE voltage hourly measurements (mV) from the MAS. The solid dashed/dotted line is the mean difference, the dashed red line is the median difference and the blue solid lines are the 2.5^{th} and 97.5^{th} percentiles. On the right, there are scatterplots between the O_3 AE voltages with the equivalence line in red and the correlation between the hourly measurements. The correlations for each pairing are also provided.

Sensor 3 appears to be agreement with the other two sensors as seen in the Bland-Altman plots in Figure 3.9 c) and e), where zero lies between the 2.5^{th} and 97.5^{th} percentiles. The mean and median values are close to each other. On the scatterplots in Figure 3.9 d) and f), some of the points are lying on the equivalence line. Sensors 1 and 3 appear to be weakly positive linearly correlated with r = 0.22, whereas Sensors 2 and 3 are weakly negatively linearly correlated with r = −0.07. Overall, the Bland-Altman analysis confirms the conclusions from the exploratory analysis from Figures 3.3 b) and 3.4 b).

O_3 - NO_2

None of the sensors are in agreement with each other as seen from the Bland-Altman plots in Figure 3.10 a), c) and e), where the 2.5^{th} and 97.5^{th} percentiles are either entirely positive or negative. The mean and median values for all measurements are close to each other. In the scatterplots in Figure 3.10 b), d) and f), the points either lie below or above the equivalence line. It is interesting to note that there is a strong positive linear relationship between the measurements for Sensors 1 and 3 with r = 0.81. Overall, the

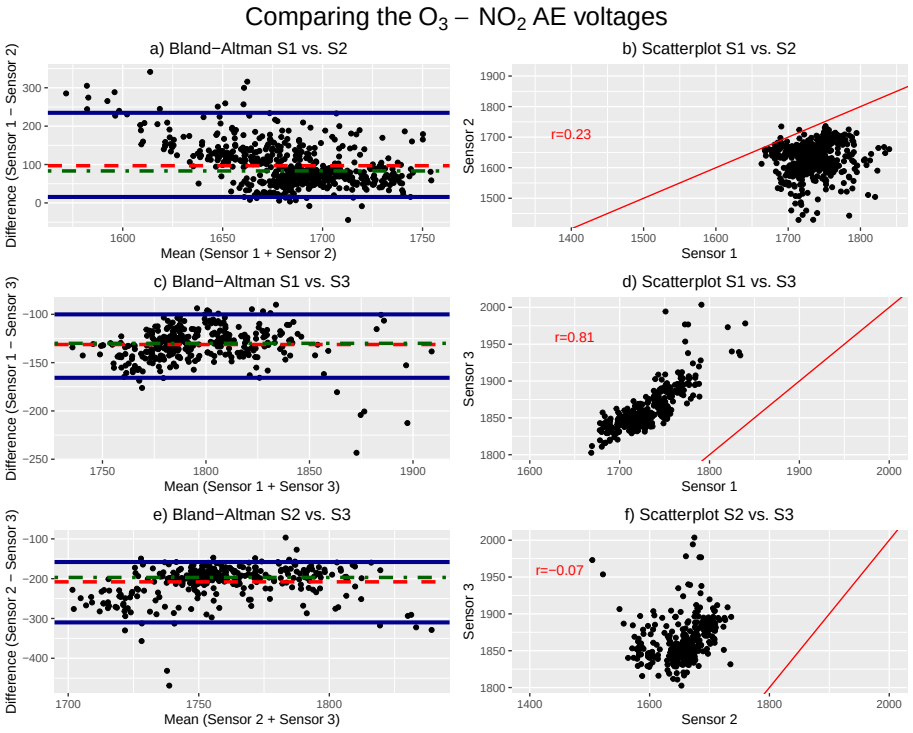

FIGURE 3.10: On the left, there are Bland-Altman plots to assess the existence of agreement between hourly measurements of the O_3 - NO_2 AE voltage hourly measurements (mV) from the MAS. The dashed/dotted green line is the mean difference, the dashed red line is the median difference and the blue solid lines are the 2.5^{th} and 97.5^{th} percentiles. On the right, there are scatterplots between the O_3 - NO_2 AE voltages with the equivalence line in red and the correlation between the hourly measurements. The correlations for each pairing are also provided.

results from Bland-Altman analysis are in agreement with the expectations based on the exploratory analysis in Figures 3.3 c) and 3.4 c).

3.2.3 WE voltages

NO_2

The NO_2 WE hourly measurements from Sensor 2 are not in agreement with the other two MAS as seen from the Bland-Altman plots in Figure 3.11 a) and e). The values between the 2.5^{th} and 97.5^{th} percentiles for the differences between Sensors 1 and 2 are entirely negative, whereas the values between the 2.5^{th} and 97.5^{th} percentiles for the differences between Sensors 2 and 3 are entirely positive. In both cases, the mean and median values are close to each other. The scatterplots in Figure 3.11 b) and f) show that the points respectively lie above or below the equivalence line. The correlation coefficients ($r = 0.89$ and $r = 0.51$) show that there is respectively a strong and moderate positive linear relationship between the measurements.

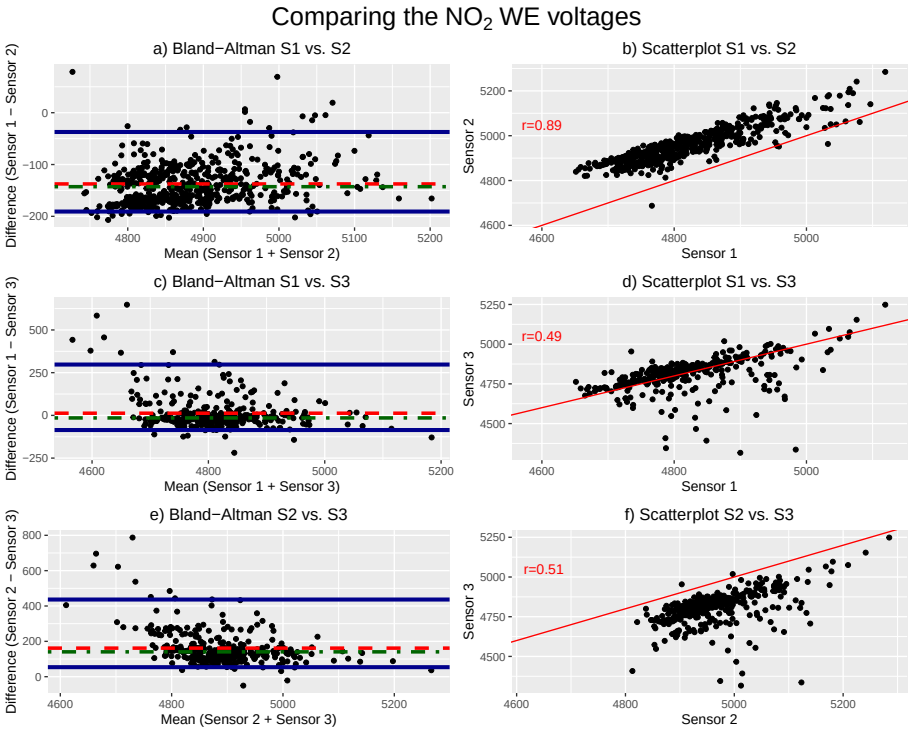

FIGURE 3.11: On the left, there are Bland-Altman plots to assess the existence of agreement between hourly measurements of the NO_2 WE voltage hourly measurements (mV) from the MAS. The dashed/dotted green line is the mean difference, the dashed red line is the median difference and the blue solid lines are the 2.5^{th} and 97.5^{th} percentiles. On the right, there are scatterplots between the NO_2 WE voltages with the equivalence line in red and the correlation between the hourly measurements. The correlations for each pairing are also provided.

However, the NO_2 WE hourly measurements between Sensors 1 and 3 are in agreement with each other as zero lies between the 2.5^{th} and 97.5^{th} percentiles as seen in the Bland-Altman plot in Figure 3.11 c). Furthermore, the mean and median values in Figure 3.11 c) are both very close to zero. However, the correlation coefficient for the NO_2 measurements from Sensors 1 and 3 is the lowest of the three NO_2 WE correlations with $r = 0.49$. This only indicates a moderate positive linear relationship between the measurements (Figure 3.11 d)). The Bland-Altman analysis confirms the expectations based on the exploratory analysis in Figures 3.3 d) and 3.4 d).

O_3

From the Bland-Altman plots in Figure 3.12 a) and c), it follows that the O_3 WE hourly measurements from Sensor 1 are in agreement with the other two MAS. For both plots, zero lies between the 2.5^{th} and 97.5^{th} percentiles. The mean and median values for the differences between Sensors 1 and 2 are quite far from zero, whereas the mean and median values for the differences between Sensors 1 and 3 are almost zero but in both cases the mean and median values appear to agree with each other. The scatterplots in Figure 3.12

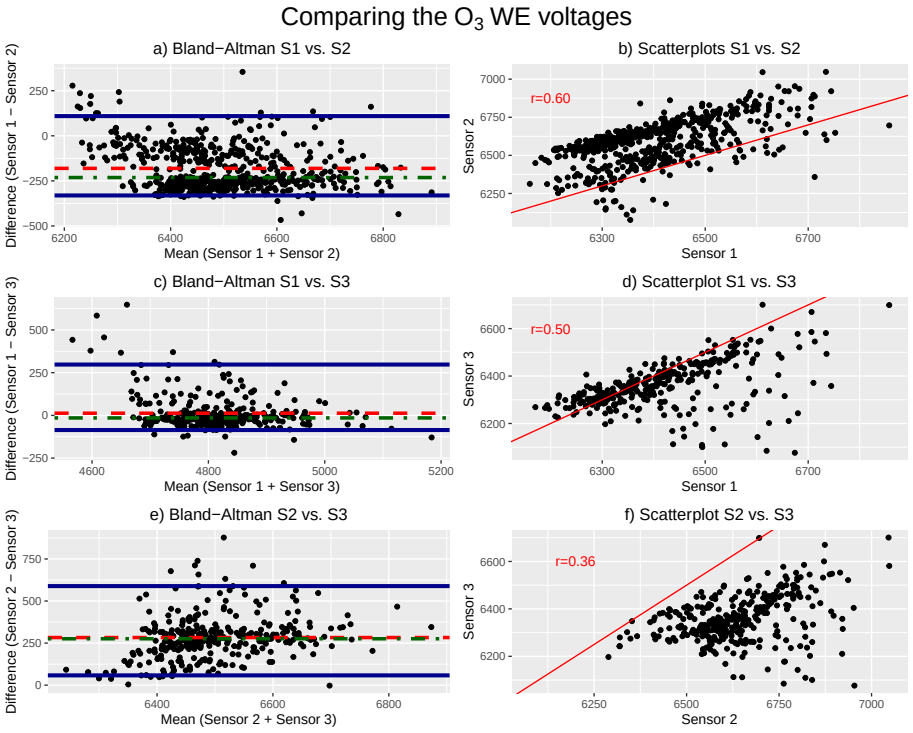

FIGURE 3.12: On the left, there are Bland-Altman plots to assess the existence of agreement between hourly measurements of the O_3 WE voltage hourly measurements (mV) from the MAS. The dashed/dotted green line is the mean difference, the dashed red line is the median difference and the blue solid lines are the 2.5^{th} and 97.5^{th} percentiles. On the right, there are scatterplots between the O_3 WE voltages with the equivalence line in red and the correlation between the hourly measurements. The correlations for each pairing are also provided.

b) and d) show that there is a moderately strong linear positive relationship between the MAS hourly measurements with r = 0.60 and r = 0.50, respectively. The agreement between the measurements from Sensors 1 and 2 is surprising given the exploratory analysis in Figures 3.3 e) and 3.4 e).

The Bland-Altman plot in Figure 3.12 e) shows that O_3 WE hourly measurements from Sensors 2 and 3 are not in agreement with each other as the values between the 2.5^{th} and 97.5^{th} percentiles are entirely positive. The mean and median values are almost identical with values slightly higher than 250 mV. On the scatterplot in Figure 3.12 f), the points lie below the equivalence line and the correlation coefficient is r = 0.36 indicating a low to moderate positive linear relationship.

O_3 - NO_2

All the Bland-Altman plots in Figure 3.13 a), c) and e) show that there is agreement between all the O_3 - NO_2 WE hourly measurements from all three of the sensors as zero lies between the 2.5^{th} and 97.5^{th} percentiles in all the plots. In all cases, the mean

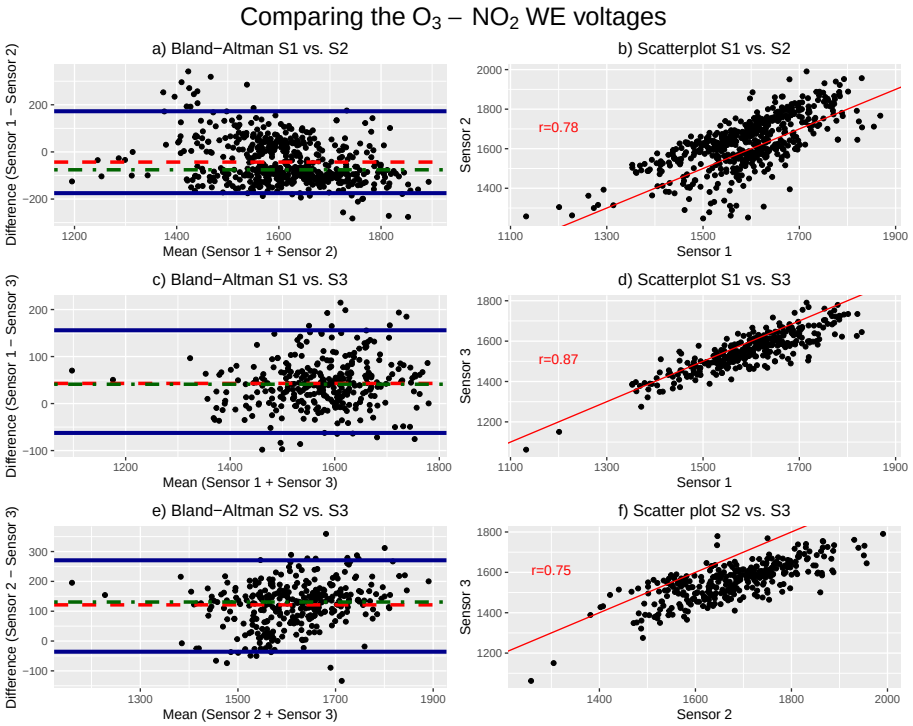

FIGURE 3.13: On the left, there are Bland-Altman plots to assess the existence of agreement between hourly measurements of the O_3 - NO_2 WE voltage hourly measurements (mV) from the MAS. The dashed/dotted green line is the mean difference, the dashed red line is the median difference and the blue solid lines are the 2.5^{th} and 97.5^{th} percentiles. On the right, there are scatterplots between the O_3 - NO_2 WE voltages with the equivalence line in red and the correlation between the hourly measurements. The correlations for each pairing are also provided.

and median values are close to each other. The correlation plots in Figure 3.13 b), d) and f) show that there appears to be strong linear positive correlation between the measurements with r varying from 0.75 to 0.87. These conclusions are expected given the exploratory analysis in Figures 3.3 f) and 3.4 f).

3.2.4 Findings

Overall, the conclusions from the Bland-Altman plots are in agreement with the exploratory analysis in Figures 3.3 and 3.4 with the exception of the O_3 WE MAS hourly measurements from Sensors 1 and 2. For the AE hourly measurements, only 3 out of the 9 Bland-Altman plots in Figures 3.8, 3.9 and 3.10 show consistency which indicates the AE voltages are not reliable. Therefore, the AE MAS hourly measurements will not be further modelled in this chapter. On the other hand, the WE MAS hourly measurements are more often than not consistent with each other (6 out of the 9 Bland-Altman plots in Figures 3.11, 3.12 and 3.13 show consistency). However, as this is not the case for all WE voltages from the sensors, this indicates that the sensors require further improvements in terms of providing reliable hourly measurements.

3.3 Relating the MAS to the reference monitor data

The consistency of the hourly measurements taken by the MAS is crucial but it is even more important to check the relationship between the MAS voltages and the true pollutant concentrations as measured by the reference monitor. To test the quality of the hourly measurements taken by the MAS in comparison to the reference monitor, linear regression models were fitted using the MAS voltage as a response and the respective reference pollution levels (NO_2 and O_3) as covariates. Due to the inconsistency in the AE voltages with each other, only the WE voltages of the MAS were modelled. Additionally, the models were fitted with the average voltage across the three sensors as response because averaging across the three MAS hourly measurements would provide less fluctuating air pollution measurements. However, the models did not perform well, and additional covariates, temperature and relative humidity, were also included in the regression. Due to the specifications of the AirSpecks package for the O_3 hourly measurements [10], there are two models for the O_3 MAS hourly measurements - one of the models will have as an additional covariate, to the four previously mentioned, the NO_2 WE hourly measurements. The reference levels for the NO_2 and O_3 pollutants were included in all the models as the basic assumption was that the MAS are measuring a mixture of the two pollutants. Additional models (with the averaged NO_2 and O_3 WE voltages from each sensor as well as the O_3 - NO_2 WE voltages for each sensors) were fitted based on the aforementioned dependencies between the MAS hourly measurements [10]. To avoid repetitiveness, the full case for the Sensor 1 measurements will be presented, whereas the other cases will contain a short table summarising the final models fitted as well as parameter estimates and their 95% CIs for the best model based on AIC and BIC values and the full model with all explanatory variables. When AIC and BIC disagree, the final model is chosen based on the BIC value as BIC favours models with a smaller number of covariates than AIC [79].

3.3.1 NO_2 WE voltage

Sensor 1

The first step in modelling the NO_2 WE hourly measurements taken by Sensor 1 was to fit a standard OLS model (see Subsection 2.2.1) with just one covariate - the reference monitor hourly measurements for NO_2. The model has an $R^2_{adj.} = 26.1\%$ which means that only 26.1% of the variability in the data is explained by the reference NO_2 measurements. The diagnostic plots for the model are presented in Figure 3.14.

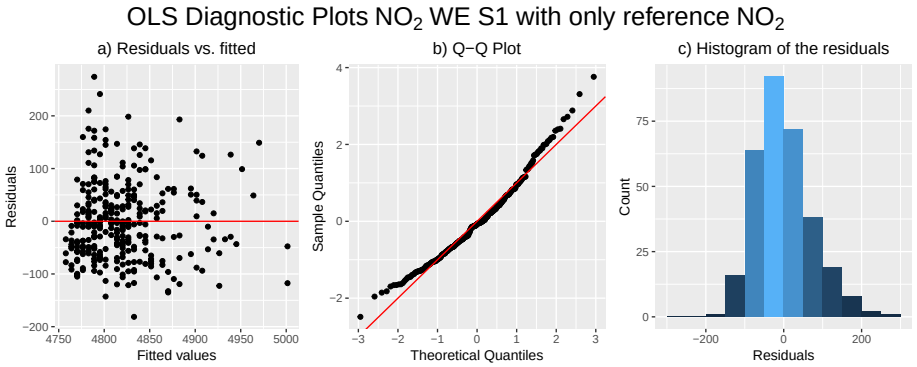

FIGURE 3.14: Diagnostic plots for the OLS fit for the model with NO_2 WE Sensor 1 hourly measurements (in mV) as a response and the reference NO_2 concentration (in μg m^{-3}) as a covariate.

The diagnostic plots in Figure 3.14 show that there are problems with the fit. On the residuals vs. fitted values plot (Figure 3.14 a)), the points are fanning out to the left, which indicates heteroscedascity problems. The points on the qq-plot (Figure 3.14 b)) are mostly following the normality line but there is curvature in the tails. The histograms of the residuals (Figure 3.14 c)) shows that the residuals are symmetric and normally distributed around zero. However, a Shapiro-Wilk test was performed and p-value of 0.03 was estimated indicating that there is some non-normality. Due to all these issues, more covariates (reference O_3, temperature and relative humidity) are added in different combinations to try and explain the variation in the data as well as fix the fit problems. The comparative summary of these models is provided in Table 3.2.

Model	$\mathbf{R^2_{adj.}}$	**DF**	**AIC**	**BIC**
$RefNO_2$	26.06%	3	3,578.66	3,589.89
Temperature	8.57%	3	3,645.13	3,656.37
Rel.Humidity	5.18%	3	3,656.53	3,667.77
$RefNO_2$+Temperature	37.31%	4	3,528.03	3,543.01
$RefNO_2$+Rel.Humidity	35.14%	4	3,538.64	3,553.62
$RefNO_2$+$RefO_3$	34.87%	4	3,539.94	3,554.93
Temperature+Rel.Humidity	8.33%	4	3,646.94	3,661.92
$RefNO_2$+Temperature+Rel.Humidity	37.29%	5	3,529.11	3,547.84
$RefNO_2$+Temperature+Rel.Humidity+$RefO_3$	40.86%	6	3,511.75	3,534.23

TABLE 3.2: Comparing the different linear models fitted with NO_2 WE voltage (in mV) from Sensor 1 as a response.

From Table 3.2, $R^2_{adj.}$, AIC and BIC agree that the best model is the one with all four covariates. The model fit suggests that the sensor has also been measuring the fluctuations in the meteorological conditions in addition to the pollutant concentrations. The diagnostic plots for the final model are examined in Figure 3.15.

The residuals vs. fitted values plot in Figure 3.15 a) does not indicate that any assumptions are broken - the points are randomly scattered around zero. There is no more

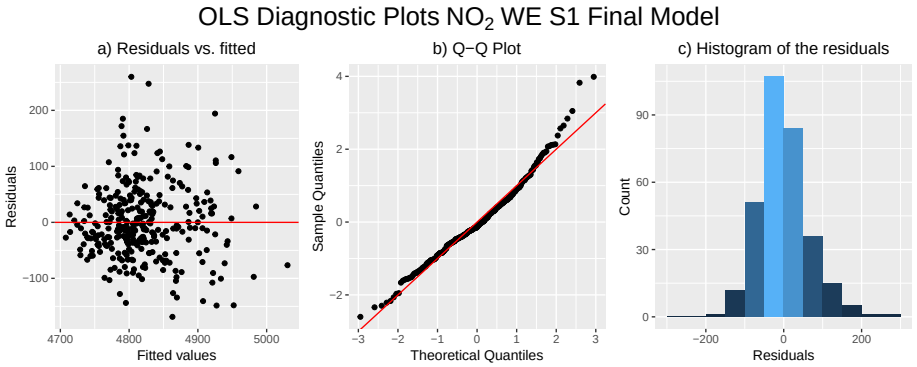

FIGURE 3.15: Diagnostic plots for the OLS fit for the model with NO_2 WE Sensor 1 hourly measurements (in mV) as a response and all four covariates (reference NO_2 (in μg m^{-3}), reference O_3 (in μg m^{-3}), temperature (in °C) and relative humidity (in %)).

fanning out which indicates the heteroscedasticity problem has been solved by adding more covariates. The qq-plot in Figure 3.15 b) shows an improvement compared to Figure 3.14 b) as the points on the bottom tail are now lying on the equivalence line. The histogram in Figure 3.15 c) shows that the residuals are symmetric and normally distributed. A Shapiro-Wilk test was again performed but the p-value was again 0.03 suggesting that the residuals are not normally distributed. However, as this is due to just a few outliers and there is an improvement from the single covariate model in Figure 3.14, this is reasonable in real world data and is not a reason for concern. Additionally, the ACF and PACF plots of the residuals are examined in Figure 3.16. The plots indicate that there is autocorrelation present in the residuals. This suggests that the linear model is not appropriate. For interpretability reasons, AR(1) and AR(2) correlation structures were applied and compared. Therefore, all models were fitted as GLS ones with the correlation structure and compared using the information criteria summarised in Table 3.3.

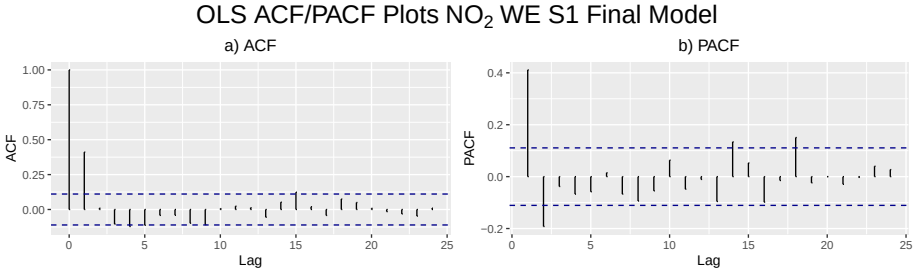

FIGURE 3.16: ACF (a) and PACF (b) of the residuals for the final linear model with Sensor 1 NO_2 WE voltage (in mV) as response and all four covariates.

According to Table 3.3, both the AIC and BIC values show that AR(2) is the preferred correlation structure although there is almost no difference between the two correlation structures. The AIC values show that the best model is the full model with all four covariates, whereas the BIC values suggest that the best model only has the two reference pollutant concentrations. Therefore, Table 3.4 summaries the coefficients and their 95%

Model	Corr. Structure	DF	AIC	BIC
$RefNO_2$	AR(1)	4	3,440.42	3,455.38
$RefNO_2$	AR(2)	5	3,435.50	3,454.20
Temperature	AR(1)	4	3,442.42	3,457.38
Temperature	AR(2)	5	3,437.26	3,455.96
Rel.Humidity	AR(1)	4	3,445.75	3,460.71
Rel.Humidity	AR(2)	5	3,440.20	3,458.90
$RefNO_2$+Temperature	AR(1)	5	3,437.28	3,455.96
$RefNO_2$+Temperature	AR(2)	6	3,432.46	3,454.87
$RefNO_2$+Rel.Humidity	AR(1)	5	3,441.25	3,459.93
$RefNO_2$+Rel.Humidity	AR(2)	6	3,436.37	3,458.80
$RefNO_2$+$RefO_3$	AR(1)	5	3,419.50	3,438.18
$RefNO_2$+$RefO_3$	AR(2)	6	3,415.54	3,437.96
Temperature+Rel.Humidity	AR(1)	5	3,441.28	3,459.97
Temperature+Rel.Humidity	AR(2)	6	3,436.42	3,458.84
$RefNO_2$+Temperature+Rel.Humidity	AR(1)	6	3,437.06	3,459.46
$RefNO_2$+Temperature+Rel.Humidity	AR(2)	7	3,432.53	3,458.66
$RefNO_2$+Temperature+Rel.Humidity+$RefO_3$	AR(1)	7	3,417.75	3,443.86
$RefNO_2$+Temperature+Rel.Humidity+$RefO_3$	AR(2)	8	3,414.06	3,443.90

TABLE 3.3: Comparing the different GLS correlation structures for the various models fitted with NO_2 WE voltage from Sensor 1 (in mV) as a response.

confidence intervals from the model with the two reference pollutants (referred to as the two covariates model) and the model with all explanatory variables (referred to as the full model) is examined. The intercept terms are not included as the main interest is the effect of the covariates on the MAS hourly measurements.

Model	Ref NO_2	Temp	Rel. Humidity	Ref O_3
Two covariates	5.32 (3.33, 7.31)	0	0	3.16 (1.82, 4.49)
Four covariates	5.09 (2.95, 7.22)	-0.43 (-8.59, 7.73)	0.32 (-1.70, 2.35)	3.40 (1.92, 4.88)

TABLE 3.4: Summary of the parameter estimates and their 95% confidence intervals for the two final models with NO_2 WE voltage (in mV) from Sensor 1 as a response.

From Table 3.4, it is clear that both the reference level pollutions are significant for both models as the 95% CIs are entirely positive. This suggests that as the pollution levels increase so does the NO_2 WE voltage. In the four covariates model, temperature and relative humidity are not significant as their 95% CIs contain zero and are almost symmetric around it. This implies that temperature and relative humidity do not have a significant effect on the NO_2 WE hourly measurements from Sensor 1. Hence, the diagnostic plots for the two covariates model are examined below in Figure 3.17.

The diagnostic plots in Figure 3.17 are very similar to those in Figure 3.15 as expected. Figure 3.17 a) shows that the points are randomly scattered and normally distributed around zero. The qq-plot in Figure 3.17 b) shows that there is still problems with the

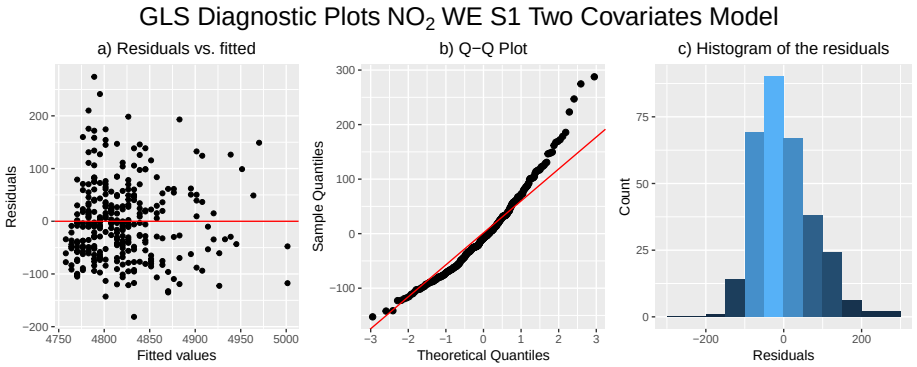

FIGURE 3.17: Diagnostic plots for the OLS fit for the model with NO_2 WE Sensor 1 hourly measurements (in mV) as a response and the reference NO_2 and reference O_3 pollutant concentrations (in μg m^{-3}) as covariates.

top tail. Additionally, the Shapiro-Wilk test indicated non-normality with a p-value of 0.03. However, the histogram in Figure 3.17 c) shows that the residuals are normally distributed around zero and the non-normality is the result of a few outliers which is reasonable when working with real world data. The ACF and PACF plots of the residuals are not re-examined as they do not reflect the change in the correlation structure of the residuals (see Subsection 2.2.1).

Sensor 2

Similarly to the modelling of the NO_2 WE voltage from Sensor 1, the OLS models with NO_2 WE voltage from Sensor 2 as a response with any variation of the covariates have autocorrelation present in their residuals. Hence, GLS models were fitted and two possible correlation structures were compared - AR(1) and AR(2). The comparison of different models is omitted for brevity. Both AIC and BIC values suggested that the AR(1) correlation structure is preferred. However, as with Sensor 1, AIC favours the full model, whereas BIC favours the model with the two reference concentrations.

Sensor 3

As with the NO_2 WE voltages from Sensors 1 and 2, the OLS models with NO_2 WE voltage from Sensor 3 as a response with any combination of the covariates have autocorrelation present in their residuals. Hence, GLS models were fitted and two possible correlation structures were compared - AR(1) and AR(2) but the comparison is omitted to avoid repetition. Both AIC and BIC values suggested that the AR(2) correlation structure is preferred. AIC once again favoured the full model with all four covariates, whereas BIC favoured a model with just two of the covariates - temperature and relative humidity.

Average NO_2 WE voltage

There is a lot of variability among the hourly measurements from the three sensors in Figure 3.3 and the Bland-Altman analysis in Figure 3.11 showed that the measurements from all the sensors are not consistent with each other except for Sensors 1 and 3. Hence, it is of interest to find out whether an averaged value across the sensors captures the fluctuations of the pollutants. However, an OLS fit of the models suffers from the same issue as the models for each of the sensors individually - there is autocorrelation present in the residuals. Model comparison for different sets of covariates and correlation structures is omitted for brevity. Both AIC and BIC agreed that the best model has four covariates (the two reference pollutant levels, temperature and relative humidity) and an AR(2) is the preferred correlation structure.

3.3.2 O_3 WE voltage

Overall, the O_3 WE voltage modelling was quite similar to the NO_2 WE one. There is autocorrelation present in the residuals when an OLS fit is applied. Therefore, the models with a GLS fit with AR(1) and AR(2) correlation structures are applied. When modelling O_3 WE voltage, besides the standard combination of covariates, a fifth covariate is added - NO_2 WE voltage from the corresponding sensor. This accounts for the aforementioned relationship between the O_3 WE and NO_2 WE voltages [10]. Similarly to the NO_2 WE voltages modelling, the full case for the first sensor is presented in full, whereas summaries are provided for hourly measurements from Sensors 2 and 3, and the averaged voltage across all three sensors.

Sensor 1

The OLS model with O_3 WE voltage from Sensor 1 as a response and the reference O_3 as an exploratory variable was fitted first. The reference O_3 is significant and the model has an $R^2_{\text{adj.}}$ of 19.04%, which indicates that the model only explains 19% of the variability in the data. Furthermore, the diagnostic plots in Figure 3.18 also highlight that the model is not a good fit to the data.

The residuals vs. fitted values plot in Figure 3.18 a) shows that the points are evenly spread around zero and randomly scattered. However, on the qq-plot in Figure 3.18 b), the points lie mostly off the normality line which also indicates problems with the fit. A Shapiro-Wilk test was performed and a p-value = 1.556e-05 was estimated indicating non-normality of the residuals. The issues are further highlighted by the histogram in Figure 3.18 c), which shows that the residuals are right-skewed rather than symmetric. In

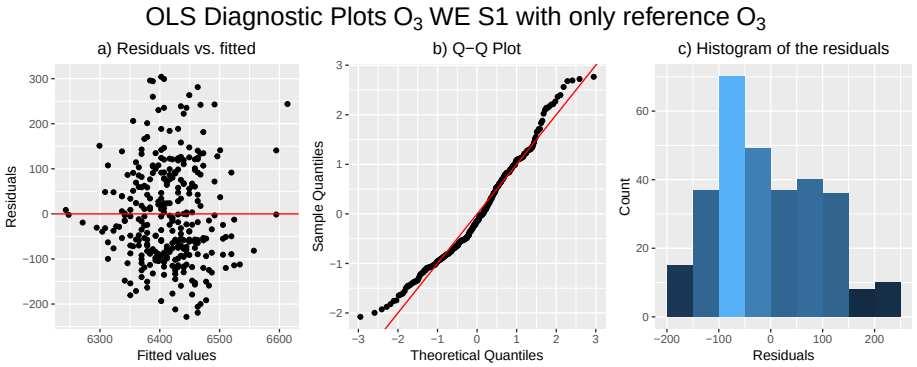

FIGURE 3.18: Diagnostic plots for the OLS fit for the model with O_3 WE Sensor 1 hourly measurements (in mV) as a response and the reference O_3 (in μg m^{-3}) as a covariate.

order to solve the issues from the plots and increase the percentage variability explained by the models, more covariates were added. A summary table with the $R^2_{adj.}$, AIC and BIC values of all the models is presented in Table 3.5.

Model	$\mathbf{R^2_{adj.}}$	**DF**	**AIC**	**BIC**
$RefO_3$	19.04%	3	3,835.55	3,846.79
Temperature	36.99%	3	3,757.08	3,768.31
Rel.Humidity	51.48%	3	3,675.26	3,686.50
NO_2 WE S1	28.79%	3	3,794.37	3,805.61
$RefO_3$ + NO_2 S1 WE	51.20%	4	3,677.05	3,692.03
$RefO_3$+Temperature	41.72%	4	3,732.67	3,747.65
$RefO_3$+Rel.Humidity	58.22%	4	3,628.48	3,643.46
$RefO_3$+$RefNO_2$	26.49%	4	3,805.32	3,820.30
Temperature+Rel.Humidity	51.22%	4	3,676.95	3,691.94
$RefO_3$+Temperature+Rel.Humidity	58.47%	5	3,627.59	3,646.32
NO_2 WE S1 +Temperature+Rel.Humidity	65.96%	5	3,565.30	3,584.03
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$	67.47%	6	3,552.09	3,574.57
$RefO_3$+NO_2 WE S1+Temperature+Rel.Humidity+$RefNO_2$	79.04%	7	3,415.58	3,441.80

TABLE 3.5: Comparing the different linear models fitted with O_3 WE voltage (in mV) from Sensor 1 as a response.

All comparison methods' results from Table 3.5 agree that the best model has all five covariates as it explains 79% of the variability in the data. Furthermore, all the results from the models seem to suggest that using the NO_2 WE voltage from Sensor 1 helps explain the variability in the data. Therefore, the diagnostic plots for that model are examined in Figure 3.19 and the ACF and PACF plots for the residuals are presented in Figure 3.20.

The diagnostic plots in Figure 3.19 show that the model is actually a good fit to the data. The residuals vs. fitted values plot in Figure 3.19 a) are randomly scattered around zero and the points on the qq-plot, Figure 3.19 b), lie mostly on the equivalence line. The histogram of the residuals, Figure 3.19 c), shows that they are symmetric and normally

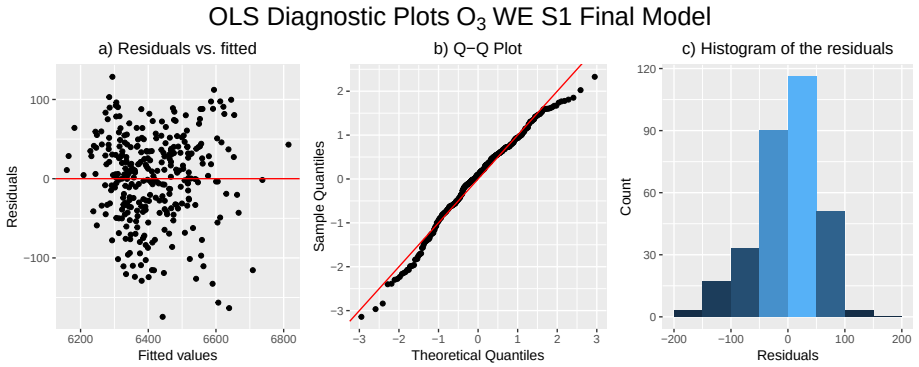

FIGURE 3.19: Diagnostic plots for the OLS fit for the model with O_3 WE Sensor 1 voltage (in mV) as a response and the five covariates (the reference O_3 (in μg m^{-3}), reference NO_2 (in μg m^{-3}), temperature (in °C), relative humidity (in %) and NO_2 WE Sensor 1 voltage).

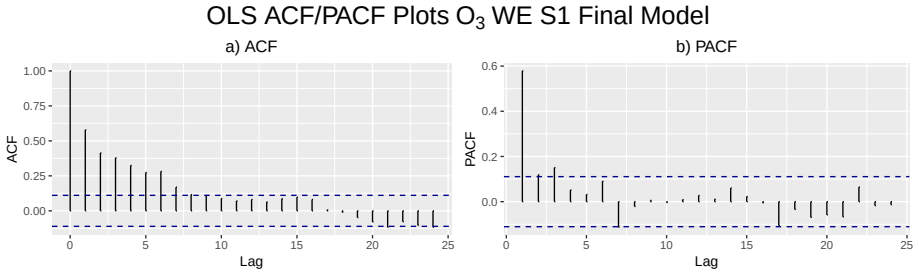

FIGURE 3.20: ACF (a) and PACF (b) of the residuals for the final linear model with Sensor 1 O_3 WE voltage (in mV) as response and all five covariates.

distributed around zero. Furthermore, a Shapiro-Wilk test was performed and p-value of 0.1 was estimated suggesting that there is no evidence of non-normality of the residuals. However, the ACF and PACF plots in Figure 3.20 a) and b) show that there is strong autocorrelation present, which means that the OLS fit is not appropriate. Therefore, all models were refitted with a GLS fit with AR(1) and AR(2) correlation structures, which were chosen for interpretability reasons. A comparison table for these models is presented in Table 3.6.

From Table 3.6, it is clear that for the models without NO_2 WE as a covariate, the best model is the one with four covariates (reference O_3 and NO_2, temperature and relative humidity) with an AR(1) structure as both AIC and BIC values show. However, for the model with all five covariates both AIC and BIC favour the AR(2) correlation structure. The estimated parameters for the models with four and five covariates are compared in Table 3.7. The intercept terms are not included as the main interest is in the effects the covariates have on the response.

From Table 3.7 it follows that adding NO_2 WE hourly measurements from Sensor 1 to the model causes the reference NO_2 to lose its significance but all other covariates remain significant for the model as their 95% CIs do not contain zero. It is interesting to note

Model	Corr. Structure	DF	AIC	BIC
$RefO_3$	AR(1)	4	3,551.27	3,566.23
$RefO_3$	AR(2)	5	3,553.16	3,571.86
Temperature	AR(1)	4	3,593.09	3,608.05
Temperature	AR(2)	5	3,594.36	3,613.06
Rel.Humidity	AR(1)	4	3,545.66	3,560.62
Rel.Humidity	AR(2)	5	3,547.59	3,566.29
NO_2WE S1	AR(1)	4	3,502.61	3,517.57
NO_2WE S1	AR(2)	5	3,492.21	3,510.91
$RefO_3$+NO_2WE S1	AR(1)	5	3,426.55	3,445.23
$RefO_3$+NO_2WE S1	AR(2)	6	3,428.54	3,450.96
$RefO_3$+Temperature	AR(1)	5	3,535.85	3,554.54
$RefO_3$+Temperature	AR(2)	6	3,536.02	3,558.44
$RefO_3$+Rel.Humidity	AR(1)	5	3,496.34	3,515.02
$RefO_3$+Rel.Humidity	AR(2)	6	3,496.17	3,518.59
$RefO_3$+$RefNO_2$	AR(1)	5	3,540.08	3,558.76
$RefO_3$+$RefNO_2$	AR(2)	6	3,541.49	3,563.89
Temperature+Rel.Humidity	AR(1)	5	3,533.14	3,551.82
Temperature+Rel.Humidity	AR(2)	6	3,535.12	3,557.54
$RefO_3$+Temperature+Rel.Humidity	AR(1)	6	3,473.68	3,496.08
$RefO_3$+Temperature+Rel.Humidity	AR(2)	7	3,473.46	3,499.59
NO_2WE S1+Temperature+Rel.Humidity	AR(1)	6	3,339.11	3,361.51
NO_2WE S1+Temperature+Rel.Humidity	AR(2)	7	3,341.10	3,367.24
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$	AR(1)	7	3,457.04	3,483.15
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$	AR(2)	8	3,457.84	3,487.68
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$+NO_2WE S1	AR(1)	8	3,270.57	3,300.39
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$+NO_2WE S1	AR(2)	9	3,266.81	3,300.35

TABLE 3.6: Comparing the different GLS models fitted with O_3 WE voltage (in mV) from Sensor 1 as a response.

Model	Ref O_3	Temp	Rel. Humidity	Ref NO_2	NO_2 WE S1
Four Covariates	6.93 (5.46, 8.40)	-21.37 (-29.30, -13.43)	-9.72 (-11.70, -7.75)	5.05 (2.87, 7.22)	0
Five Covariates	4.64 (3.47, 5.81)	-24.30 (-30.38, -18.21)	-10.55 (-12.12, -8.99)	0.37 (-1.37, 2.10)	0.72 (0.63, 0.80)

TABLE 3.7: Summary of the parameter estimates and their 95% confidence intervals for the two final models with O_3 WE voltage (in mV) from Sensor 1 as a response.

that the parameter estimates and their respective 95% CIs between the two models are different to each other, although they remain either entirely positive or negative. Hence, it could be concluded that in both models, the effect of the covariates are similar. As both AIC and BIC agree, the best model is the one with five covariates and therefore, the diagnostic plots for it are examined in Figure 3.21.

The diagnostic plots in Figure 3.21 show that the model is a good fit. The residuals vs. fitted values in Figure 3.21 a) are randomly scattered and evenly distributed around zero. The points on the qq-plot in Figure 3.21 b) lie in a straight line. A Shapiro-Wilk test was performed with an estimated p-value of 0.06 indicating that there is no evidence

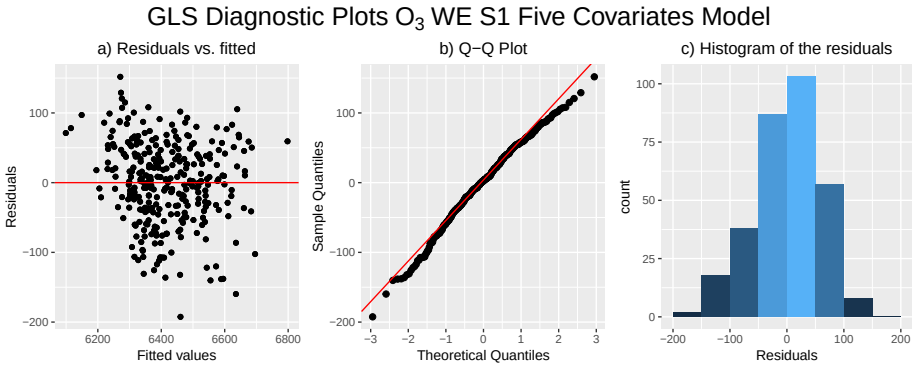

FIGURE 3.21: Diagnostic plots for the GLS fit with AR(1) correlation structure for the model with O_3 WE Sensor 1 voltage (in mV) as a response and all five covariates.

that the residuals are not normally distributed. Lastly, the histogram of the residuals in Figure 3.21 c) are symmetric around zero. The ACF and PACF plots of the residuals are not presented as they cannot show the change of adding a correlation structure (see Subsection 2.2.1).

Sensor 2

The OLS models with O_3 WE voltage from Sensor 2 as a response have the same problem as the models with O_3 WE voltage from Sensor 1 as a response - there is autocorrelation present in the residuals. Therefore, the models were refitted with AR(1) and AR(2) correlation structures but their comparison is omitted for brevity. The best model was the one with all five covariates and with an AR(2) correlation structure as both AIC and BIC values confirmed this. However, when the models without NO_2 WE hourly measurements are examined, it appears that the best model is the one with four covariates (reference O_3 and NO_2, temperature and relative humidity) and an AR(2) correlation structure. It is expected that similar models are chosen for the O_3 WE hourly measurements from Sensors 1 and 2 given that the Bland-Altman analysis (Figure 3.12) showed the measurements from the two sensors are consistent with each other.

Sensor 3

The OLS models with O_3 WE voltage from Sensor 3 have autocorrelation present in the residuals. Therefore, two correlation structures, AR(1) and AR(2), were compared using AIC and BIC values but to avoid repetition the comparison is omitted. It appeared that the models with NO_2 WE voltage from Sensor 3 in them as a covariate are performing better than the others. For most models, AIC and BIC favour the AR(1) correlation structure but the best model appears to be the one with all five covariates in it with

an AR(2) correlation structure. However, when the models without NO_2 WE voltage are examined, according to the AIC, the best model appears to be the one with four covariates (reference O_3 and NO_2 as well as temperature and relative humidity) with an AR(1) correlation structure, whereas according to BIC, the best model has reference O_3, temperature and relative humidity with an AR(1) correlation structure (referred to as the three covariates model).

Average O_3 WE voltage

The OLS models with the averaged O_3 WE as a response contain autocorrelation in the residuals so they were refitted as GLS fits with two different correlation structures (AR(1) and AR(2)), which are omitted for brevity. Both AIC and BIC favoured the AR(1) correlation structure except for the largest model with five covariates, where according to AIC the more complex AR(2) correlation structure is favoured. Overall, the best model is the one with all five covariates. However, when the models with the NO_2 WE hourly measurements averaged across the three sensors are ignored, the best model is the one with all four covariates (reference O_3 and NO_2, temperature and relative humidity) with both AIC and BIC favouring the AR(1) correlation structure.

3.3.3 O_3 - NO_2 WE voltage

An alternative to modelling the O_3 WE voltage using the NO_2 WE voltage as a covariate is to model the difference between the two voltages. Therefore, for each of the sensors, the difference between the O_3 and NO_2 WE voltages will be modelled using the reference NO_2 and O_3 pollution concentrations as well as the temperature and relative humidity measurements. As with the NO_2 WE and O_3 WE modelling, the full case for Sensor 1 will be presented, whereas only the summary of the final models is presented for the other two sensors and the averaged values to avoid repeatability.

Sensor 1

The OLS model with O_3 - NO_2 WE voltage from Sensor 1 as a response with the reference O_3 pollutant concentration as a covariate has a $R^2_{\text{adj.}} = 31.89\%$, indicating that almost 32% of the variability in the response has been explained by the covariate. The diagnostic plots in Figure 3.22 show that the model is a good fit to the data as the points on the residuals vs. fitted values plot (Figure 3.22 a)) are randomly scattered around zero, almost all the points on the qq-plot (Figure 3.22 b)) lie perfectly on the equivalence line and the histogram of the residuals (Figure 3.22 c)) shows the residuals

are symmetric around zero. Furthermore, a Shapiro-Wilk test was performed with a p-value of 0.97 which suggests the residuals are normally distributed. However, more covariates are added in order to increase the percentage of explained variability. The models are summarised in Table 3.8.

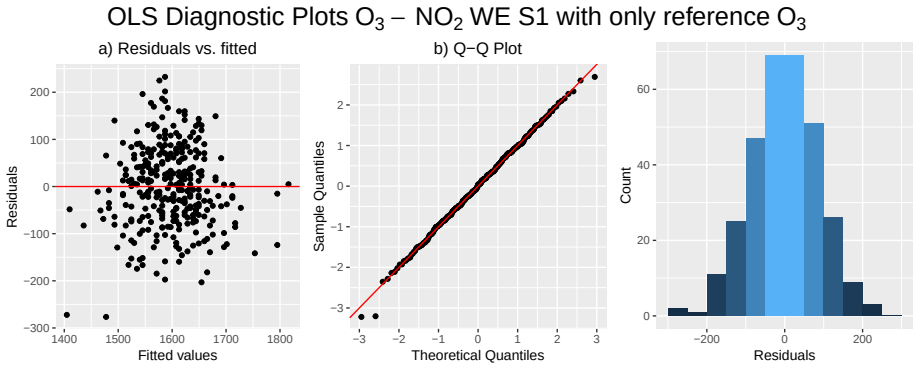

FIGURE 3.22: Diagnostic plots for the OLS fit for the model with O_3 - NO_2 WE Sensor 1 voltage (in mV) as a response and the reference O_3 (in μg m^{-3}) as a covariate.

Model	$R^2_{adj.}$	DF	AIC	BIC
$RefO_3$	31.89%	3	3,684.79	3,696.03
Temperature	21.92%	3	3,727.57	3,738.81
Rel.Humidity	41.90%	3	3,635.04	3,646.28
$RefO_3$+Temperature	39.21%	4	3,649.21	3,664.19
$RefO_3$+Rel.Humidity	58.90%	4	3,526.67	3,541.66
$RefO_3$+$RefNO_2$	33.73%	4	3,676.22	3,691.21
Temperature+Rel.Humidity	43.10%	4	3,628.52	3,643.50
$RefO_3$+Temperature+Rel.Humidity	65.82%	5	3,470.00	3,488.73
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$	66.44%	6	3,465.26	3,487.74

TABLE 3.8: Comparing the different linear models fitted with O_3 - NO_2 WE voltage (in mV) from Sensor 1 as a response.

All methods of comparison in Table 3.8 agree that the best model is the one with all four covariates. Therefore, the diagnostic plots for the model, in Figure 3.23, are examined. The residuals vs. fitted values plot (Figure 3.23 a)) shows the model is a good fit to the data as the points are randomly scattered around zero and the histogram of the residuals (Figure 3.23 c)) shows that the residuals are normally distributed around zero. However, the qq-plot (Figure 3.23 b)) suggests that there are problems with the fit as many of the points at both tails are off the normality line. Overall, the diagnostic plots in Figure 3.23 are not as good as those in Figure 3.22. Furthermore, there is autocorrelation present in the residuals as seen from the ACF and PACF plots in Figure 3.24 for both the only reference O_3 covariate model and the full model with all four covariates. In order to account for the autocorrelation, the models were refitted using a GLS fit with different correlation structures (for interpretability reasons once again AR(1) and AR(2)). The models are compared using AIC and BIC values in Table 3.9.

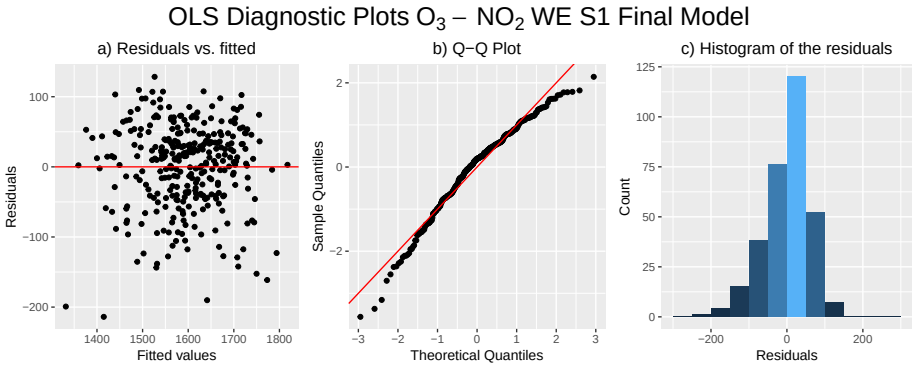

FIGURE 3.23: Diagnostic plots for the OLS fit for the model with O_3 - NO_2 WE from Sensor 1 voltage (in mV) as a response and all four covariates (reference NO_2 (in μg m^{-3}), reference O_3 (in μg m^{-3}), temperature (in °C) and relative humidity (in %)).

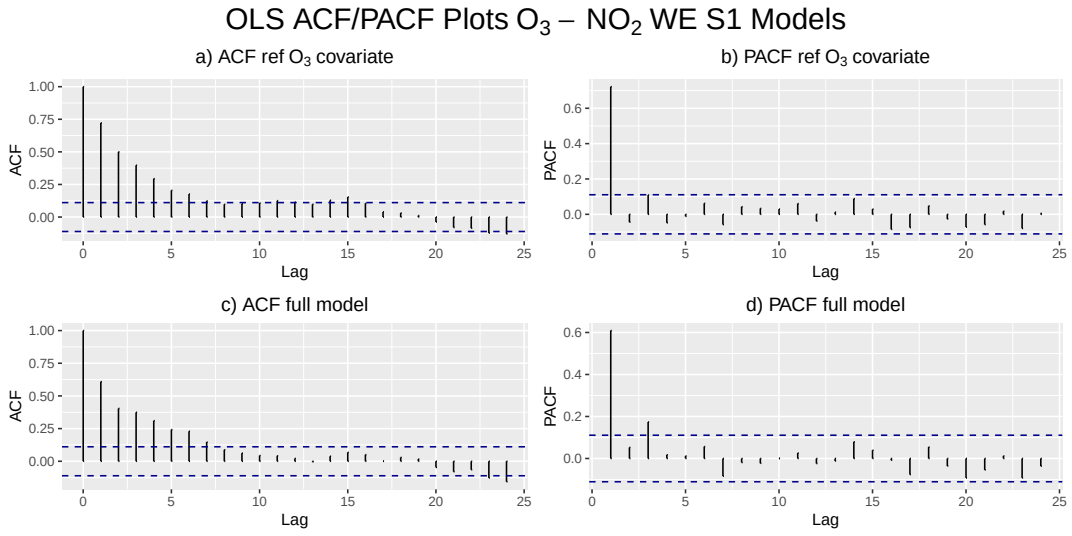

FIGURE 3.24: ACF and PACF plots of the residuals for the reference O_3 covariate model a) and b) and the full (all four covariates) linear model c) and d) with O_3 - NO_2 WE voltage (in mV) from Sensor 1 as response.

According to the AIC and BIC values from Table 3.9, AR(1) is the more commonly preferred correlation structure. However, AIC and BIC disagree which is the best model - AIC favours the full model with all four covariates, whereas BIC favours the model with only three covariates (reference O_3 pollution concentration, temperature and relative humidity). To further compare the two models, the estimates and their respective 95% CIs are compared in Table 3.10. The intercepts are not included as the effects of the covariates on the response are only of interest.

In Table 3.10, it is clear that the estimate for the reference NO_2 pollutant concentration is not significant in the model with four covariates, therefore, the model with three covariates (all of them are significant as the 95% CIs do not contain zero) is the better fit to the data. The estimate for the reference O_3 pollutant concentration is positive which means that as the true concentration of O_3 was rising, so was the O_3 - NO_2

Model	Corr. Structure	DF	AIC	BIC
$RefO_3$	AR(1)	4	3,448.14	3,463.10
$RefO_3$	AR(2)	5	3,449.27	3,467.96
Temperature	AR(1)	4	3,474.84	3,489.79
Temperature	AR(2)	5	3,469.95	3,488.65
Rel.Humidity	AR(1)	4	3,402.18	3,417.14
Rel.Humidity	AR(2)	5	3,402.01	3,420.71
$RefO_3$+Temperature	AR(1)	5	3,425.33	3,444.02
$RefO_3$+Temperature	AR(2)	6	3,427.24	3,449.66
$RefO_3$+Rel.Humidity	AR(1)	5	3,355.77	3,374.45
$RefO_3$+Rel.Humidity	AR(2)	6	3,357.73	3,380.15
$RefO_3$+$RefNO_2$	AR(1)	5	3,447.99	3,466.67
$RefO_3$+$RefNO_2$	AR(2)	6	3,449.12	3,471.54
Temperature+Rel.Humidity	AR(1)	5	3,347.43	3,366.11
Temperature+Rel.Humidity	AR(2)	6	3,348.63	3,371.05
$RefO_3$+Temperature+Rel.Humidity	AR(1)	6	3,298.80	3,321.20
$RefO_3$+Temperature+Rel.Humidity	AR(2)	7	3,299.80	3,325.93
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$	AR(1)	7	3,298.36	3,324.47
$RefO_3$+Temperature+Rel.Humidity+$RefNO_2$	AR(2)	8	3,298.96	3,328.80

TABLE 3.9: Comparing the different GLS models fitted with the O_3 - NO_2 WE voltage (in mV) from Sensor 1 as a response.

Model	Ref O_3	Temp	Rel. Humidity	Ref NO_2
Three covariates	3.93 (3.00, 4.86)	-26.51 (-33.21, -19.82)	-10.98 (-12.71, -9.25)	0
Four covariates	3.56 (2.31, 4.81)	-25.90 (-32.73, -19.07)	-10.91 (-12.64, -9.18)	-0.80 (-2.62, 1.02)

TABLE 3.10: Summary of the parameter estimates and their 95% confidence intervals for the two final models with O_3 - NO_2 WE voltage (in mV) from Sensor 1 as a response.

WE voltage from Sensor 1. The reference NO_2 pollutant concentration estimate is negative, although not significant, but this suggests that as the NO_2 concentration has increased, the difference between the O_3 and NO_2 WE voltage has decreased, which is to be expected. Relative humidity and temperature have negative estimates. Hence, as relative humidity and temperature increase, the difference between O_3 - NO_2 WE voltage from Sensor 1 decreases. Lastly, the diagnostic plots for the model with three covariates are examined.

The diagnostic plots for the final model in Figure 3.25 indicate the model is a good fit to the data. The residuals vs. fitted values plot in Figure 3.25 a) shows that the points are randomly scattered around zero. The qq-plot in Figure 3.25 b) is an improvement on the previous diagnostic plots for the linear model in Figure 3.23 b) but there are a few points on the bottom tail that are off the normality line. However, the histogram of the residuals, Figure 3.25 c), shows that they are symmetric around zero. The p-value from a Shapiro-Wilk test was estimated to be 0.07 indicating there is no evidence to suggest that the residuals are not normally distributed. The ACF/PACF plots of the residuals

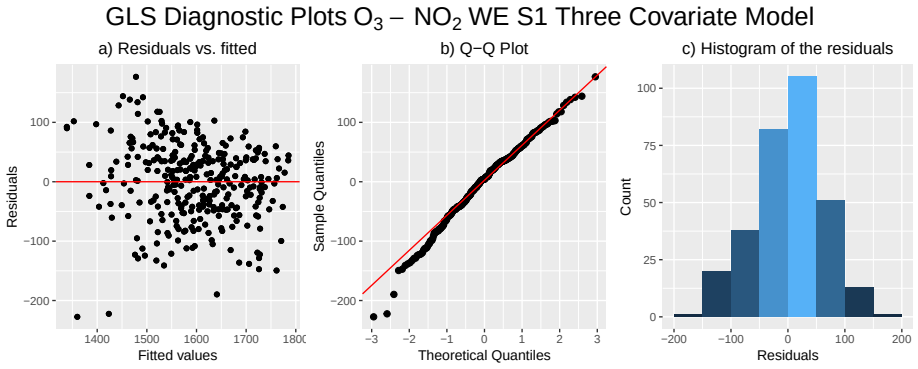

FIGURE 3.25: Diagnostic plots for the GLS fit for the model with O_3 - NO_2 WE from Sensor 1 voltage (in mV) as a response and reference O_3 (in μg m^{-3}), temperature (in °C) and relative humidity (in %) as covariates.

for the GLS fit are not presented as they do not reflect the change in the correlation structure (see Subsection 2.2.1).

Sensor 2

The OLS fit for the O_3 - NO_2 WE voltage from Sensor 2 contains autocorrelation in the residuals. The models, therefore, have to be refitted as a GLS fit with AR(1) and AR(2) correlation structures but the comparison is omitted for brevity. For the simpler models with less covariates, both AIC and BIC preffered the AR(1) correlation structure. However, for the models with three and four covariates, the AR(2) correlation structure is chosen. Once again, AIC favours the full model with all four covariates, whereas BIC prefers the model with three (reference O_3 pollution concentration, temperature and relative humidity) covariates.

Sensor 3

The ACF and PACF plots for all OLS models for the O_3 - NO_2 WE voltage from Sensor 3 as a response contain autocorrelation. Hence, the models were refitted using a GLS fit with both AR(1) and AR(2) correlation structures but the comparison is omitted to avoid repetition. Both AIC and BIC favoured the simple AR(1) correlation structure. Furthermore, both the information criteria agreed that the best model is the one with all four covariates.

Averaged O_3 - NO_2 WE voltage

The averaged difference between the O_3 and NO_2 WE voltages were averaged across the three sensors and OLS fits were used to model them. However, there was autocorrelation

present in the residuals of all models. Therefore, the models were refitted as GLS fits with two different correlation structures - AR(1) and AR(2) but the comparison is omitted for brevity. AIC seems to struggle picking between the two correlation structures, whereas BIC mostly favours AR(1). This is not surprising as the BIC favours simpler models. However, both information criteria picked an AR(1) model as the best. AIC favoured the full model with all four covariates, whereas BIC favoured the model with three covariates (reference O_3 pollutant concentration, temperature and relative humidity).

3.3.4 Findings

After models were fitted to the different covariates, the parameter estimates, their 95% CI, p- and t-values from the full models (GLS fit with the respective correlation structure) were examined and compared to look for common trends in Tables 3.11 and 3.12. This is done in order to establish and explain the contribution of all the different covariates to the response.

Pollutant	Sensor	Ref NO_2	Ref O_3	Temp.	Rel. Humidity	NO_2 WE
NO_2	S1	**5.09 (2.95, 7.22)**	**3.40 (1.92, 4.88)**	-0.43 (-8.59, 7.73)	0.32 (-1.70, 2.35)	
	S2	**3.90 (1.94, 5.85)**	**2.52 (1.16, 3.88)**	-4.97 (-12.47, 2.53)	-0.24 (-2.17, 1.68)	
	S3	2.55 (-0.14, 5.24)	1.36 (-0.61, 3.34)	-5.46 (-17.04, 6.13)	**6.22 (3.27, 9.16)**	
	Avg.	**2.98 (1.01, 4.96)**	**2.78 (1.35, 4.22)**	-4.67 (-13.00, 3.66)	**2.34 (0.24, 4.44)**	
O_3	S1	**5.05 (2.87, 7.22)**	**6.93 (5.46, 8.40)**	**-21.37 (-29.30, -13.43)**	**-9.72 (-11.70, -7.52)**	
	S2	**3.10 (0.89, 5.32)**	**5.72 (4.21, 7.24)**	**-36.98 (-45.16, -28.08)**	**-14.70 (-16.86, -12.54)**	
	S3	1.31 (-0.98, 3.60)	**3.27 (1.66, 4.89)**	**-17.37 (-26.38, -8.36)**	**-3.89 (-6.23, -1.56)**	
	Avg.	**3.01 (1.08, 4.93)**	**5.13 (3.80, 6.46)**	**-27.25 (-34.53, -19.97)**	**-9.80 (-11.66, -7.94)**	
O_3	S1	0.37 (-1.37, 2.10)	**4.64 (3.47, 5.81)**	**-24.30 (-30.38, -18.21)**	**-10.55 (-12.12, -8.99)**	**0.72 (0.63, 0.80)**
	S2	0.04 (-1.82, 1.90)	**3.84 (2.58, 5.11)**	**-35.74 (-42.34, -29.14)**	**-15.00 (-16.75, -13.25)**	**0.64 (0.54, 0.74)**
	S3	-0.98 (-2.63, 0.67)	**2.69 (1.58, 3.80)**	**-14.42 (-20.39, -8.45)**	**-6.48 (-8.06, -4.91)**	**0.54 (0.47, 0.60)**
	Avg.	0.16 (-1.33, 1.66)	**3.81 (2.80, 4.82)**	**-25.36 (-30.71, -20.01)**	**-10.78 (-12.17, -9.39)**	**0.60 (0.52, 0.68)**
O_3 - NO_2	S1	-0.80 (-2.62, 1.02)	**3.56 (2.31, 4.81)**	**-25.90 (-32.73, -19.07)**	**-10.91 (-12.64, -9.18)**	
	S2	-1.22 (-3.19, 0.74)	**2.88 (1.53, 4.23)**	**-35.57 (-42.86, -28.28)**	**-15.16 (-17.07, -13.25)**	
	S3	**-2.66 (-4.67, -0.66)**	**2.09 (0.72, 3.46)**	**-11.37 (-18.78, -3.95)**	**-7.28 (-9.14, -5.42)**	
	Avg.	-1.07 (-2.75, 0.62)	**2.79 (1.62, 3.96)**	**-26.06 (-32.50, -19.62)**	**-11.63 (-13.28, -9.98)**	

TABLE 3.11: Summary of the parameter estimates and their 95% confidence intervals for all models. All significant values are bolded.

Pollutant	Sensor	Ref NO_2	Reference O_3	Temp.	Rel. Humidity	NO_2 WE
NO_2	S1	4.69 **(0)**	4.52 **(0)**	-0.10 (0.92)	0.32 (0.75)	
	S2	3.92 **(0)**	3.65 **(0)**	-1.30 (0.19)	-0.25 (0.80)	
	S3	1.87 (0.06)	1.35 (0.18)	-0.93 (0.35)	4.15 **(0)**	
	Avg.	2.97 **(0)**	3.81 **(0)**	-1.10 (0.27)	2.19 **(0.03)**	
O_3	S1	4.56 **(0)**	9.26 **(0)**	-5.30 **(0)**	-9.70 **(0)**	
	S2	2.76 **(0)**	7.42 **(0)**	-8.90 **(0)**	-13.37 **(0)**	
	S3	1.12 (0.26)	3.99 **(0)**	-3.79 **(0)**	-3.28 **(0)**	
	Avg.	3.08 **(0)**	7.60 **(0)**	-7.36 **(0)**	-10.39 **(0)**	
O_3	S1	0.41 (0.68)	7.82 **(0)**	-7.85 **(0)**	-13.27 **(0)**	16.76 **(0)**
	S2	0.04 (0.97)	6.00 **(0)**	-10.65 **(0)**	-16.90 **(0)**	12.41 **(0)**
	S3	-1.17 (0.24)	4.77 **(0)**	-4.75 **(0)**	-8.12 **(0)**	15.96 **(0)**
	Avg.	0.22 (0.83)	7.44 **(0)**	-9.33 **(0)**	-15.24 **(0)**	14.96 **(0)**
O_3 - NO_2	S1	-0.87 (0.38)	5.61 **(0)**	-7.46 **(0)**	-12.39 **(0)**	
	S2	-1.22 (0.22)	4.21 **(0)**	-9.60 **(0)**	-15.62 **(0)**	
	S3	-2.61 **(0.01)**	3.01 **(0)**	-3.02 **(0)**	-7.69 **(0)**	
	Avg.	-1.24 (0.21)	4.70 **(0)**	-7.96 **(0)**	-13.88 **(0)**	

TABLE 3.12: Summary of the t- and p-values for all models. The t-values are on the top row, whereas the p-values are in parenthesis in the bottom row and the significant p-values are bolded.

In all models, the main pollutant from the reference monitor is significant (the 95% confidence interval does not contain zero) but there is always at least one other significant covariate as seen in Table 3.11. The only exception is the model for the NO_2 WE voltage from Sensor 3, for which the only significant covariate is relative humidity. This is also confirmed by the p-values presented in Table 3.12. However, both reference pollutants are significant for all models (except for Sensor 3), which suggests that the MAS are in fact measuring a combination of both pollutants. The parameters in Table 3.11 are positive, which means that as both reference pollutant levels increase, so do the MAS' voltages for both pollutants. It is clear that the main pollutant has a bigger t-value in all cases except for Sensor 3 and averaged NO_2 WE voltage, indicating that the main pollutant has a bigger influence on the response as shown in Table 3.12. It also has to be noted that most non-significant parameters have 95% confidence intervals which are not symmetric around zero suggesting that there might be a weak relationship, which the models are unable to capture.

From Table 3.11, it is clear that for most NO_2 WE voltage models, temperature and relative humidity are not significant as their 95% confidence intervals contain zero. Furthermore, the p-values presented in Table 3.12 are larger than 0.05 suggesting that the weather did not have a significant effect on the NO_2 WE voltage MAS' hourly measurements. Additionally, the t-values for temperature and relative humidity for the NO_2 WE voltage models are close to zero as shown in Table 3.12. However, for the O_3 WE voltage models, the t-values for temperature and relative humidity have really high values and appear to be the most influential variables in the models, suggesting that the MAS hourly measurements' for the O_3 WE voltage were heavily influenced by the weather. It is interesting to note that when modelling the difference O_3 - NO_2 WE voltage, the results are very similar to those for the standard O_3 WE voltage model as temperature is the most influential variable followed by relative humidity and only then comes the reference O_3 level (Table 3.12). Due to the effect of NO_2 on the MAS' hourly measurements for the O_3 WE voltage in [10], an additional variable was added to the model - the NO_2 WE voltage for the respective sensor as there were reasons to believe that the O_3 WE voltage reflects the variation in the NO_2 WE voltage as well. As a result of this, the reference NO_2 loses its significance for all of the models for the O_3 WE voltage. From Table 3.11, it can be seen that in those models, the NO_2 WE voltage is always significant and has a positive estimate suggesting that as the NO_2 WE voltage increases so does the O_3 WE voltage. Furthermore, for Sensors 1 and 3, the NO_2 WE voltage has the biggest influence based on its t-values in Table 3.12, and for Sensor 2 and the averaged values, it is the second most influential covariate after relative humidity in contrast to the results for the model without the NO_2 WE voltage as a covariate.

3.4 Conclusion

In this chapter, the ability of deploying MAS (in this case, ALPHASENSE B2 electrochemical sensors unit) as implemented in the AirSpeck systems "out of the box" (i.e. without prior laboratory calibration) was evaluated by comparing the hourly measurements of three such sensors for NO_2 and O_3 with each other as well as their respective reference monitors. The MAS voltage (AE and WE) outputs for two pollutants, NO_2 and O_3 showed clear correlations with reference observations based on a deployment without prior calibration or validation, as it would typically be applied in the context of citizen science applications utilising sensor packages with limited prior expert knowledge. While this is promising and indicates that the electrochemical sensors on the market produce relevant data for environmental observation, the strength of the relationship and the indication of several covariates suggest that while the cost of the sensor

hardware is low, substantial time and expertise is required to derive robust, reliable and quality controlled data from such sensor packages.

During the preliminary analysis in Section 3.1 and the Bland-Altman analysis in Section 3.2, it was found that the AE voltages for both pollutants were inconsistent with each other (Figures 3.3, 3.4, 3.8, 3.9 and 3.10) and, therefore, they were not used for further modelling. Two thirds of the measurements from the MAS' WE voltage were found to be in agreement with each other, although there were moderate to high correlations between the measurements (Figures 3.11, 3.12 and 3.13). There appeared to be a linear relationship between the MAS' and the AURN reference sensor's hourly measurements (Figure 3.7) and, therefore, linear regression was used to check how well the MAS measurements are associated with the AURN reference sensor and measure air pollution. However, all models have problems with non-independent errors, which requires a GLS fit. To improve the quality of the models, temperature and relative humidity were used as covariates besides the AURN reference sensor measurements. Correlation structures were chosen based on minimising the AIC and BIC with preference to BIC as it favours models with a smaller number of covariates. In all models, the main pollutant from the reference monitor was always significant. However, both pollutant concentrations are significant in all models except for those for Sensor 3, indicating that a combination of the two pollutants was measured by the MAS. For the NO_2 models, the reference NO_2 hourly measurements have the highest t-value (Table 3.12) in almost all cases, hence, it has the biggest influence on the MAS' hourly measurements. However, it is interesting to note that although the Bland-Altman analysis had shown that the NO_2 WE voltages from Sensors 1 and 3 are consistent with each other, but not with the hourly measurements from Sensor 2, the models show that the NO_2 WE voltages from Sensors 1 and 2 are influenced by the changes in the reference NO_2 hourly measurements, whereas the NO_2 WE voltages from Sensor 3 are reflecting the changes in relative humidity. For the O_3 models, temperature followed by relative humidity have the highest t-values (Table 3.12) which means that the weather was heavily influencing the O_3 hourly measurements by the MAS. Two statistical approaches were used to investigate the influence of the NO_2 WE voltage on the O_3 WE voltage - one was using the NO_2 WE voltage as a covariate when modelling O_3 WE voltage and the other was taking the differences between the two voltages. It is interesting to note that the model for O_3 WE voltage, which includes NO_2 WE voltage as a covariate, the response is mostly influenced by NO_2 WE voltage and relative humidity (Table 3.12). In contrast, the models for O_3 - NO_2 WE voltage support the conclusions from the models for O_3 WE voltage with four covariates with the only difference being that the reference NO_2 hourly measurements are no longer significant.

In general, it appears that the MAS hourly measurements are not always in agreement with each other, but the MAS WE voltages have managed to capture the changes in the pollutants' concentrations. However, the MAS are performing better for measuring NO_2 than for O_3. The MAS' voltages, especially for O_3, seem to have been heavily influenced by the meteorological conditions. Based on this, it has to be concluded that when operating MAS packages "out of the box", caution needs to be applied depending on the pollutants measured. Overall, the MAS measurements are not yet of the quality to supplement high quality sensors such as the AURN ones.

Chapter 4

Exploratory analysis of the Aberdeen and Glasgow NO_2 data

Chapter 4 presents the exploratory analysis of the NO_2 monitoring and ADMS-Urban simulated data in Aberdeen and Glasgow data sets used further on in this thesis. Section 4.1 presents the NO_2 data for Aberdeen outlining the monitoring system, checks for breaches in the regulation based on the monitoring data and explores the simulated data. Section 4.2 mirrors the organisation of Section 4.1 but focuses on the 2015 NO_2 monitoring data for Glasgow. Section 4.3 provides a concluding discussion.

4.1 Aberdeen

4.1.1 Monitoring system

Aberdeen is the third largest city in Scotland known for its economic importance as centre of the oil industry in the country. This results in traffic in the city and it is of interest to identify if there are locations of air pollution levels above the regulatory limit. As part of the UK wide AURN network, the Aberdeen monitoring system consists of six active monitoring stations in the city. Using [8], a short review of the stations is presented below with a map with the locations of the stations in Figure 4.1. Further details and pictures of the stations are available at the Scottish Air Quality website [8].

- **Anderson Drive** - the station is located four metres away from the kerbside. Anderson Drive is a roadside station. The station takes hourly measurements of NO_2 and PM_{10}.

- **Errol Place** - the station is located thirty metres away from the nearest road. Errol Place is an urban background station. The station takes hourly measurements of NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2 (sulphur dioxide).
- **King Street** - the station is located two metres away from a major road with heavy traffic from heavy goods vehicles (HGVs). The station is a roadside station. The station takes hourly measurements of NO_2, PM_{10} and $PM_{2.5}$.
- **Market Street 2** - the station is located on the pavement near Aberdeen Harbour. Market Street 2 is a roadside station. The station takes hourly measurements of NO_2, PM_{10} and $PM_{2.5}$.
- **Union Street** - the station is located two metres from the kerbside with many buses passing nearby. Union Street is a roadside station. The station takes hourly measurements of NO_2 and PM_{10}.
- **Wellington Road** - the station is located four metres from a major road, close to the roundabout at Queen Elizabeth II Bridge with heavy traffic, mostly consisting of HGVs. Wellington Road is a roadside station. The station takes hourly measurements of NO_2, PM_{10} and $PM_{2.5}$.

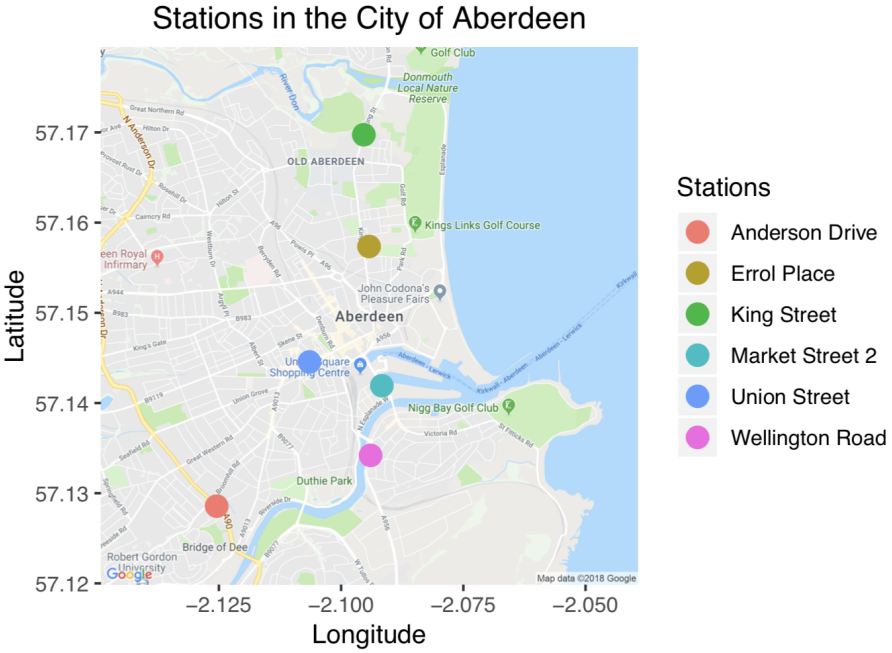

FIGURE 4.1: Map for the six AURN monitoring stations across Aberdeen in 2012 [92].

It has to be noted that the pollutant concentrations for all the stations are reported as rounded to the nearest whole number (in μg m^{-3}) in Aberdeen. Therefore, although the pollutant concentrations are continuous, the data are actually discretised.

4.1.2 NO_2 monitoring in 2012

DEFRA have identified NO_2 as the most difficult to tackle pollutant in the whole of the UK, requiring its own separate strategy [61]. Aberdeen is heavily influenced by road traffic due to the oil industry making the modelling of N_2 concentrations in the city of key interest. Hence, a variety of graphical and numerical summaries of the 2012 data are provided. Based on this initial analysis, it is aimed to answer two questions about the compliance with the EU legislation [76]:

(i) Has the hourly NO_2 concentration limit of 200 μg m^{-3} been breached more than 18 times (with consecutive breaches counted individually) in 2012?

(ii) Is the average annual mean for NO_2 concentration above the limit of 40 μg m^{-3}?

To answer question (i), Table 4.1 provides a count of the breaches of the hourly limit of 200 μg m^{-3} over the six Aberdeen monitoring stations as well as displaying the number of missing observations and the total number of observations each station has taken through the year. The maximum number of hourly concentrations in a leap year is 8784. In Table 4.1, it is clear that the number of breeches does not exceed the regulation, although observations above 200 μg m^{-3} are present - 1 at Union Street and 10 at Wellington Road. It has to be noted that the monitor at King Street is missing over one thousand observations due to the monitor not working for a month in the autumn.

Station	Breaches	Missing	Total
Anderson Drive	0	518	8689
Errol Place	0	581	8538
King Street	0	1068	8666
Market Street 2	0	381	8514
Union Street	1	228	8674
Wellington Road	10	391	6591

TABLE 4.1: A count of the number of breaches of the hourly concentration limit of 200 μg m^{-3} in Aberdeen in 2012. The missing values and total number of observations per station for the year are also provided.

Next, to help visualise the data and check for any abnormalities, time series plots for the hourly concentrations of each of the stations are presented in Figure 4.2. It is clear that high NO_2 hourly concentration at Union Street occured mid-May and does not coincide with an occurrence at Wellington Road. The occurrences at Wellington Road appear in February and April. It also has to be noted that the concentrations at Market Street 2 are close to the 200 μg m^{-3} limit, although the limit is never breached. The data appear consistent with the assumption that these stations are located close to roads with heavy

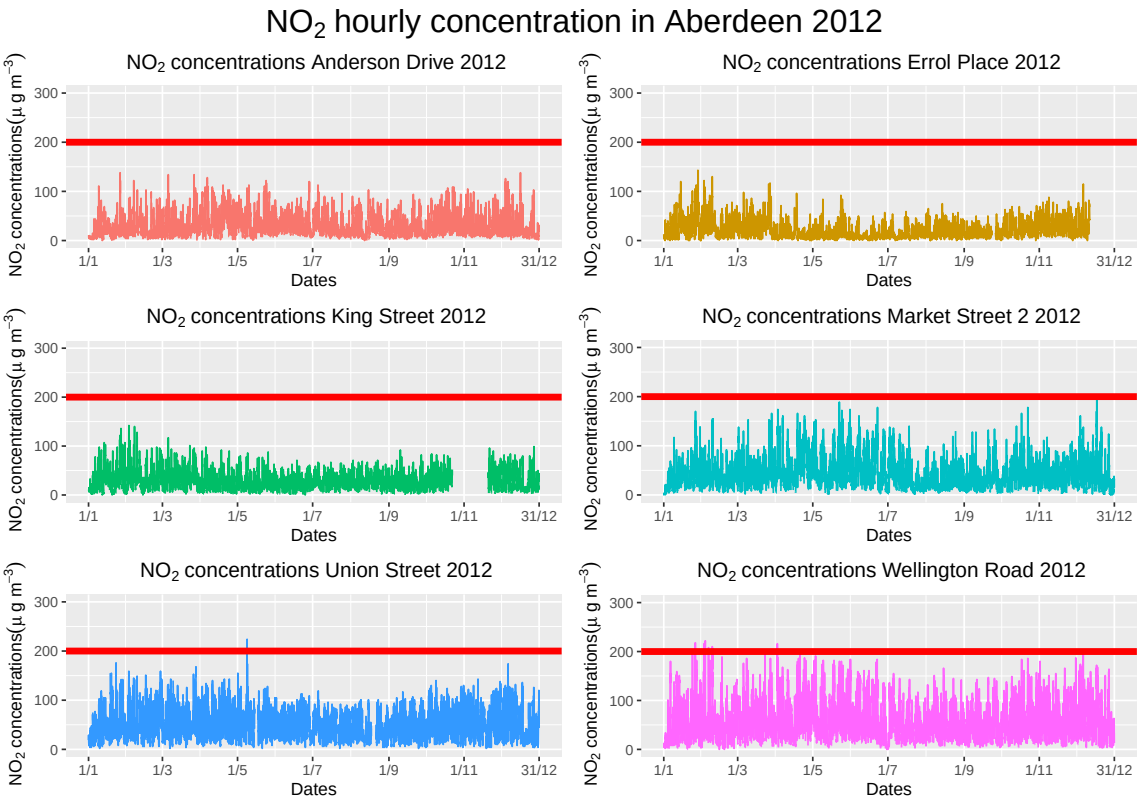

FIGURE 4.2: Time series plot for the hourly NO_2 concentrations (μg m^{-3}) for each of the six AURN monitoring stations in Aberdeen in 2012. The hourly NO_2 limit of 200 μg m^{-3} is represented by a solid red line.

traffic with many HGVs required for the oil industry. Unsurprisingly, the Errol Place concentrations are lowest, given that the station is an urban background station.

To check for skewness in the data, the histograms for the measurements at each station are examined in Figure 4.3. The histograms for all six stations show right-skew so log and square root transformations were applied and the log-transformation was chosen based on the histograms in Figure 4.4. Overall, the histograms appear more symmetrical.

Since the hourly concentrations do not indicate a non-compliance issue, the annual means were examined in order to answer question (ii) and are presented in Table 4.2, where the values above the regulation are highlighted in red. Since there are missing observations for each of the stations, confidence intervals for the mean are also provided to give a range of plausible values for the true mean. A standard 95% confidence interval is provided. However, the standard confidence interval does not account for the correlation between the observations of time series data. Therefore, a 95% bootstrap interval and an adjusted for the correlation interval are also provided (estimated as described in Subsection 2.1.4),

However, before calculating the 95% correlation adjusted confidence intervals, the PACF plots (calculated as described in Subsection 2.1.2) for the time series of each of the

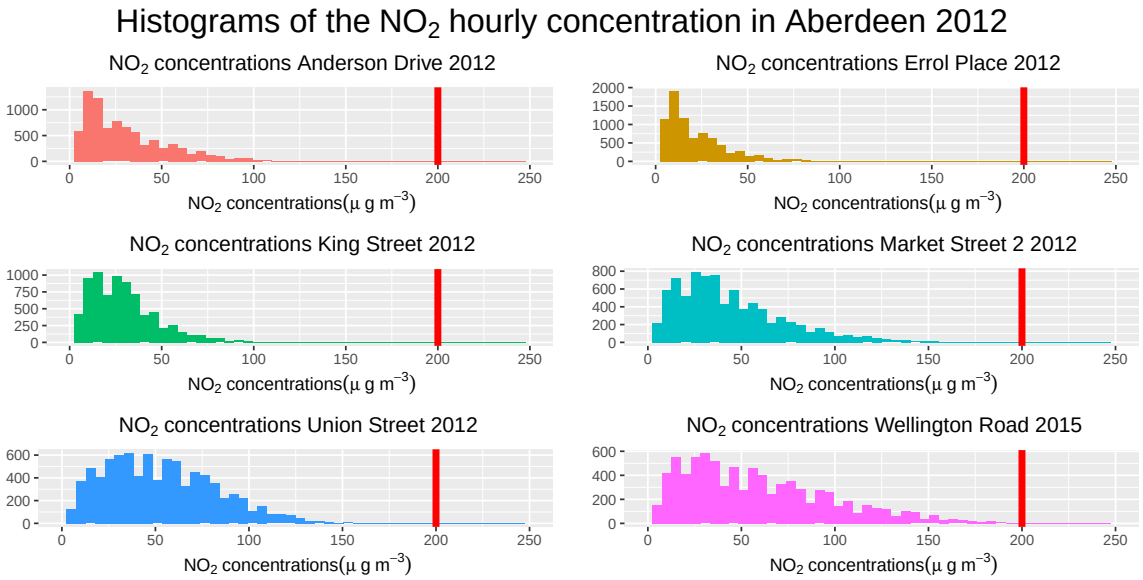

FIGURE 4.3: Histograms for the hourly NO_2 concentrations (μg m^{-3}) for each of the six AURN monitoring stations in Aberdeen in 2012. The hourly NO_2 limit of 200 μg m^{-3} is represented by a solid red line.

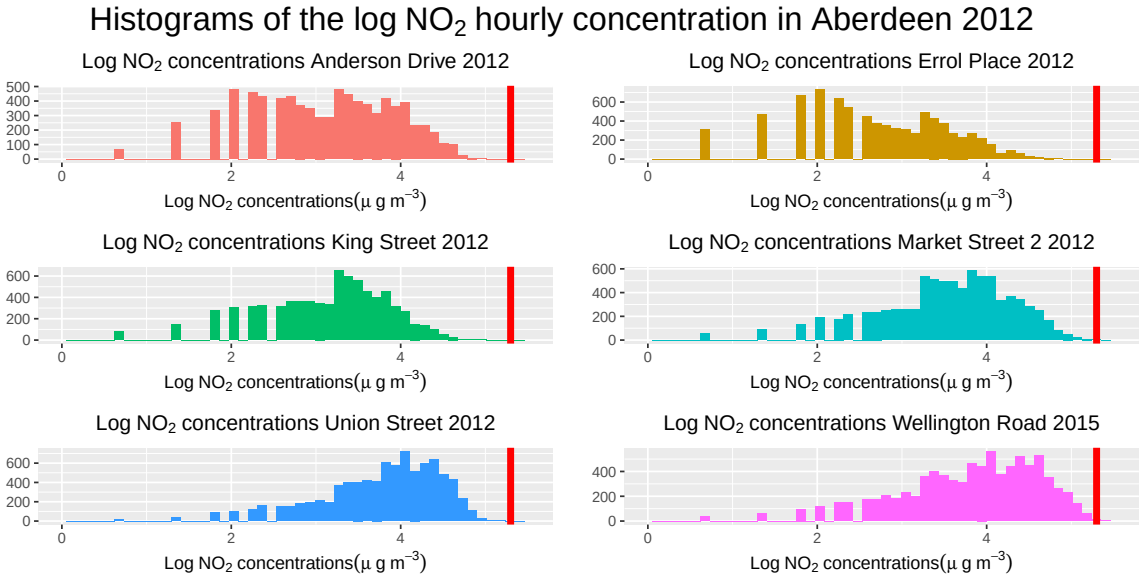

FIGURE 4.4: Histograms for the hourly log NO_2 concentrations (μg m^{-3}) for each of the six AURN monitoring stations in Aberdeen in 2012. The hourly log NO_2 limit of log(200) μg m^{-3} is represented by a solid red line.

stations must be examined to check the type of AR process of the data in Figure 4.5. For all stations, for the first lag the observations have strong positive autocorrelation, which is expected for hourly measurements of pollution concentrations. There is a possible diurnal seasonality which is expected as during the day concentrations are more influenced by the emissions whereas during the night the hourly concentrations are more dependant on the previous hour's concentrations. Overall, there is significant autocorrelation for two or three lags, i.e. the correlation is above the standard error bars

and then similarly for the last few hours of the day. However, the standard error bars are calculated based on the size of the time series (8784). As the size of the time series gets larger, the standard errors get closer to zero. This results in very narrow error bands, which should in turn be examined as relative rather than exact error bands. As a result of the narrow error bands, more lags are likely to appear marginally significant although this is just by chance since the error bands do not depend on the structure of the data but only its size. It is interesting to note that the second lag for all stations is negative which is expected when observing an AR process with a very strong positive autocorrelation at the first lag. Therefore, an AR(1) process is chosen as a reasonable simplification for the estimation of the 95% correlation adjusted confidence intervals.

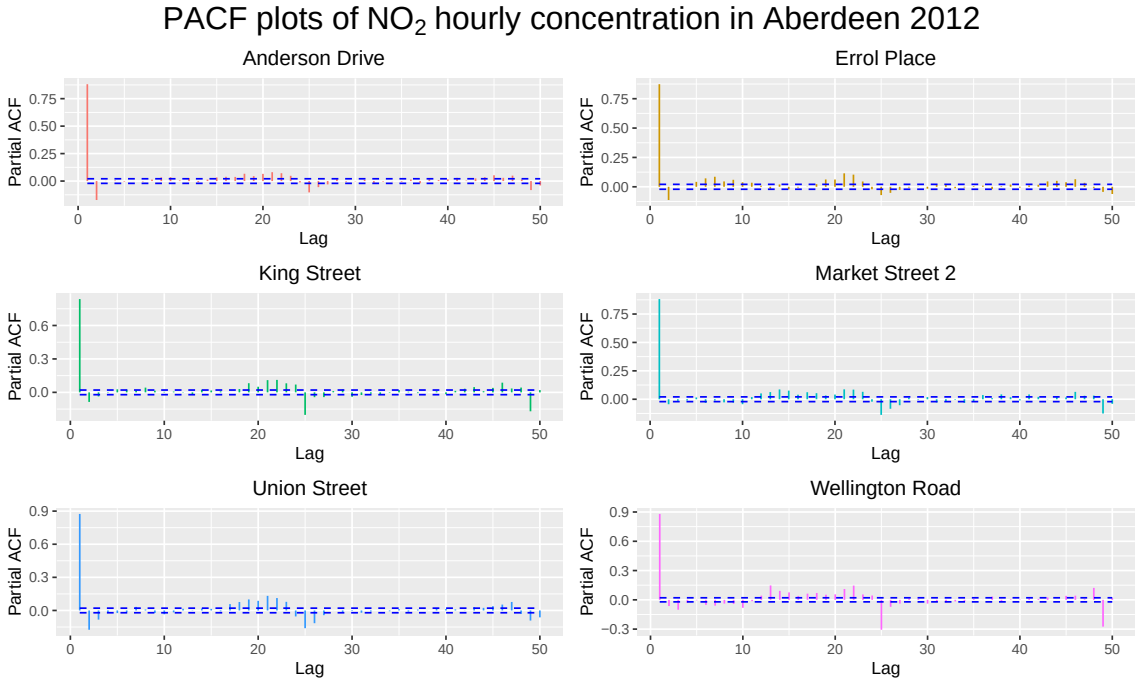

FIGURE 4.5: PACF plots for the time series of the hourly NO_2 concentrations (in μg m^{-3}) for each of the six AURN monitoring stations in Aberdeen in 2012 up to lag 50.

Station	Annual Mean	St. interval	Bootstrap	Corr. adjusted
Anderson Drive	30.38	(29.90, 30.86)	(29.91, 30.88)	(20.59, 32.18)
Errol Place	21.00	(20.63, 21.36)	(20.62, 21.39)	(19.60, 22.40)
King Street	29.17	(28.77, 29.56)	(28.76, 29.60)	(27.84, 30.49)
Market Street 2	**44.04**	**(43.42, 44.65)**	**(43.42, 44.68)**	**(41.64, 46.43)**
Union Street	**52.84**	**(52.23, 53.46)**	**(52.23, 53.48)**	**(50.45, 55.24)**
Wellington Road	**59.02**	**(58.20, 59.85)**	**(58.20, 59.85)**	**(55.64, 62.40)**

TABLE 4.2: Comparing the annual mean and the three different types of 95% intervals (standard confidence interval, bootstrap and correlation adjusted confidence) for the hourly NO_2 concentrations (μg m^{-3}) across the six AURN monitoring stations in Aberdeen in 2012.

Three types of intervals are presented in Table 4.2. Expectedly, the standard confidence interval is too narrow because it does not adjust for the correlation in the data. The

bootstrap intervals are almost identical to the standard confidence interval. The bootstrapping was not performed using blocks of data to avoid the potential influence of the repetition of periods of high concentrations. The adjusted for correlation 95% confidence interval is wider than the others which makes it the most realistic interval of the three given. From Table 4.2, it is clear that for three of the stations (Market Street 2, Union Street and Wellington Road), the annual average is above the limit. Market Street 2 has breached the limit by 10% of the regulation, whereas Union Street has breached by 32% and Wellington Road by almost by 50%. Furthermore, the intervals for the annual means at these stations do not contain the limit of 40 μg m^{-3} indicating that these three stations are located at places, where the air pollution does not comply with the regulation. The other three stations (Anderson Drive, Errol Place and King Street) do not breach the 40 μg m^{-3} and their respective confidence intervals do not contain the limit either indicating that the air pollution regulations are not likely to be breached there.

However, the data from the monitoring stations are right-skewed (as seen in the histograms in Figure 4.3). Therefore, the mean values could easily be misleading as they are affected by the skewness in the data. Hence, the median values for the concentrations at each station are examined as the median is a quantile measure and is estimated based on the ordered pollutant concentrations. In a similar way to the mean estimation, different types of intervals for the median values of the six Aberdeen stations are provided as shown in Subsection 2.1.4. The simplest approach is to produce a quantile interval for the data by ordering the observations from smallest to largest and noting the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ quantiles as described in [87]. There are also bootstrap and 95% confidence interval alternatives. These three types of intervals for the medians are presented in Table 4.3.

Station	Annual Median	Quantile	Bootstrap	Corr. adjusted
Anderson Drive	23.00	(4.00, 88.00)	(23.00, 25.00)	(23.00, 25.00)
Errol Place	15.00	(2.00, 67.00)	(15.00, 15.00)	(15.00, 15.00)
King Street	25.00	(4.00, 75.00)	(25.00, 27.00)	(25.00, 27.00)
Market Street 2	38.00	**(6.00, 117.00)**	(36.00, 38.00)	(36.00, 38.00)
Union Street	**50.00**	**(8.00, 117.00)**	**(48.00, 50.00)**	**(48.00, 50.00)**
Wellington Road	**52.00**	**(8.00, 151.00)**	**(50.00, 52.00)**	**(50.00, 52.00)**

TABLE 4.3: Comparing the annual median and the three different types of 95% intervals (quantile, bootstrap and correlation adjusted confidence) for the hourly NO_2 concentrations (μg m^{-3}) across the six AURN monitoring stations in Aberdeen in 2012.

In Table 4.3, three different intervals for the median are presented. The quantile interval is too wide and provides limited information about the median in comparison to the other two intervals. The bootstrap and the confidence intervals are in agreement with

each other producing identical results. Furthermore, for both the bootstrap and the adjusted confidence intervals, in most cases, the median value is the same as the end of the interval. This is because all measurements are rounded to the nearest integer and, hence, there is a lot of repetition in the observed values for each station.

The median values for all the stations (Table 4.3) are always lower than the means (Table 4.2). The biggest change is for Market Street 2, where the median as well as the bootstrap and correlation adjusted intervals are within the regulatory limit, whereas for the mean the regulation was breached. In terms of the two main compliance questions, it appears that the recorded hourly observations comply with the EU regulations but the annual means for three monitoring stations (Market Street 2, Union Street and Wellington Road) are identified as "at risk".

4.1.3 ADMS-Urban simulations

ADMS-Urban simulation model

The monitoring stations provide information about the pollutant concentrations around the city. However, that information is insufficient to model the varying NO_2 concentration across the city. This issue is addressed by Directive 2008/50/EC [76], which sets the use of air quality models in addition to the monitoring data. The Cambridge Environmental Research Consultants (CERC) [37] have developed a series of air quality models called Atmospheric Dispersion Modelling Systems (ADMS). ADMS-Urban is the most comprehensive one as it models the pollutant concentrations in urban areas [36]. ADMS-Urban is a computer simulation model, which can be used to estimate the hourly pollutant concentrations at different city locations even at locations where there are no monitoring stations present. The ADMS-Urban model is often used for "developing and testing policy on air quality; the development of air quality action plans; investigation of air quality management and planning options for a wide range of sources including transport sources; source apportionment studies; air quality and health impact assessments of proposed developments and use of the model for the provision of detailed street-level air quality forecast"[36]. For a set of inputs, the ADMS-Urban simulation model provides an hourly forecast for one or multiple pollutant concentrations for a year. The simulation model takes the following as inputs:

- **City Outline** - The road network in the city; the width of each road in the road network, which is important for the dispersion of the pollutants; heights of buildings as they form the city's terrain in terms of street canyons;

- **Traffic** - Observations of traffic flows at key junctions across the city between 07:00 and 19:00. The observations are collected on only one day of the year based on which an average flow of vehicles is determined for each road section as well as a diurnal (daytime) cycle;
- **Road and Background Emissions** - Emission rates are calculated using background emissions (such as commercial and domestic sources) and the average flow for each road section (calculated based on the traffic data) using the EMIT tool developed by CERC. The diurnal cycle is also used by EMIT to create a multiplying emission rate by an hour factor to adjust for the varying emissions throughout the day. For more information on EMIT, visit http://www.cerc.co.uk/environmental-software/EMIT-tool.html. EMIT produces hourly emission readings for a year. However, it has to be stressed that the emissions are created based on a diurnal cycle where the 24 hours are multiples of a certain baseline;
- **Meteorological Data** - Hourly measurements of temperature, wind speed and wind direction for a year. The meteorological data come from the MET office Bishopton station. Hence, the meteorological data are the same for the whole city. However, modelling techniques are used to provide variation in the meteorological data based on geographical locations; and
- **Chemistry** - Information about the interactions between different pollutants in the air.

ADMS-Urban is a deterministic air quality simulation model. Therefore, for the same set of inputs, the simulation will always produce the same NO_2 hourly concentrations and annual averages, which are comparable to EU regulation [124]. However, ADMS-Urban has some uncertainty in terms of predicting the exact observed NO_2 concentrations built into it because of the inputs. A major uncertainty in the inputs comes from the fact that the average flow of emissions is determined based on observations from one day in the year. Therefore, it has to be stressed that the emissions are created based on a diurnal cycle where the 24 hours are multiples of a certain baseline. Nonetheless, the EMIT tool has some built-in uncertainties to add naturalism. Overall, the uncertainty is most visible when running the ADMS-Urban model under the actual conditions through a set period of time. Hence, it has to be noted that although ADMS-Urban is a deterministic simulation model, there are inherited uncertainties from some of the inputs. While these inputs are not changed in the runs used in this thesis, it is important to establish a framework with which these inputs can later be investigated. However, there is a discrepancy between the simulated hourly predictions for NO_2 in ADMS-Urban and the observed hourly concentrations for NO_2. Even though there are discrepancies

between the simulated and actual NO_2 hourly concentrations, the ADMS-Urban simulations track well the cycle of the NO_2 concentrations and that results in very similar annual mean concentrations for the simulation and the actual data. More information on these comparisons is available in [33]. Nevertheless, the ADMS-Urban model is popular for studies of different pollutants: [159] investigates CO concentrations in the city of Ravenna, Italy, [152] investigates PM_{10} concentrations in London and [57] investigates NO_2 concentrations in Kaunas city, Lithuania.

SEPA uses the current version of the ADMS-Urban 4.1 model as an addition to the monitoring stations as the EU directive [76] prescribes. The ADMS-Urban model could be used to predict the NO_2 hourly concentration for a year for given "point, line, area, volume and grid source models" [36]. However, the ADMS-Urban model has two major drawbacks. One is that to use the ADMS-Urban model, it is highly advisable that the user undergoes training. The second drawback of the ADMS-Urban model is that it takes a long time to run if all points across a city are to be evaluated. There are many inputs to be considered as seen above, but the key interest lies in understanding how the changes in the meteorological data and the information about the road network including the emission rate for each road section (which for brevity will be referred to as emissions) affect the NO_2 pollutant concentration. Instead of having to re-run the whole ADMS-Urban simulation model for a change in at most three inputs, it would be computationally much easier to use an emulator, a general description for which has been provided in Section 2.4.

In order to create an emulator, a number of ADMS-Urban simulation runs are required. The emulation of deterministic computer models (such as ADMS-Urban) is discussed in [169]. The main issue when emulating deterministic models is that for the same inputs, the same outputs are produced resulting in a lack of random noise. However, [169] points out that "modelling of a computer code as if it was a realization of a stochastic process, ..., gives basis for the quantification of uncertainty [around the predictions from the fitted model] and a statistical framework for design and analysis". Therefore, [169] recommends modelling using Gaussian Processes to establish the framework for modelling non-deterministic (stochastic) computer simulations.

In order to perform emulations, simulations are chosen based on a Latin Hypercube design as discussed in Subsection 2.4.1 and applied in [82]. The simulated data was produced by SEPA before being provided. For SEPA, it is important to investigate the changes in hourly NO_2 pollutant concentrations, as well as the NO_2 annual average concentrations, across the city while the meteorological conditions such as temperature, wind speed and wind direction are changing. However, SEPA believed that wind speed and temperature have a weak positive correlation. Therefore, it was decided that only

one of the two variables is going to be used to create the emulator. Wind speed was preferred to include in the emulator over temperature because it has a more immediate effect on the pollutant dispersion. Wind direction is crucial to keep because changing the wind direction will change the terrain in which the pollutants are diffused. Therefore, wind direction is also included in the emulator. Finally, the emissions themselves are included in the emulator as the emulator has to be able to account for changes in the pollutant concentrations as well as the meteorological conditions. It is crucial to note that due to licensing issues, there is no actual meteorological data available for Aberdeen presented in this thesis.

ADMS-Urban annual simulations for Aberdeen

In 2012, SEPA used ADMS-Urban to simulate 98 scenarios for each of the six monitoring stations in Aberdeen. The 98 scenarios were chosen using a Latin Hypercube (LHC) design as described in [82]. The LHC design is created by varying emissions (from -50% to +30%), wind speed (from -20% to +20%) and wind direction (from -15° to +15°) in comparison to the baseline set by the observed data in 2012. The ranges of the plausible values on which the inputs are varied were chosen by SEPA based on prior knowledge. The new inputs are generated such that all hourly values are multiplied by the percentage change for emissions or wind speed, or for wind direction, the degree change has been added to all hourly observations. The LHC design is presented in Figure 4.6.

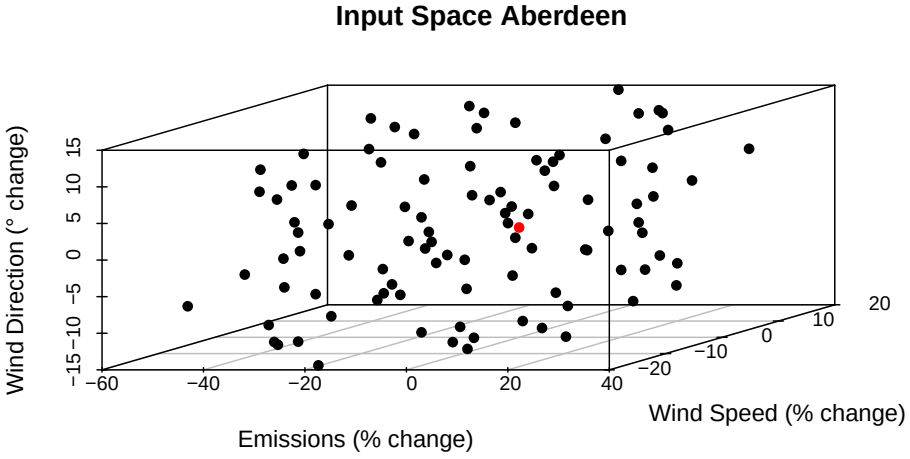

FIGURE 4.6: Input space for the ninety-eight simulations of the annual NO_2 average concentrations (μg m^{-3}) in Aberdeen. The point $(0, 0, 0)$ representing the true values for emissions (% change), wind speed (% change) and wind direction (° change) is in red.

The actual NO_2 annual average concentrations (as blue triangles) are compared to the boxplots for the NO_2 annual averages from simulations in Figure 4.7. It is expected that the actual annual average will be slightly higher than the majority of the simulated ones

as the point $(0, 0, 0)$ is not the centre of the design sample space. The two stations for which there are no simulations breaching the 40 μg m^{-3} annual regulation are Anderson Drive and Errol Place. While the true NO_2 concentration for Anderson Drive is lower than 52% of simulated values, all the simulated values for Errol Place are higher than the observed one. Hence, these two stations are not "at risk". Previously, there has been no evidence to label King Street as an "at risk" monitoring station but there are several simulations that breach the 40 μg m^{-3} annual regulation. However, the true concentration is lower than 99% of the simulated values suggesting that ADMS-Urban tends to over-predict the NO_2 annual concentrations at King Street. For the Market Street 2 monitoring station, the true NO_2 annual average is higher than 42% of the simulated ones. The majority of the simulations are above the 40 μg m^{-3} limit which is consistent with the station being identified as "at risk". The simulations for Union Street are also predominantly higher than the 40 μg m^{-3} limit, although only 21% are higher than the true NO_2 annual average. At Wellington Road, although the majority of the simulated values are above the regulatory limit of 40 μg m^{-3}, the true concentration is higher than any of the simulated values suggesting that the ADMS-Urban model is under-predicting the NO_2 annual average at the location. Overall, the simulated NO_2 averages by the ADMS-Urban model are reasonable for three monitoring stations (Anderson Drive, Market Street 2 and Union Street), higher for two monitoring stations (Errol Place and King Street), and lower for one monitoring station (Wellington Road).

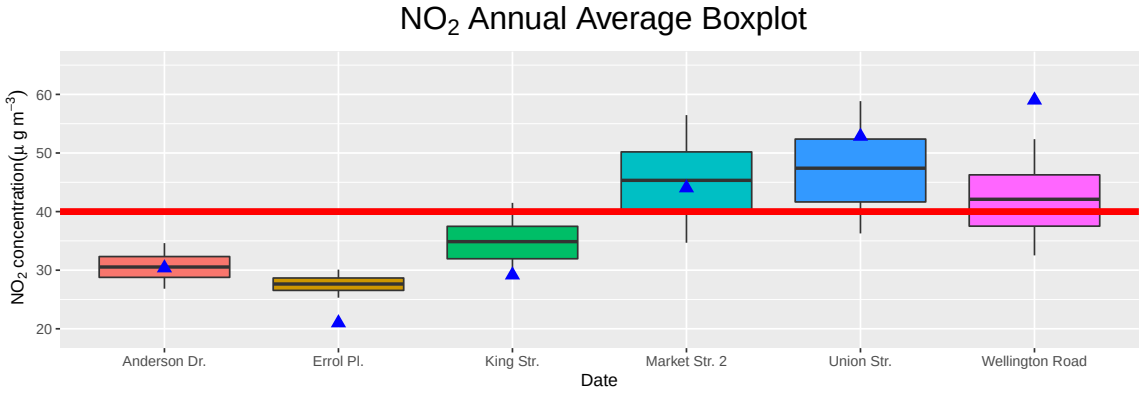

FIGURE 4.7: Boxplots for the NO_2 annual averages (μg m^{-3}) from ninety-eight simulations of ADMS-Urban for each of the six monitoring stations in Aberdeen. The yearly NO_2 limit of 40 μg m^{-3} is represented by a solid red line. The true annual average in 2012 for each station is signified by a blue triangle.

4.1.4 Variograms

As previously stated, the ADMS-Urban simulation scenarios for Aberdeen are chosen using a 3-dimensional LHC design using variations of **emissions** (from -50% to +30%), **wind speed** (from -20% to +20%) and **wind direction** (from -15° to +15°) to the

baseline observations in 2015. There is a need to investigate whether there is evidence of correlation between the points (defining the simulation scenarios) in the input space using a variogram. An exploratory tool with which the presence of spatial correlation in the LHC locations, which define the simulation scenarios, can be assessed is a variogram plot, where the semivariance is plotted against the distances between points. In Subsection 2.3.1, it was established that the **sill** provides information about the limiting value of the variogram as the distance between points goes to infinity, the **range** is the distance for which the variogram reaches the sill and the **nugget** is the limiting value of the variogram at distances close to zero. All variograms in this subsection are computed using the **gstat** [149] package in R [157]. These variograms will be used to provide an initial impression of likely estimates for each of the hyperspatial range parameters for the three inputs forming the LHC space. The three variables will be referred to as the **hyperspatial range parameters** for the rest of the thesis. The variograms of the simulated NO_2 annual average concentration for each of the stations in Aberdeen for the three inputs are examined in Figures 4.8, 4.9 and 4.10. It has to be noted that Monte Carlo envelops were not added to the variograms as the points of the variograms do not reach a peak and plateau afterwards.

Firstly, the simulated NO_2 annual average residuals (from an intercept only model) variograms for each of the stations in Aberdeen based only on the emissions input are examined in Figure 4.8. Although it is expected that the points will reach a peak and plateau afterwards, on all the variograms the points are constantly increasing. Since the LHC space was designed to have spatial correlation between the scenarios, this result is not surprising but it means that the sill and the range parameters appear to be going to infinity within the observed space. The variograms do not have a nugget effect which means there is no measurement error. For the stations with overall higher NO_2 annual concentrations (Market Street 2, Union Street and Wellington Road), the semivariance has much larger values in comparison with the other three stations (Anderson Drive, Errol Place and King Street).

Next, the simulated NO_2 annual average residuals (from an intercept only model) variograms for each of the monitoring stations in Aberdeen based only on the wind speed input are examined in Figure 4.9. Although the variograms in Figure 4.9 also indicate the presence of spatial correlation, there are quite a few differences between the variograms in Figure 4.9 and those in Figure 4.8. Most obviously, the points on all the variograms in Figure 4.9 appear to plateau at a range of 30 but instead of plateauing afterwards, the points are decreasing. This indicates that the hyperspatial range parameter for wind speed will be easier to calculate than the one for emissions. Furthermore, there appears to be a nugget effect present for four stations (King Street, Market Street

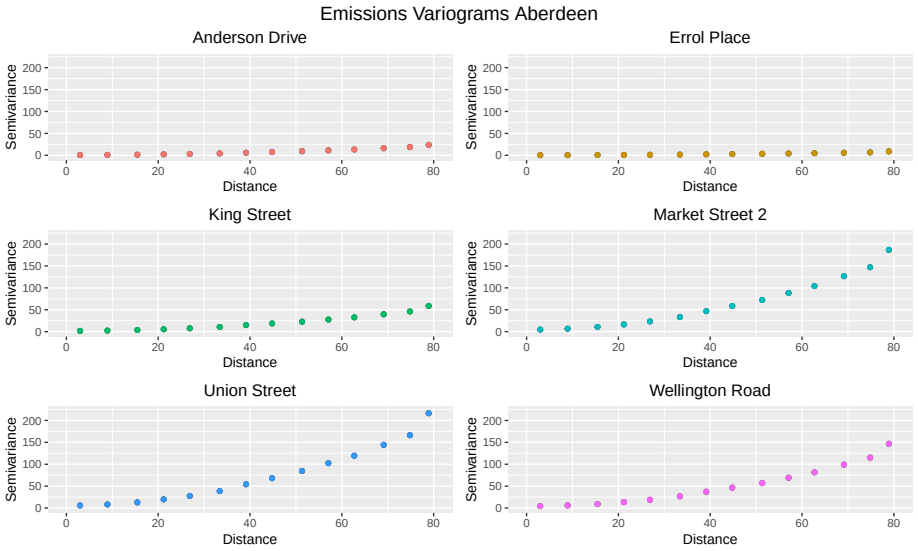

FIGURE 4.8: Emission variograms for the six monitoring stations in Aberdeen.

2, Union Street and Wellington Road) suggesting a small measurement error in the wind speed.

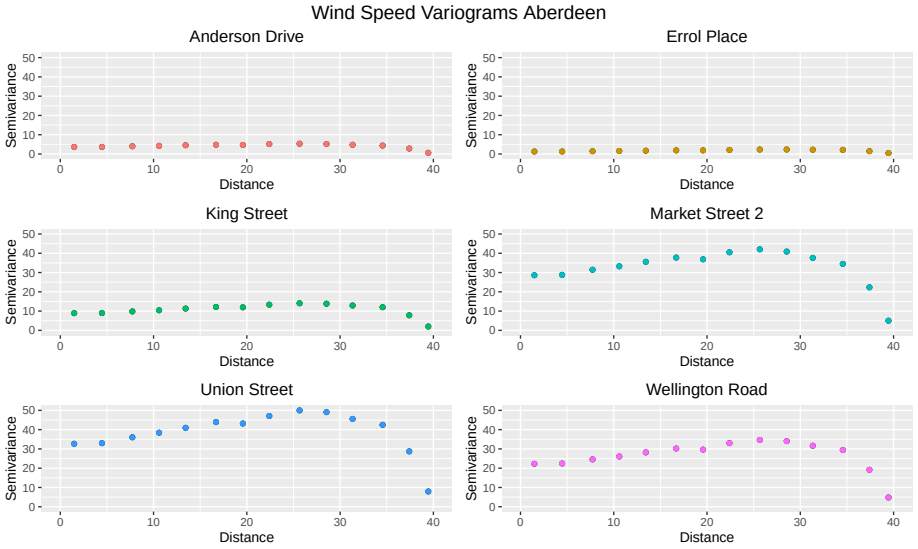

FIGURE 4.9: Wind speed variograms for the six monitoring stations in Aberdeen.

Lastly, the simulated NO_2 annual average residuals (from an intercept only model) variograms for each of the monitoring stations in Aberdeen based only on the wind direction input are examined in Figure 4.10. The wind direction variograms appear to plateau at about a value of 20. However, there seems to be an increase in the semivariance values just before a distance of 30 and since that is outside the scope of the LHC, it is possible that the hyperspatial range parameter based on the wind direction would go to infinity in a similar fashion to the range parameter for emissions. Therefore, it would be difficult to estimate the wind direction hyperspatial range parameter. As

with the wind speed variograms in Figure 4.9, there is a nugget effect suggesting there is a small measurement error in the wind direction.

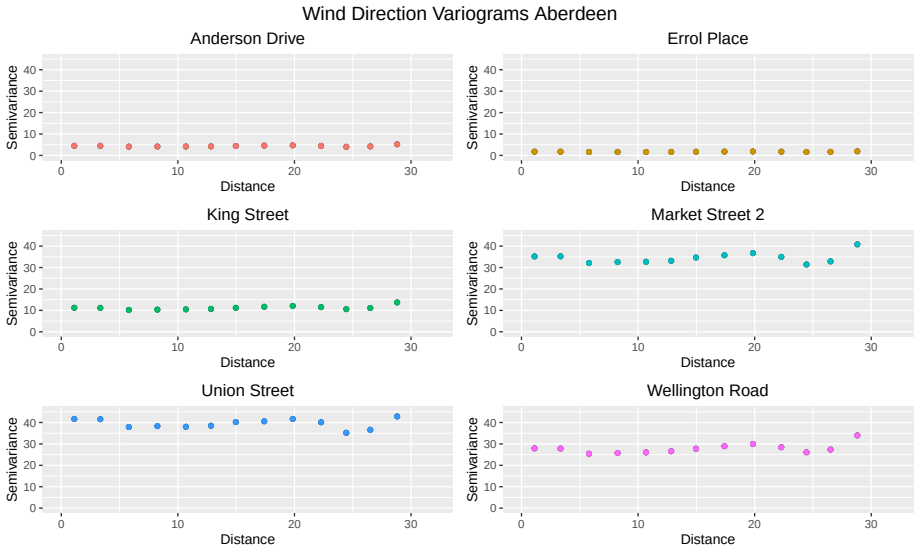

FIGURE 4.10: Wind Direction variograms for the six monitoring stations in Aberdeen.

Since the hyperspatial range parameters are modelled together in the following chapters, it would be beneficial to examine the joint effect of the three inputs for each station. To do this, 3D variograms can be used for all the stations. A 3D variogram is an extension of the standard 2D variogram [148] in the **gstat** package in R. In order to estimate a variogram, it is necessary to calculate the distances between the coordinates for each location. In the 2D variogram case, the distances for each location are estimated based on two coordinates (x, y), whereas in the 3D case, the distance for each location is estimated based on three coordinates (x, y, z). Referring back to the semi-variogram estimation for two spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$,

$$\widehat{\gamma}_Y(\mathrm{h}_k^m) = \frac{1}{2|N(\mathrm{h}_k)|} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathrm{h}_k)} [y(\mathbf{s}_i) - y(\mathbf{s}_j)]^2 \,, \tag{2.60}$$

the spatial locations are now defined to have coordinates $(a_{\mathbf{s}_i}, b_{\mathbf{s}_i}, c_{\mathbf{s}_i})$ and $(a_{\mathbf{s}_j}, b_{\mathbf{s}_j}, c_{\mathbf{s}_j})$, respectively. The variogram is intrinsically stationary and it is isotropic so only the distance between two points is important, i.e. $h = ||(a_{\mathbf{s}_i}, b_{\mathbf{s}_i}, c_{\mathbf{s}_i}) - (a_{\mathbf{s}_j}, b_{\mathbf{s}_j}, c_{\mathbf{s}_j})||$. Therefore, the interpretation of the 3D variogram is the same as for a 2D variogram. Hence, 3D variograms for each of the monitoring stations in Aberdeen based on the coordinates in the LHC design (Figure 4.6) are presented in Figure 4.11.

The 3D variograms for the simulated NO_2 annual average residuals (from an intercept only model) for each of the monitoring stations in Aberdeen in Figure 4.11 are almost identical to the emissions variograms based only on the emissions input in Figure 4.8. The constantly raising semivariance is an indicator that there is high correlation between

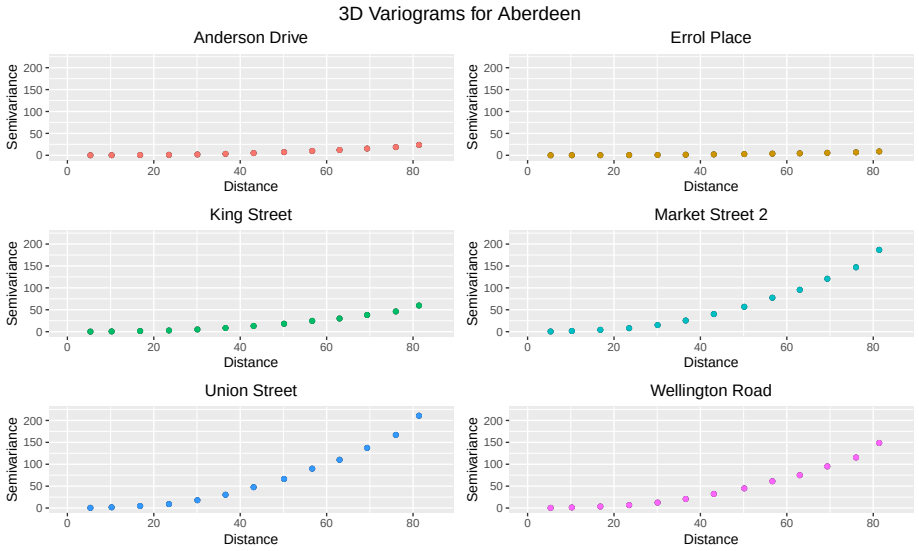

FIGURE 4.11: 3D variograms for the six monitoring stations in Aberdeen.

the different ADMS-Urban scenarios which would aid predictions in the LHC sample space. It is interesting to note that even though some of the individual variograms had nugget effects (Figures 4.9 and 4.10), there is no nugget present on any of the 3D variograms for any of the stations. Similarly to the previous variograms, the semivariance seems to increase for monitoring stations with higher NO_2 annual average concentrations. Overall, it appears that there is hyperspatial correlation between the scenarios with the main driving factor being emissions.

4.1.5 Findings

There are six monitoring stations in Aberdeen measuring the NO_2 concentrations. Three of them (Market Street 2, Union Street and Wellington Road) are labelled as "at risk" as their true NO_2 annual recordings in 2012 are above the 40 μg m^{-3} regulation limit as seen in Table 4.2. Using the ADMS-Urban simulation model, 98 scenarios for different emissions, wind speed and wind direction were created to explore how these factors affect the NO_2 annual concentrations. The simulations appear to be reasonable for three monitoring stations (Anderson Drive, Market Street 2 and Union Street), whereas for Errol Place and King Street the predictions are higher than the true recordings, and for Wellington Road lower than the true recording. Lastly, variograms were used to assess the presence of spatial correlation with the LHC input space. It was found that there is evidence for hyperspatial correlation which is predominantly affected by the emissions.

4.2 Glasgow

4.2.1 Monitoring system

Glasgow is the most populous city in Scotland. However, the city faces one of the highest pollution levels in the whole of the UK [21]. To protect the health of its citizens, there is a need to monitor more closely the air pollution around Glasgow in order to be able to reduce the current concentrations. The air pollution monitoring system in Glasgow is part of the UK wide AURN network. The system has seven road active monitoring stations in the city and one station outside the city. A quick review of the stations is presented below based on the descriptions in [8], which is followed by a map with the locations of the stations in Figure 4.12. Further information and pictures of the stations are available at the Scottish Air Quality website [8].

- **Burgher Street** - the station is located four metres away from a major road in the east end of the city. Burgher Street is a roadside station. The station takes hourly measurements of NO_2 and PM_{10}.
- **Byres Road** - the station is located five metres away from a major road in the west end of the city. Byres Road is a roadside station. The station takes hourly measurements of NO_2, PM_{10} and $PM_{2.5}$.
- **Central Station** - the station is located half a meter away from a major road in the city centre, next to Glasgow Central train station. The monitor is at a transport hub and there are high buildings which create prerequisites for high air pollution. The station is a kerbside type. The station takes hourly measurements of NO_2.
- **Dumbarton Road** - the station is located 1.5 metres away from a major road in the west end of the city. Dumbarton Road is a roadside station. The station takes hourly measurements of NO_2, PM_{10} and $PM_{2.5}$.
- **Great Western Road** - the station is located close to a major road in the west end of the city, close to a subway station. Great Western Road is a roadside station. The station takes hourly measurements of NO_2.
- **High Street** - the station is located five and a half metre from a major road, next to the biggest hospital in the city. High Street is a roadside station. The station takes hourly measurements of NO_2.

- **Townhead** - the station can be accessed through a residual access road but the closest road is approximately 122 metres away. Townhead is an urban background station. The station takes hourly measurements of NO_2 and O_3.

- **Waulkmillglen Reservoir** - the station is 700 metres to the North West of the M77. Waulkmillglen Reservoir is a rural station. The station takes hourly measurements of NO_2, O_3, PM_{10} and $PM_{2.5}$

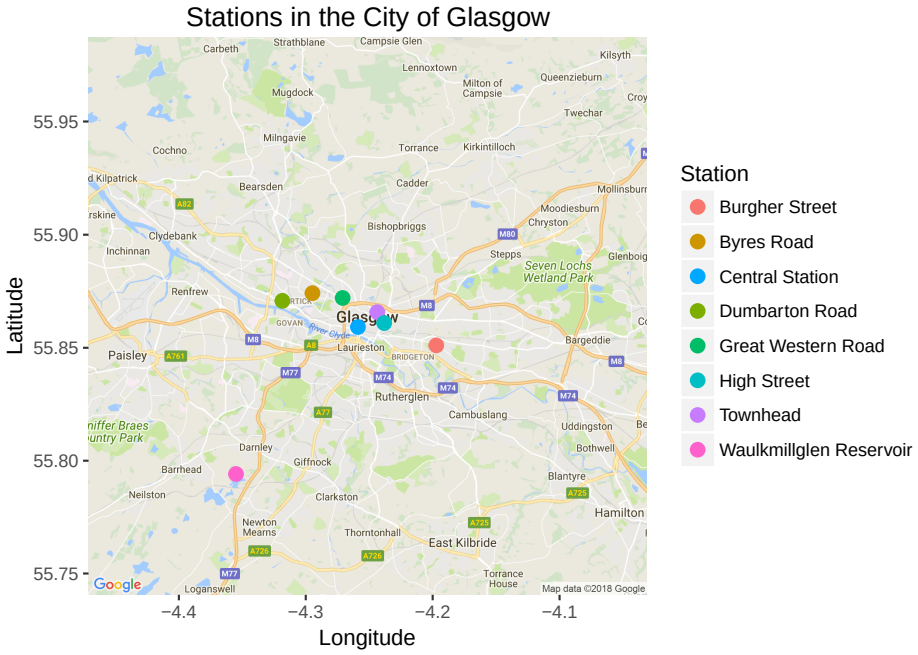

FIGURE 4.12: Map of the eight AURN monitor stations across the City of Glasgow in 2015 [92].

The pollutant concentrations for all stations are reported as rounded to the nearest whole number except for Waulkmillglen Reservoir. This specific station is used for background monitoring and, therefore, the pollutant concentrations at Waulkmillglen Reservoir are reported to a higher precision, namely one decimal point. Therefore, although the pollutant concentrations are continuous, the data are actually discretised in a similar way to the Aberdeen data.

4.2.2 NO_2 monitoring in 2015

One of the biggest problems SEPA faces when tackling the air pollution in Glasgow is the NO_2 concentrations as stated in [61]. Therefore, a variety of graphical and numerical summaries of the 2015 data are provided. Using this preliminary analysis, two major questions concerning the compliance with the EU legislation [76] are addressed:

(i) Has the hourly NO_2 concentration limit of 200 μg m^{-3} been breached more than 18 times in 2015?

(ii) Is the average annual mean for NO_2 concentration above the limit of 40 μg m^{-3}?

In order to answer question (i), Table 4.4 provides a count of the breaches of the hourly limit of 200 μg m^{-3} for the eight monitoring stations, as well as displaying the number of missing observations and the total number of observations each station has taken through 2015 in Glasgow. The maximum number of hourly observations in a non-leap year is 8760. From Table 4.4, it is clear that the hourly regulation has not been breached but at Central Station four hourly concentrations have been above 200 μg m^{-3}. A key thing to notice is that the High Street and Waulkmillglen Reservoir stations have more than 1500 missing values, which are due to long periods of time (over a month), during which the monitors were not working. To visualise the data and check for any other abnormalities, time series plots for the hourly concentrations of each of the stations are presented in Figure 4.13.

Station	**Breaches**	**Missing**	**Total**
Burgher Street	0	71	8689
Byres Road	0	222	8538
Central Station	4	94	8666
Dumbarton Road	0	246	8514
Great Western Road	0	86	8674
High Street	0	2169	6591
Townhead	0	396	8364
Waulkmillglen Reservoir	0	1688	7072

TABLE 4.4: A count of the number of breaches of the hourly concentration limit of 200 μg m^{-3} in Glasgow in 2015. The missing values and total number of observations per station for the year are also provided.

From Figure 4.13, the breaches of the 200 μg m^{-3} hourly limit are easy to see. All four events have happened at the Central Station monitor. Central Station is the station where the highest concentrations have been recorded through the year. Waulkmillglen Reservoir has recorded lower concentrations than all other stations, followed by Townhead. This is expected given that these two stations are rural and urban background locations, respectively. Furthermore, the plots for Byres Road and Dumbarton Road show visible gaps: for Byres Road (starting around 1/12) and for Dumbarton Road (starting around 1/9). Possible reasons for these visible gaps of missing data are that the monitors were under repair or the data were not transmitted. Overall, all the times series in Figure 4.13 show similar characteristics - increased pollution concentration in the winter months and a decrease in the summer months, which is expected based on the previously discussed characteristics of NO_2 in Subsection 1.2.1. In mid-January (between 15/01 and 24/01), all monitors have recorded increased pollution concentrations.

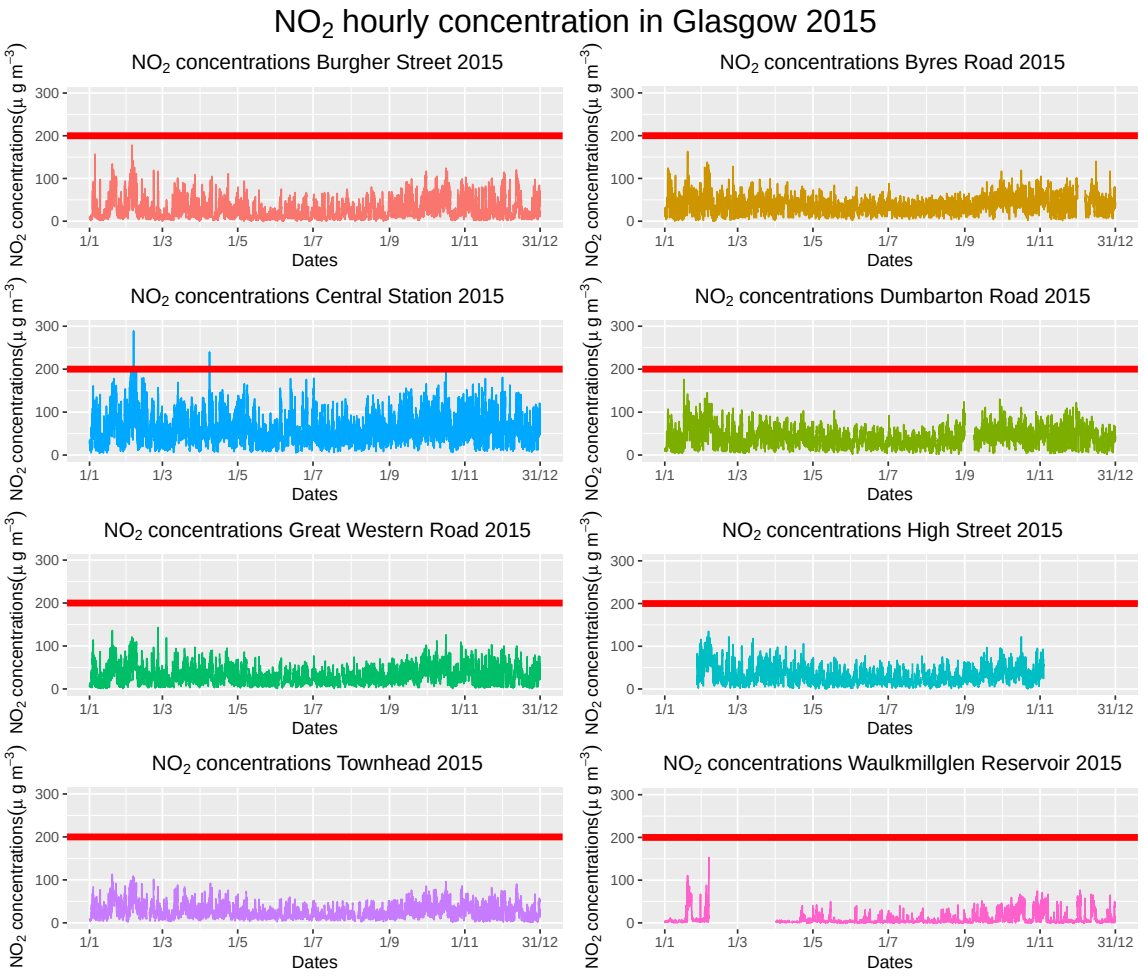

FIGURE 4.13: Time series plot for the hourly NO_2 concentrations (μg m^{-3}) for each of the eight AURN monitoring stations in Glasgow in 2015. The hourly NO_2 limit of 200 μg m^{-3} is represented by a solid red line.

A similar trend occurs in the beginning of February. These time periods need further investigation.

Another point of discussion is the fact that the hourly pollutant concentrations are skewed. The skew in the data is best seen by examining the histograms for the measurements for each of the stations in Figure 4.14. The histograms show that the distributions for the NO_2 measurements for all the stations are right skewed. Log and square root transformations were compared and the log-transformation was chosen based on the histograms presented in Figure 4.15. The distributions for all histograms appear more symmetrical. Waulkmillglen Reservoir is the only exception where the distribution continues to be right skewed which is not surprising given the small values recorded at the location.

Because there were no non-compliance issues with the hourly concentrations of NO_2, the annual means were checked in order to answer question (ii). They are presented in

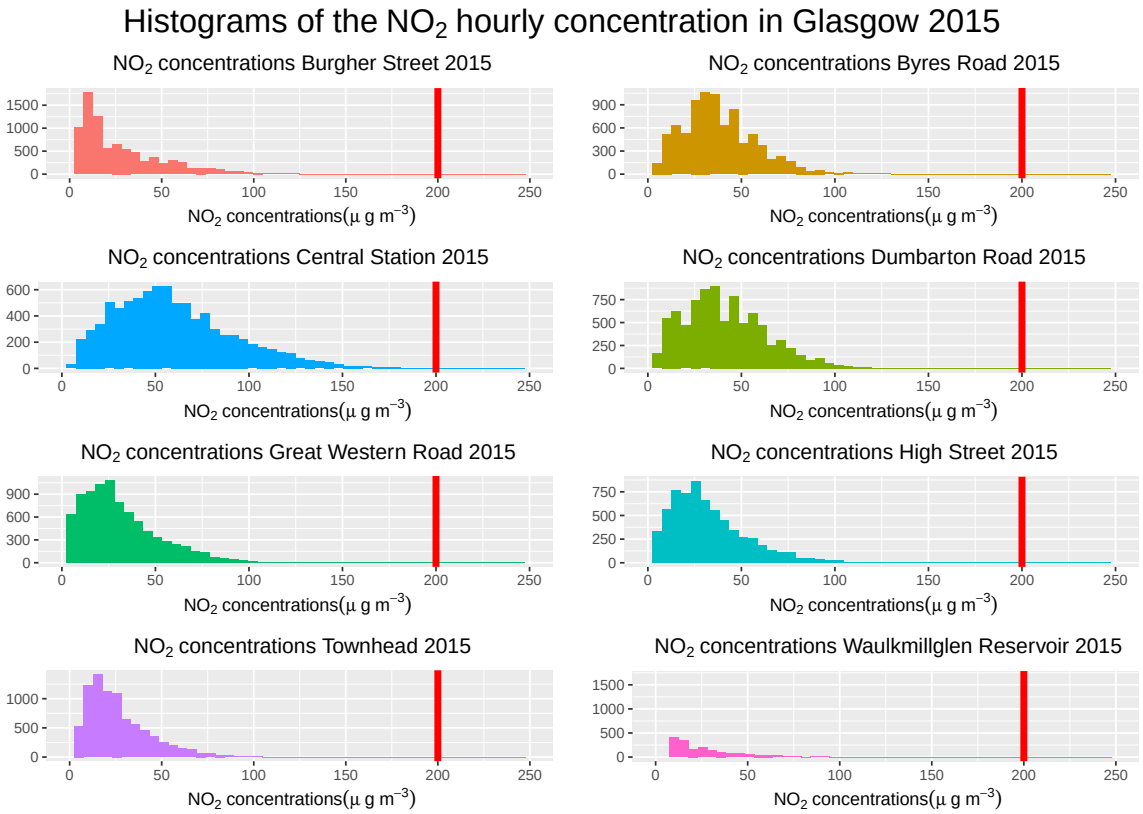

FIGURE 4.14: Histograms for the hourly NO_2 concentrations (μg m^{-3}) for each of the eight AURN monitoring stations in Glasgow in 2015. The hourly NO_2 limit of 200 μg m^{-3} is represented by a solid red line.

Table 4.5. The values above the regulation have been highlighted in red. The whole data set is not available as there are missing observations for all the stations. Therefore, confidence intervals for the mean would be useful to provide a range of possible values for the true mean. A standard 95% confidence interval is provided. However, that standard confidence interval does not take into account the correlation between the observations, so a 95% bootstrap interval and an adjusted for correlation are also provided (estimated as described in Subsection 2.1.4).

Before calculating the 95% correlation adjusted confidence interval, the PACF plots for the time series of each of the stations must be examined to determine the type of AR process of the data. The PACF plots (calculated as described in Subsection 2.1.2) for all stations are presented in Figure 4.16. The PACF plots, similarly to Aberdeen, show that at all stations, for the first lag the observations have very strong positive autocorrelation. Overall, the autocorrelations for three or four lags are significant (above the standard error bars). Similarly to Aberdeen, there is a possible diurnal seasonality present as the concentrations around the 20$^{\text{th}}$ hour appear significant. An interesting feature is that around the 500$^{\text{th}}$ lag for Waulkmillglen Reservoir appears highly significant but this is the result of the fact that this station measures background emissions, there is much less

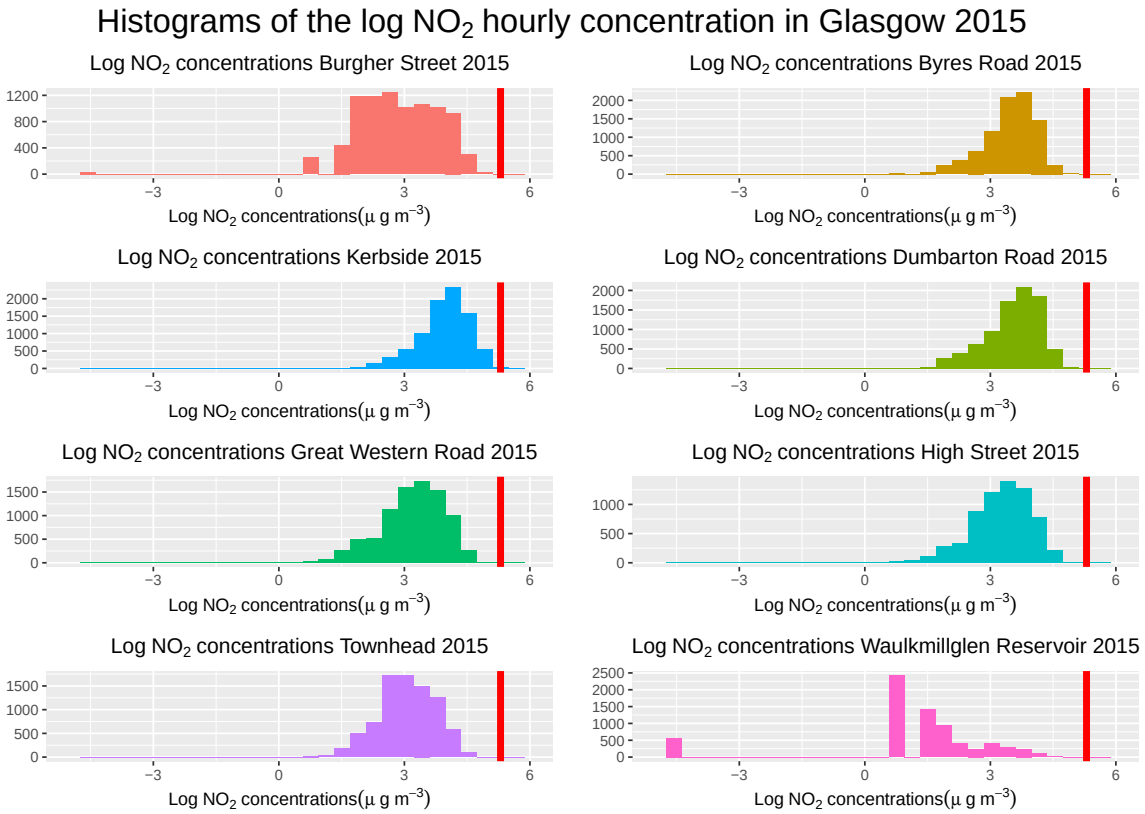

FIGURE 4.15: Histograms for the hourly log NO_2 concentrations (μg m^{-3}) for each of the eight AURN monitoring stations in Glasgow in 2015. The hourly log NO_2 limit of log(200) μg m^{-3} is represented by a solid red line.

variability in the observed concentrations. For all stations, except for Central Station, the second lag is negative as in the Aberdeen case which is expected when there is a very strong positive autocorrelation at the first lag. However, due to the relative nature of interpretation of the error bands for large time series, an AR(1) process is a reasonable simplification for the estimation of the 95% correlation adjusted confidence intervals.

Station	Annual Mean	St. interval	Bootstrap	Corr. adjusted
Burgher Street	26.67	(26.19, 27.15)	(26.19, 27.17)	(24.49, 28.84)
Byres Road	37.82	(37.40, 38.23)	(37.41, 38.23)	(36.07, 39.57)
Central Station	**60.39**	**(59.71, 61.07)**	**(59.73, 61.06)**	**(58.13, 62.66)**
Dumbarton Road	**41.43**	**(40.96, 41.90)**	**(40.99, 41.88)**	**(39.34, 43.53)**
Great Western Road	31.15	(30.72, 31.58)	(30.75, 31.57)	(29.39, 32.91)
High Street	32.39	(31.96, 32.82)	(31.95, 32.88)	(30.45, 34.33)
Townhead	26.21	(25.85, 26.56)	(25.84, 26.55)	(24.56, 27.85)
Waulkmillglen Reservoir	8.62	(8.34, 8.90)	(8.32, 8.96)	(7.22, 10.02)

TABLE 4.5: Comparing the annual mean and the three different types of 95% intervals (standard confidence interval, bootstrap and correlation adjusted confidence) for the hourly NO_2 concentrations (μg m^{-3}) across the eight AURN monitoring stations in Glasgow in 2015.

Table 4.5 presents three different types of intervals for the annual mean. The standard

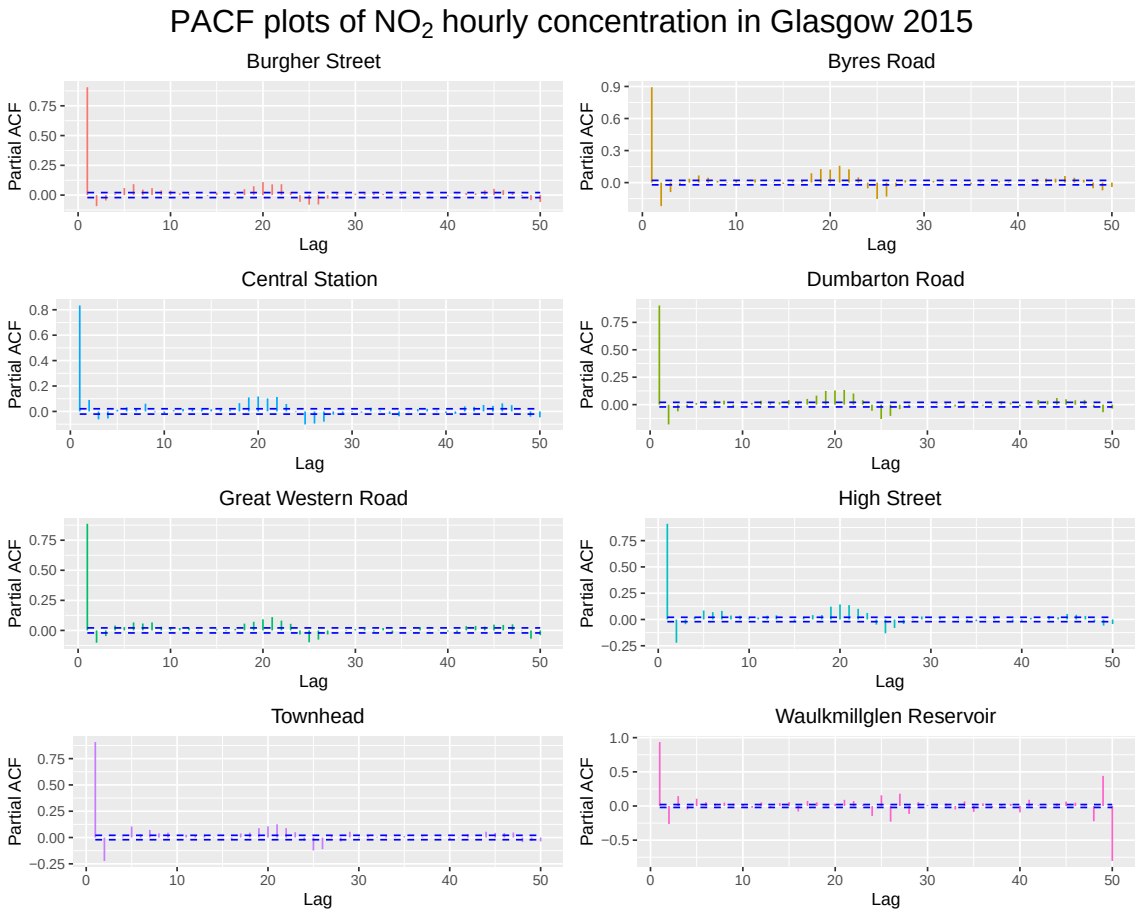

FIGURE 4.16: PACF plots for the time series of the hourly NO_2 concentrations (μg m^{-3}) for each of the eight monitoring stations in Glasgow in 2015 up to lag 50.

confidence interval is too narrow because it does not adjust for the correlation in the data. The bootstrap intervals are almost identical to the standard confidence interval. However, the adjusted for correlation 95% confidence interval is wider than the other two which makes it the most realistic interval of the three presented. These findings are in agreement with the Aberdeen calculations in Table 4.2. From Table 4.5, it is obvious that in 2015 two stations (Central Station and Dumbarton Road) have failed to comply with the EU regulation. For both stations, the recorded NO_2 annual average concentration is above the limit of 40 μg m^{-3}. Central Station has an annual average which is 50% higher than the regulation. For Dumbarton Road, the mean is just above the regulation. However, the correlation adjusted 95% confidence interval suggests that it is possible for the true mean to be below 40 μg m^{-3} as the interval includes values under 40. As the data from the monitoring stations are right-skewed (as seen in Figure 4.14), which could result in misleading mean values, the median values for concentrations at each station are also calculated in Table 4.6.

Table 4.6 presents the three different intervals. The quantile interval is too wide and

Station	Median	Quantile	Bootstrap	Conf. interval
Burgher Street	19.00	(2.00, 86.00)	(17.00, 19.00)	(17.00, 19.00)
Byres Road	36.00	(8.00, 82.00)	(34.00, 36.00)	(34.00, 36.00)
Central Station	**55.00**	**(12.00, 136.00)**	**(54.00, 56.00)**	**54.00, 56.00)**
Dumbarton Road	38.00	**(8.00, 94.00)**	**(38.00, 40.00)**	**(38.00, 40.00)**
Great Western Road	27.00	(4.00, 82.00)	(28.00, 29.00)	(28.00, 29.00)
High Street	28.00	(5.00, 84.00)	(28.00, 29.00)	(28.00, 29.00)
Townhead	22.00	(5.00, 69.00)	(21.00, 22.00)	(21.00, 22.00)
Waulkmillglen Reservoir	3.80	(0.00, 51.60)	3.80, 3.80)	(3.80, 3.80)

TABLE 4.6: Comparing the annual median and the three different types of 95% intervals (quantile, bootstrap and confidence interval) for the hourly NO_2 concentrations (μg m^{-3}) across the eight monitoring stations in Glasgow in 2015.

provides limited information about the median. The bootstrap and the confidence intervals are in agreement with each other. Furthermore, for both the bootstrap and the confidence intervals, in most cases, the median is the same as one end of the intervals. This is because all measurements were rounded to the nearest integer and, hence, there is a lot of repetition in the observed values for each station. These results are also in agreement with what was observed for the Aberdeen data in Table 4.3.

Overall, the median values for all stations (Table 4.6) are always lower than the means (Table 4.5) but not very different from the mean values except for the Waulkmillglen Reservoir where the median value is 2.25 times lower than the mean. For Central Station, the median is lower than the mean but 55 μg m^{-3} is still 38% larger compared to the regulation. It is important to note that the median value for Dumbarton Road is below the 40 μg m^{-3} regulation, whereas the mean value was above the regulation. Based on the results from Tables 4.5 and 4.6, two stations (Central Station and Dumbarton Road) are identified as "at risk".

4.2.3 ADMS-Urban simulations

For a better understanding of the NO_2 hourly concentrations across Glasgow, the ADMS-Urban simulation model (introduced in Subsection 4.1.3) was used to produce 100 simulation scenarios for each one of the eight monitoring stations. The scenarios were produced by varying only three inputs in the model - wind speed, wind direction and emissions. As with Aberdeen, a LHC design was used to vary these three inputs. For the **emissions**, the values were varied from -100% to $+20\%$, **wind speed** was varied from -20% to $+20\%$, and **wind direction** was varied from $-15°$ to $15°$ in comparison to the observed baseline values in 2015. The input boundaries were chosen by SEPA experts as values outside these regions are highly unlikely to occur. The variations for

the three inputs were chosen using the LHC, making them an optimal random sample of combinations across the sampled space given in Figure 4.17.

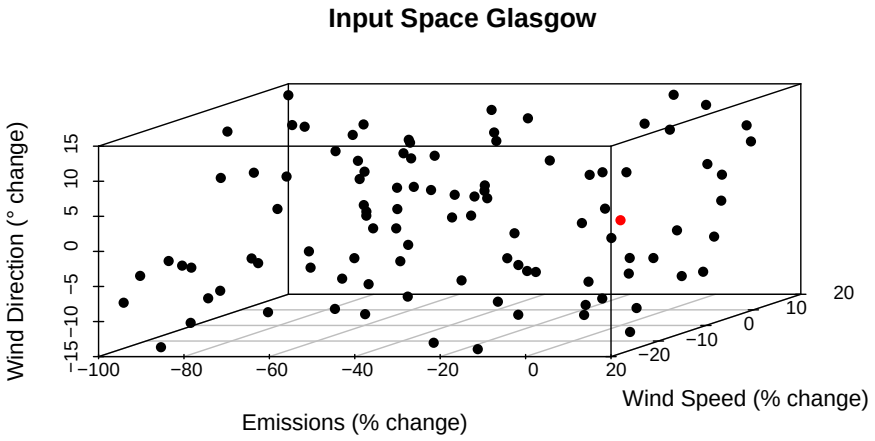

FIGURE 4.17: Input space for the one hundred simulations of the year long time series of the hourly NO_2 concentrations (μg m^{-3}) in Glasgow. The point $(0, 0, 0)$ representing the true values for emissions (% change), wind speed (% change) and wind direction (° change) is in red.

The main difference between the ADMS-Urban simulations for Aberdeen and Glasgow is that for Glasgow, the hourly concentrations for a full year are available so both the annual average and hourly regulations will be examined. It is important to note that once the simulation scenarios were produced, for each of them there were 35 missing hourly NO_2 concentrations at the same place for each of the scenarios and each of the stations. The missing data in the simulation scenarios were a result of missing input data, which is not available to examine in this thesis. 35 missing values are only 0.4% of the whole simulation, which is almost a negligible amount of missing data so the missing data were imputed without having a visible impact on the exploratory analysis. When there is only one NO_2 hourly concentration missing, the average of the neighbouring NO_2 hourly observations was taken and assigned to the missing observation. When a series of missing NO_2 hourly concentration was present, the average of the hourly NO_2 concentrations for the same hour of two neighbouring days was taken and assigned to the missing observation.

Firstly, the NO_2 hourly concentrations at each station are examined by plotting the boxplot for every 50^{th} hour in the time series for each station in Figures 4.18 and 4.19. The 50^{th} hours were chosen as they show the largest variability in hourly concentrations. In colour, the actual hourly NO_2 concentration for this specific station is superimposed. It is expected that the actual hourly NO_2 concentrations will lie within the boxplots. Given that the design is not centred around the $(0, 0, 0)$ point as seen in Figures 4.18 and 4.19, the actual concentrations are expected to lie above the hourly respective median. However, in Figures 4.18 and 4.19 it appears that the ADMS-Urban hourly simulated NO_2 concentrations are not similar to the actual NO_2 concentrations as the points for the

actual concentrations are not lying within the range of the boxplot or even the outliers but rather above or below all simulated values for 42% of the plotted hours across all stations. Additionally, the actual concentration points do not differ by a constant offset to the median of the simulated values. For all stations except for Burgher Street, the highest simulated values are higher than the observed ones. Overall, it appears that the ADMS-Urban model has struggled with accurately simulating the hourly concentrations.

Looking at the plots for each individual station in Figures 4.18 and 4.19, there are a few points of interest. The plot for Burgher Street is the one where the majority of points (60%) are outside each of the boxes for the simulations for each specific hour highlighting the fact that the ADMS-Urban model has struggled most with the predictions for Burgher Street in comparison with the other stations. Furthermore, the highest actually recorded hourly NO_2 concentration is higher than the highest simulated values. For Central Station, the majority of the actual concentrations are within the boxplots for the simulated values (68%) which is likely an effect from the fact that the boxplot spreads are larger than the spreads for any other station. As Central Station was identified as a "at risk" location, it is good that the ADMS-Urban simulations show scenarios where the hourly 200 μg m^{-3} regulation is breached. The simulations for Dumbarton Road are in contrast to those at Burgher Street - the actual NO_2 concentrations are not as high in comparison to Burgher Street, but the simulations are. The ADMS-Urban simulations for Dumbarton Road indicate the station is an "at risk" location where the hourly 200 μg m^{-3} regulation could be breached with simulations above 200 μg m^{-3}. Great Western Road is the only station where the actual highest NO_2 concentration is the same hour where the simulations have the highest values. For High Street and Waulkmillglen Reservoir, there are a lot of missing values from the actual hourly NO_2 concentrations to compare the performance of the simulations to the actual values. Even though the highest High Street simulation is higher than the actual recorded values, the actual recordings appear above the boxplots in 31% of the cases, whereas for Waulkmillglen Reservoir half the points are outside the boxplots. Overall, for all stations the actual points are more often above (38%) than below (5%) the boxplots.

After comparing the hourly NO_2 ADMS-Urban simulations to the actual hourly NO_2 concentrations for each of the eight monitoring stations in Glasgow, the annual averages are also compared to the boxplots for the annual average simulations in Figure 4.20. As with the hourly concentrations, it is expected that the actual annual average is higher than the majority of the simulated ones due to the design of the sample space where the baseline conditions point $(0, 0, 0)$ is not the centre.

In Figure 4.20, for all stations except for Burgher Street, there is at least one simulation model where the NO_2 annual average is higher than the observed one. For Burgher

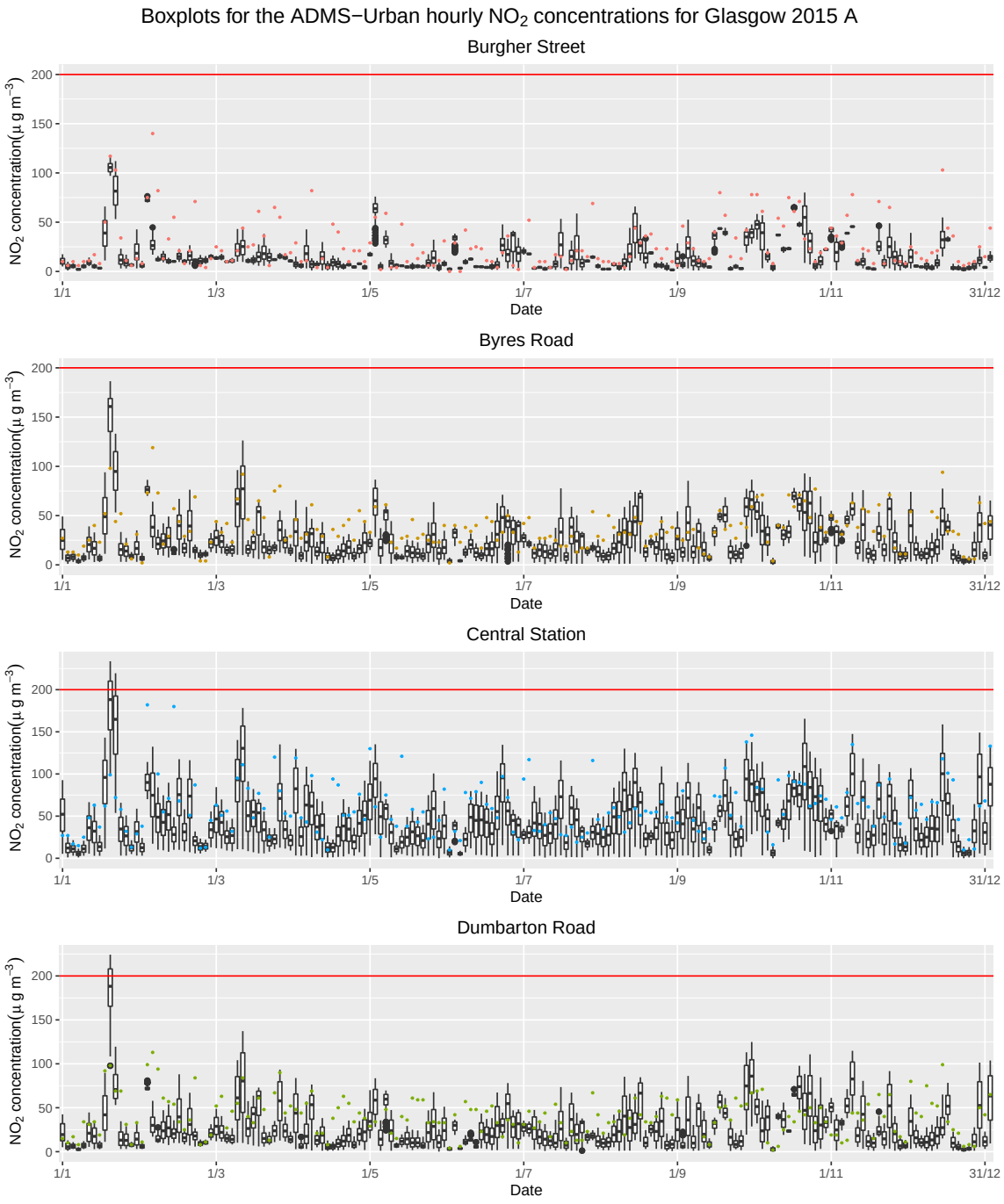

FIGURE 4.18: Boxplots for the NO_2 hourly concentration (μg m^{-3}) for every 50$^{\text{th}}$ hour from ADMS-Urban for Burgher Street, Byres Road, Central Station and Dumbarton Road. The true NO_2 concentrations for the corresponding hour for each station are the coloured points on top of the boxplots. The hourly NO_2 limit of 200 μg m^{-3} is represented by a red line.

Street, all the ADMS-Urban simulations are smaller than the actually observed annual average. For Byres Road, the majority of the annual averages (97%) from the simulations are lower than the actual annual average, indicating that ADMS-Urban is struggling with simulating the annual averages for Byres Road. Central Station has been previously identified as an "at risk" location where the regulations are very likely to be broken with

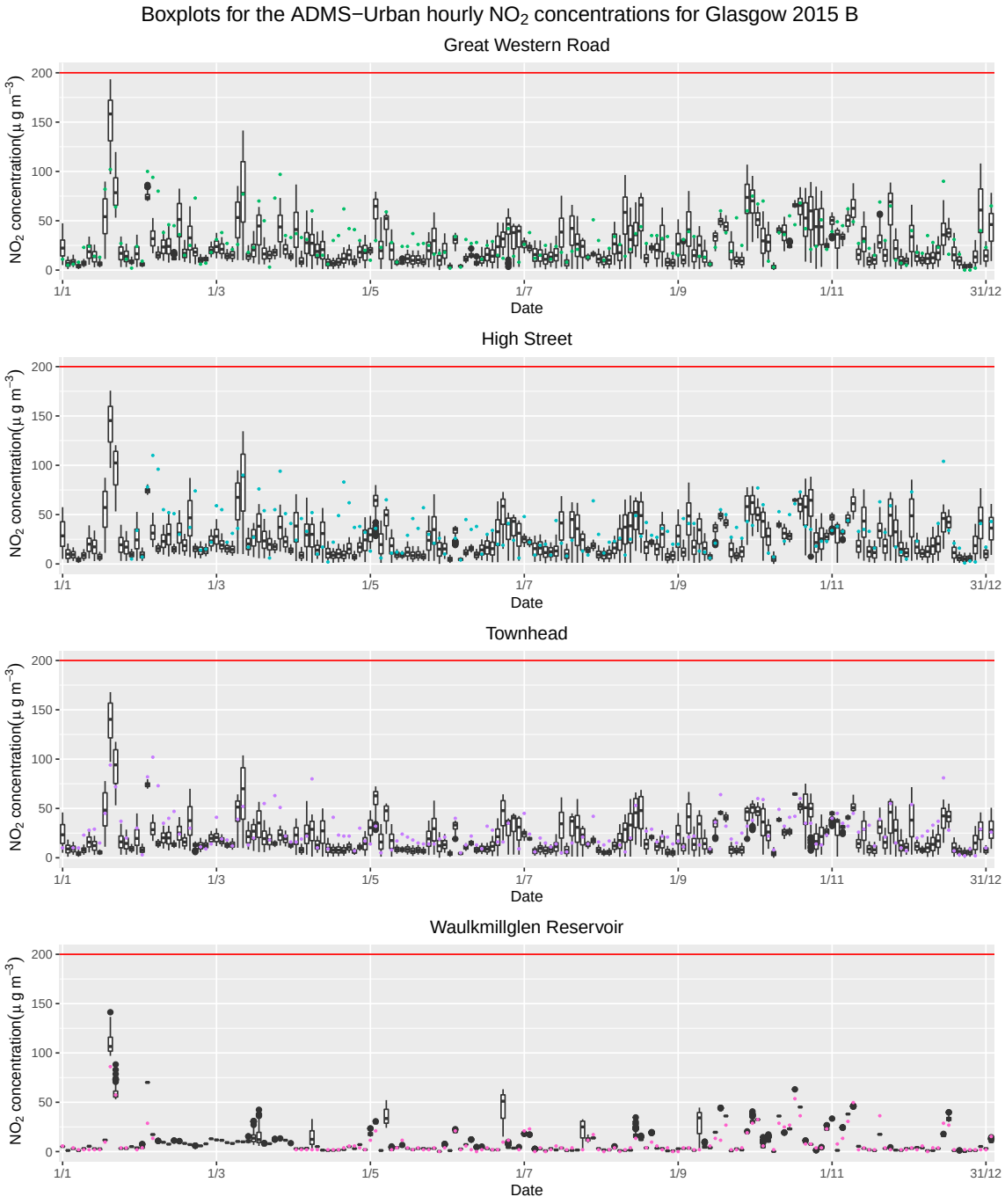

FIGURE 4.19: Boxplots for the NO_2 hourly concentration (μg m^{-3}) for every 48$^{\text{th}}$ hour from ADMS-Urban for Great Western Road, High Street, Townhead and Waulkmillglen Reservoir. The true NO_2 concentrations for the corresponding hour for each station are the coloured points on top of the boxplots. The hourly NO_2 limit of 200 μg m^{-3} is represented by a red line.

the majority of the interquartile range is above the 40 μg m^{-3} line. The large spread of the Central Station box indicates that for some scenarios it is possible for compliance with the regulation. For Dumbarton Road, there are several simulations above the 40 μg m^{-3} regulation (8%). For Great Western Road, the actual annual average is below the limit and so are the simulated annual averages, whereas for High Street, there are a few

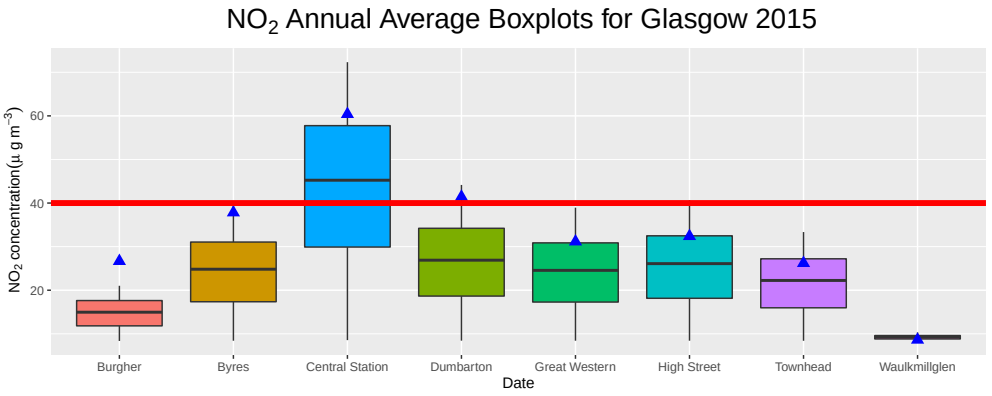

FIGURE 4.20: Boxplots for the NO_2 annual averages (μg m^{-3}) from one hundred simulations of ADMS-Urban for each of the eight monitoring stations in Glasgow. The yearly NO_2 limit of 40 μg m^{-3} is represented by a red line. The true annual average for 2015 for each station is the blue triangle.

simulations (3%) above the 40 μg m^{-3} limit. For Townhead, the annual averages from both the actual values and the simulations are visibly below the 40 μg m^{-3} limit. The annual averages for Waulkmillglen Reservoir appear quite similar and the actual annual average is in the middle of the simulated values. Overall, Figure 4.20 suggests that the simulated annual concentrations of the ADMS-Urban model appear more accurate than the hourly time series for every monitoring station in Glasgow, which is expected given that there is more variation at the hourly level. Overall, all boxplots for the simulated NO_2 annual averages are approximately symmetrical and there are no outliers suggesting that the NO_2 simulated annual average for each station are normally distributed.

4.2.4 Variograms

The ADMS-Urban simulation scenarios for Glasgow are chosen based on a 3-dimensional LHC created by varying **emissions** (from -100% to $+20\%$), **wind speed** (from -20% to $+20\%$) and **wind direction** (from -15° to $+15^\circ$) to the baseline observed values in 2015. As with Aberdeen, variogram plots are used to explore the presence of spatial correlation between the LHC points used to simulate the NO_2 annual averages for each of the monitoring stations in Glasgow.

To begin, the simulated NO_2 annual average variograms residuals (from an intercept only model) for each of the monitoring stations in Glasgow based on emissions only are examined in Figure 4.21. Similarly to the Aberdeen variograms for emissions only in Figure 4.8, although it is expected that the points will reach a peak and plateau afterwards, the points on all variograms are constantly increasing due to the design of the LHC space to have spatial correlation between all scenarios. The sill and the range parameters appear to be going to infinity in the observed space. None of the variograms

have a nugget effect indicating the lack of measurement error. As Central Station is the monitoring station with the largest recorded NO_2 concentrations, it has the largest semivariance in comparison to all other stations.

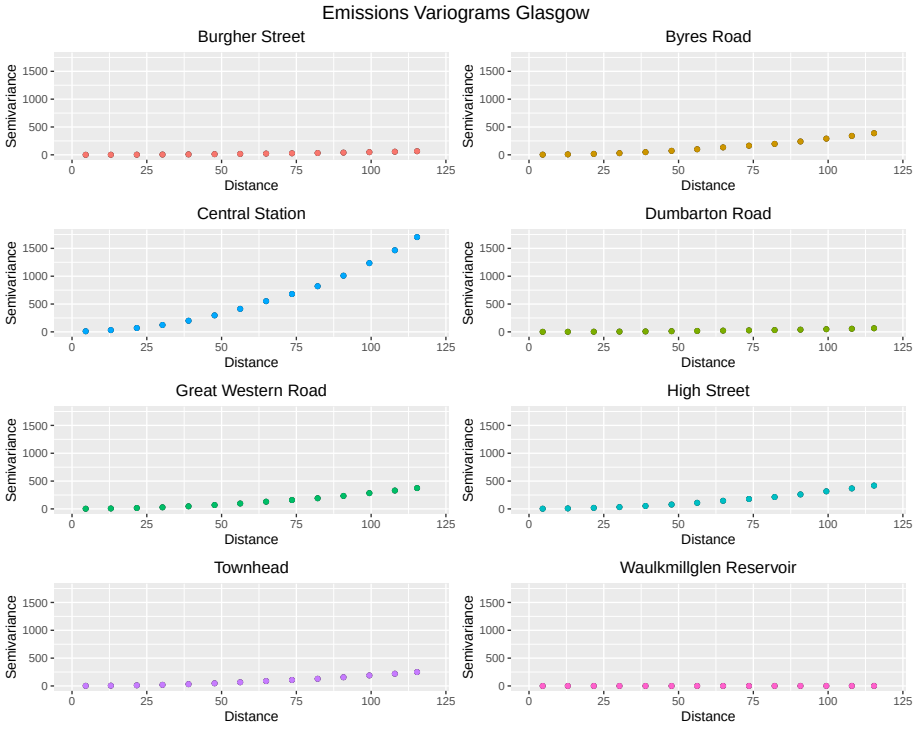

FIGURE 4.21: Emission variograms for the eight monitoring stations in Glasgow.

Following, the simulated NO_2 annual average residuals (from an intercept only model) variograms for each of the monitoring stations in Glasgow based only on the wind speed input are produced in Figure 4.22. The variograms show that there is spatial correlation present with almost all points plateauing quite evenly until the distance of 30, when the points have a spike followed by a drop. This suggests that the hyperspatial range parameter for wind speed might be easier to estimate than the emissions one. However, as opposed to the emissions variogram in Figure 4.21, there is a nugget effect for all stations except for Burgher Street and Waulkmillglen Reservoir indicating that there might be measurement error in the wind speed. Overall, the results for Glasgow are similar to the wind speed Aberdeen variograms in Figure 4.9.

Next, the simulated NO_2 annual average residuals (from an intercept only model) variograms for each of the monitoring stations in Glasgow based only on the wind direction input are presented in Figure 4.23. For all stations, the points appear constant throughout. This indicates that the wind direction hyperspatial range parameter will not be easy to estimate. As with the wind speed variograms in Figure 4.22, there is a nugget effect for all stations except for Burgher Street and Waulkmillglen Reservoir indicating

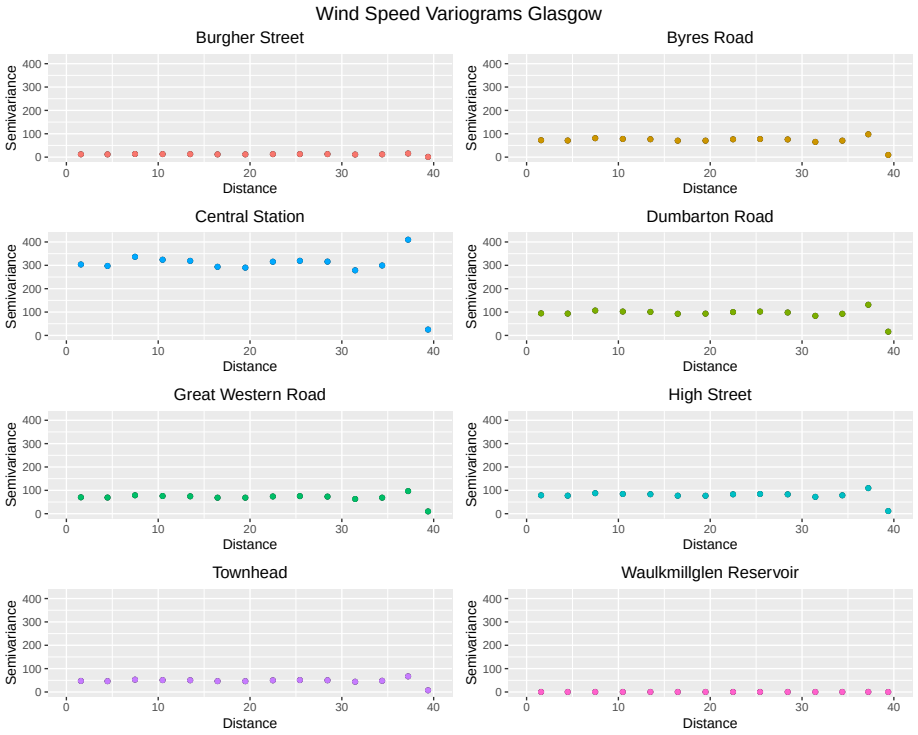

FIGURE 4.22: Wind speed variograms for the eight monitoring stations in Glasgow.

that there might be a small measurement error in the wind direction. It is interesting to note that the results are not similar to those from the wind direction Aberdeen variograms in Figure 4.10 indicating a difference between the two designs.

FIGURE 4.23: Wind direction variograms for the eight monitoring stations in Glasgow.

Lastly, 3D variograms for the simulated NO_2 annual average residuals (from an intercept only model) for each of the monitoring stations in Glasgow are shown in Figure 4.24. The variograms are very similar to the emissions variograms in Figure 4.21 suggesting that the emissions input has the largest impact on the correlation between the ADMS-Urban scenarios. Overall, it appears that there is spatial correlation between the scenarios which would assist the predictions in the LHC space. As with the Aberdeen 3D variograms in Figure 4.11, although the wind speed and wind direction variograms showed nugget effects, the 3D variograms for Glasgow do not contain any nuggets.

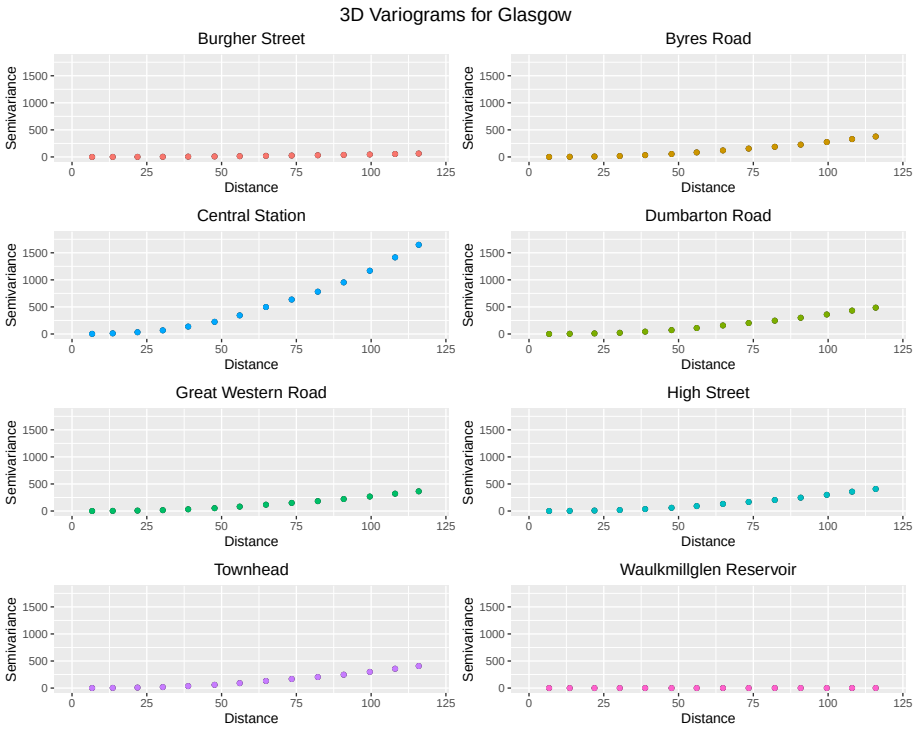

FIGURE 4.24: 3D variograms for the eight monitoring stations in Glasgow.

Additionally, variograms for each hour in the year were examined to check whether there is spatial correlation between the hourly concentrations from the ADMS-Urban scenarios. The variograms are very similar to the ones for the annual averages and therefore, omitted to prevent repetition. Overall, it appears that the NO_2 hourly concentration variograms have spatial correlation which would aid forecasting of hourly time series for the NO_2 concentrations.

4.2.5 Findings

Eight monitoring stations were used to measure NO_2 concentrations in Glasgow in 2015. Only Central Station and Dumbarton was labelled as "at risk" monitoring stations their its true NO_2 annual recording in 2015 is above the 40 μg m^{-3} as shown in Table 4.5. The ADMS-Urban model was used to create 100 simulation scenarios for different emissions,

wind speed and wind direction in order to explore how these factors influence the NO_2 annual concentrations. For all stations except for Waulkmillglen Reservoir, the true annual averages are larger than the majority of the simulated values. In terms of NO_2 hourly concentrations, the highest simulated values tend to be higher than those actually observed but overall, 43% of the simulated hourly values are either larger or smaller than the true NO_2 recordings in 2015. These results are expected given the ranges chosen for the varying the inputs of the LHC design. Using variograms, it was assessed that there is spatial correlation within the LHC input space and it was found that the hyperspatial correlation is mostly affected by the emissions.

4.3 Conclusion

In this chapter, the exploratory analysis for the Aberdeen and Glasgow data sets used in the rest of the thesis was performed. For Aberdeen, three monitoring stations (Market Street 2, Union Street and Wellington Road) were identified as "at risk" based on breaching the NO_2 annual regulation limit of 40 μg m^{-3}, whereas for Glasgow there are two such station (Central Station and Dumbarton Road). In order to explore the effects of emissions, wind speed and wind direction, the ADMS-Urban simulation model was used to create 98 and 100 simulation scenarios for Aberdeen and Glasgow, respectively. It was found that there is evidence of spatial correlation between the scenarios which would aid prediction in the LHC sample space. Based on the hourly regulations, although at some stations in both Aberdeen and Glasgow, there were hourly concentrations above 200 μg m^{-3}, no breaches were observed in either city.

Chapter 5

Univariate modelling of the NO_2 annual average using Gaussian Processes

In this chapter, the main focus is on creating a statistical model for the ADMS-Urban model for predicting the NO_2 annual average concentrations across the monitoring stations in Aberdeen and Glasgow based on the simulation runs from ADMS-Urban presented in Chapter 4. By fitting a variety of statistical models, it is aimed to find the best prediction model for the simulations and thus allow quicker evaluation of ADMS-Urban under different emissions and meteorological scenarios. The prediction power of the models will be measured using in- and out-of-sample Root Mean Squared Prediction Error (RMSPE). The out-of-sample RMSPE is calculated using a 10-fold Cross Validation (CV) procedure which splits the data into training and test sets as described in Chapter 2. This chapter will be organised as follows: Section 5.1 provides the theoretical background on Gaussian Processes (GP) which have been fitted in R using the **DiceKriging** package, then Section 5.2 presents a motivation study of modelling the NO_2 annual average concentrations in Aberdeen for six monitoring stations. The work in this section is based on [82], but here a short comparative study for an alternative way of fitting GP models is presented. Section 5.3 presents a new study using linear regression and GP models fitted for the annual averages for the 100 simulations for each station in Glasgow individually. Section 5.4 provides a concluding discussion.

5.1 Theoretical background on the modelling

In this chapter, the linear regression modelling will be done using the method described in Subsection 2.2.1. Then, correlated errors will be introduced in the modelling using GP models as suggested in the case of deterministic computer codes by [169]. The GP models are fitted to model the simulated ADMS-Urban NO_2 annual average concentrations for each station individually using the **DiceKriging** package [165] in R [157]. GP are used because they take into account the fact that there is correlation between the points in the Latin Hypercube (LHC) space. Furthermore, GP processes fit a zero variance for the points at which simulations are run as they are points for which the estimate is known.

The general model for each station is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\,, \tag{5.1}$$

where:

- $\mathbf{y} = [y_1, \ldots, y_n]^\top$ is a vector $(n \times 1)$ containing the response values, i.e. the simulated NO_2 annual average for one station with n being the number of simulations;
- $\mathbf{X}$ is a matrix $(n \times p)$ which is the design matrix for an intercept and $p-1$ fixed effect parameters. $\mathbf{X}$ contains the three input (emissions, wind speed, wind direction) variables as well as combinations of the inputs such as squared terms, interactions, etc. For $i = 1, \ldots, 100$, each row of $\mathbf{X}$ is $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip-1})$ which contains an intercept term $x_{i0} = 1$, and specific values for each variable x_{ij} where $j = 1, \ldots, p-1$;
- $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^\top$ is a vector $(p \times 1)$ which contains the fixed effect parameter estimates; and
- $\mathbf{z} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a vector $(n \times 1)$, which has mean zero $(n \times 1)$ and variance-covariance matrix $\boldsymbol{\Sigma}$ $(n \times n)$. $\boldsymbol{\Sigma}$ is built using different covariance models (which will be used interchangeably with kernels for the rest of this chapter).

$\boldsymbol{\Sigma}$ is built on stationary kernels, which only depend on the distance (h) between two points $\mathbf{x}_i$ and $\mathbf{x}_j$. For higher than 1-dimensional (1-d) input space, the tensor products of 1-d kernels are used. For this example, the covariance has general form:

$$\Sigma(h) = \sigma^2 \prod_{k=1}^{d} g(h_k, \boldsymbol{\theta}_k) + \lambda^2\,, \tag{5.2}$$

where

- σ^2 is the partial sill;
- $g(\cdot)$ is a function of distance; and
- λ^2 is the nugget effect.

From the **DiceKriging** package, four simplified versions of the Matérn function (Subsection 2.3.1) will be compared as recommended by [63] - the Matérn with parameter $\nu = \frac{1}{2}$ called the **exponential**, the Matérn with parameter $\nu = \frac{3}{2}$, the Matérn with parameter $\nu = \frac{5}{2}$ and the Matérn with parameter $\nu = \infty$, commonly referred to as the **Gaussian**. The four kernels are presented below for the 1-d case:

- **Matérn with $\boldsymbol{\nu}$=$\frac{\mathbf{1}}{\mathbf{2}}$** is $g(h,\theta) = \exp\left(-\frac{|h|}{\theta}\right)$, which gives:
  $\text{Cov}(z_i, z_j) = \sigma^2 \prod_{k=1}^{p} \left[\exp\left(-\frac{||x_{ik}-x_{jk}||}{\theta_k}\right)\right] + \lambda^2$;
- **Matérn with $\boldsymbol{\nu}$=$\frac{\mathbf{3}}{\mathbf{2}}$** is $g(h,\theta) = \left(1+\frac{\sqrt{3}|h|}{\theta}\right)\exp\left(-\frac{\sqrt{3}|h|}{\theta}\right)$, from where:
  $\text{Cov}(z_i, z_j) = \sigma^2 \prod_{k=1}^{p} \left[\left(1+\frac{\sqrt{3}||x_{ik}-x_{jk}||}{\theta_k}\right)\exp\left(-\frac{\sqrt{3}||x_{ik}-x_{jk}||}{\theta_k}\right)\right] + \lambda^2$;
- **Matérn with $\boldsymbol{\nu}$=$\frac{\mathbf{5}}{\mathbf{2}}$** is $g(h,\theta) = \left(1+\frac{\sqrt{5}|h|}{\theta}+\frac{5h^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}|h|}{\theta}\right)$. Hence:
  $\text{Cov}(z_i, z_j) = \sigma^2 \prod_{k=1}^{p} \left[\left(1+\frac{\sqrt{5}||x_{ik}-x_{jk}||}{\theta_k}+\frac{5||x_{ik}-x_{jk}||^2}{3\theta_k^2}\right)\exp\left(-\frac{\sqrt{5}||x_{ik}-x_{jk}||}{\theta_k}\right)\right] + \lambda^2$;
  and
- **Matérn with $\boldsymbol{\nu}$=$\boldsymbol{\infty}$** is $g(h,\theta) = \exp\left(-\frac{h^2}{2\theta^2}\right)$ and therefore:
  $\text{Cov}(z_i, z_j) = \sigma^2 \prod_{k=1}^{p} \left[\exp\left(-\frac{||x_{ik}-x_{jk}||^2}{2\theta_k^2}\right)\right] + \lambda^2$.

In the multidimensional case, the hyperspatial range parameter has the form $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]^\top$ where d is the number of dimensions. The response vector $\mathbf{y}$ is assumed to be normally distributed with mean $\mathbf{X}^\top\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$. The model is estimated by evaluating the likelihood:

$$\text{L}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \lambda^2; \mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right). \tag{5.3}$$

In this case, the covariance matrix $\boldsymbol{\Sigma}$ is redefined through the correlation matrix $\mathbf{R}$ $(n \times n)$ such that $\boldsymbol{\Sigma} = \sigma^2\mathbf{R} + \lambda^2\mathbb{I}_n$. Let $v = \sigma^2 + \lambda^2$ be the total variance in the data and $\alpha = \frac{\sigma^2}{\sigma^2+\lambda^2}$ is the proportion of variance explained in $\mathbf{z}$. Hence, $\mathbf{R}_\alpha = \alpha\mathbf{R} + (1-\alpha)\mathbb{I}_n$ can be used to redefine the covariance matrix $\boldsymbol{\Sigma} = v\mathbf{R}_\alpha$. Therefore, closed form solutions for $\boldsymbol{\beta}$ and v are expressed using $\boldsymbol{\theta}$ and λ^2:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{R}_\alpha^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{R}_\alpha^{-1} \mathbf{y}\,, \tag{5.4}$$

$$\widehat{v} = \frac{1}{n} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top \mathbf{R}_\alpha^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,. \tag{5.5}$$

However, $\boldsymbol{\theta}$ and λ^2 cannot be found explicitly and a numerical method should be applied. In the **DiceKriging** pacage, the BFGS algorithm is used (Subsection 2.2.4). Using this method of fitting GP models, the four different kernels will be compared using the in- and out-of-sample RMSPE by creating a model for every station individually. Since **DiceKriging** does not provide a standard error estimate for the parameters, the GLS estimate for the parameters' standard error is used such that $\text{se}(\widehat{\boldsymbol{\beta}}) = \text{diag}\left(\sqrt{\sigma^2 \left(\mathbf{X}^\top \mathbf{R}_\alpha^{-1} \mathbf{X}\right)^{-1}}\right)$ [78].

5.2 Modelling the individual stations in Aberdeen

In this section, different modelling techniques for the 2012 Aberdeen NO_2 annual average as simulated with ADMS-Urban are presented. The Anderson Drive monitoring station was chosen to be presented in full due to its location being further away from the other five stations, the models for which are summarised to avoid repetition.

5.2.1 Linear regression modelling of the Aberdeen data

The simplest model used is a linear regression (Subsection 2.2.1) with three covariates - the three inputs (emissions, wind speed and wind direction) used to create the LHC design. The distributions of the annual NO_2 averages for each station were checked using the boxplots in Figure 4.7 and it was found that for each station, the boxplot appears symmetric and therefore, normality can be assumed. Therefore, a linear regression model is fitted for each of the six monitoring stations individually. The aim is to find the simplest model explaining the relationship between the covariates and the NO_2 annual average at every station. For the modelling of all station, the wind direction change is only 30°, which is a small segment of a circle and can be treated as linear.

Anderson Drive

Before modelling is performed, a pairs plot (Figure 5.1) is used to check the relationships between the response (the NO_2 annual average as simulated by ADMS-Urban for Anderson Drive) and the covariates (the percentage change in emissions and wind speed,

and the degree change in wind direction for the different ADMS-Urban scenarios) as well as between the covariates themselves. The top row of the pairs plot in Figure 5.1 shows that there appears to be a strong positive relationship between the simulated NO_2 annual averages and emissions, whereas there appears to be a weak negative relationship between the simulated values and wind speed. Furthermore, there is evidence for a very weak positive linear relationship with wind direction. These conclusions are supported by the Pearson correlation coefficients (see Subsection 2.1.2) with the response (and their respective 95% CIs), which are 0.94 (0.90, 0.96) for emissions, -0.43 (-0.58, -0.25) for wind speed, and 0.18 (-0.02, 0.36) for wind direction, which indicates there is no significant relationship but the interval is almost entirely positive suggesting that the relationship might be very weak and hard to capture.

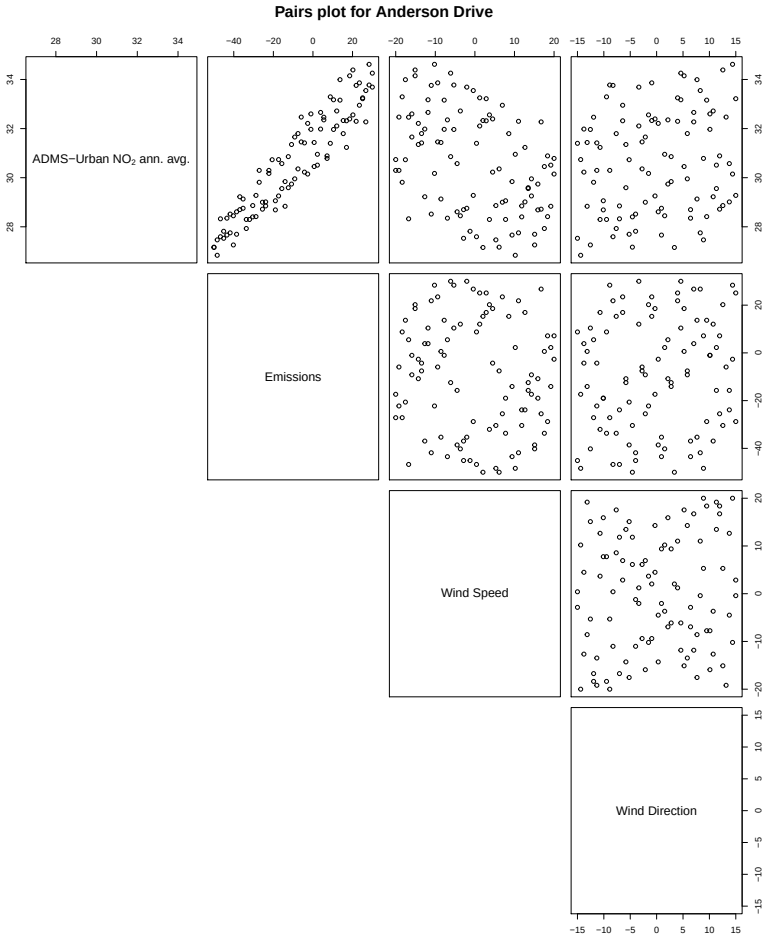

FIGURE 5.1: Pairs plot for the LHC inputs (emissions (% change), wind speed (% change) and wind direction (° change)) and the NO_2 annual average (μg m^{-3}) from the ninety-eight simulations from ADMS-Urban for the Anderson Drive station in Aberdeen.

From the second and third rows of Figure 5.1, it is clear that there is no evidence of relationships between any of the covariates as the points on all three plots appear to be randomly scattered. This is further confirmed by the Pearson correlation coefficients

(and their respective 95% CIs), which are -0.11 (-0.30, 0.09) for emissions and wind speed, 0.13 (-0.07, 0.32) for emissions and wind direction and 0.11 (-0.09, 0.30) between wind speed and wind direction. Zero lies in all three intervals, hence, there is no evidence for multicollinearity. Hence, a model with the three covariates was fitted and Table 5.1 provides its summary. All three covariates are highly significant as all the p-values are $<$ 2e-16. The emissions estimate is positive, which indicates that the higher the emissions for the year, the higher NO_2 annual average concentration for the year. The wind speed estimate is negative, which is expected based on the pairs plot in Figure 5.1, which indicates that the higher the wind speed in a year, the lower the NO_2 annual average concentration which is logical as the wind speed increases the dispersion of the pollutants. The wind direction estimate is positive, which again is expected based on the initial impressions from the pairs plots in Figure 5.1 and the Pearson correlation coefficients. The positive coefficient is most likely due to the city outline around the monitoring location.

Coefficient	**Estimate**	**Stand. Error**	**p-value**
Intercept	31.36	0.02	$<$ 2e-16
Emissions	0.08	0.0008	$<$ 2e-16
Wind Speed	-0.06	0.002	$<$ 2e-16
Wind Direction	0.02	0.002	$<$ 2e-16

TABLE 5.1: Summary of the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction ($^\circ$ change) as covariates.

Lastly, the diagnostic plots for the model are presented in Figure 5.2. The diagnostic plots indicate that the model is a good fit. The residuals vs. fitted values in plot a) are randomly scattered and there is very small variation in the points as the residuals range between -0.30 to 0.30. Furthermore, the points on the actual vs. fitted values plot in b) are lying on the equivalence line y=x. The residuals on the qq-plot are lying on the normality line indicating a good fit. The residuals are roughly symmetric as seen in plot d). A Shapiro-Wilk test was performed and a p-value of 0.66 was estimated indicating that normality is reasonable. Lastly, the plots with the residuals vs. each of the covariates (emissions, wind speed and wind direction) in plots e), f) and g) show random scatter of the points, evenly distributed around the zero line. The model has $R^2_{\text{adj.}} = 99.28\%$, which also indicates very good fit. This result is not surprising given the deterministic nature of ADMS-Urban and therefore, no further modelling is conducted and this is the final model.

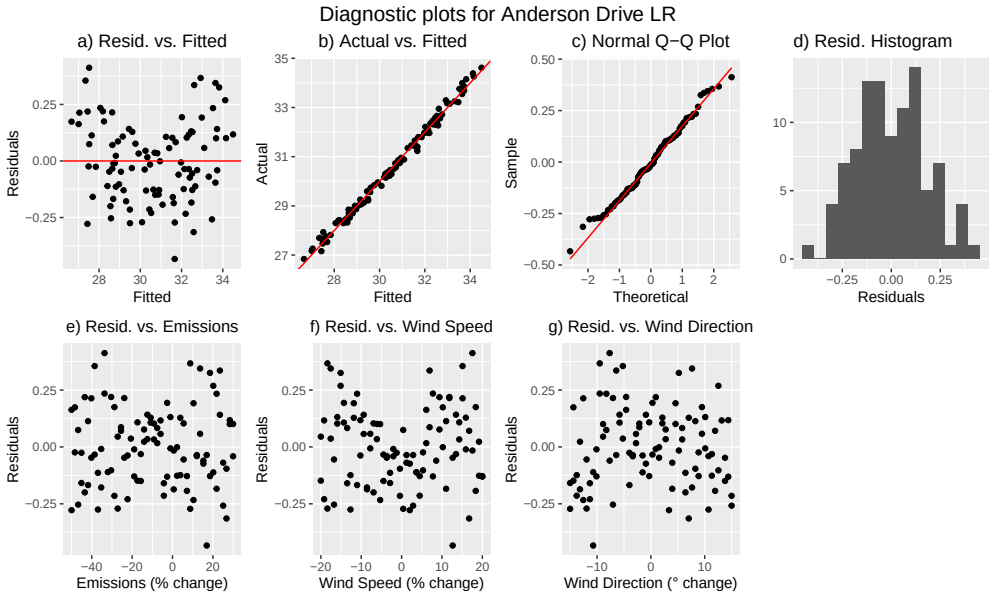

FIGURE 5.2: Diagnostic plots for the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

Errol Place

The model for the simulated NO_2 annual average at Errol Place with the three covariates is a good fit to the data based on the diagnostic plots (similar to the ones for the Anderson Drive station in Figure 5.2) and $R^2_{adj.} = 99.10\%$. Hence, the final linear model only has three covariates - the three inputs from the LHC. In a similar fashion to Anderson Drive, the parameter estimates for the emissions and wind direction are positive, whereas the parameter estimate for wind speed is negative.

King Street

In a similar way to the previous other two stations, at the King Street station, the model with just three covariates is a good fit to the data (based on the diagnostics plots similar to the ones in Figure 5.2) and has $R^2_{adj.} = 99.24\%$. Therefore, no further covariates were added to the model. Once again, the parameter estimates for emissions and wind direction are positive, whereas the wind speed parameter estimate is negative.

Market Street 2

The model for the Market Street 2 station with just three covariates is a good fit to the data as indicated by the diagnostic plots (alike to the plots presented in Figure 5.2) and

$R^2_{\text{adj.}} = 99.44\%$. No further covariates were used to model the response. The parameter estimates are positive for emissions and wind direction and negative for the wind speed.

Union Street

The model with just the three inputs as covariates for Union Street has diagnostic plots similar to those in Figure 5.2 and $R^2_{\text{adj.}} = 99.50\%$ which indicate the model is a good fit for the data. The parameter estimates are positive for emissions and wind direction but negative for the wind speed.

Wellington Road

The Wellington Road model is the same as the model for the other five monitoring stations in Aberdeen - the three inputs are significant and the diagnostic plots (similar to the ones in Figure 5.2) indicate a good fit and $R^2_{\text{adj.}} = 99.27\%$. Again, similarly, to the models for the other stations, the parameter estimates for emissions and wind direction are positive, whereas the parameter estimate for wind speed in negative.

Overall Comparison

In Table 5.2, the in- and out-of sample RMSPE based on 10-fold CV are presented for each of the six stations. The stations with lower NO_2 annual average (Anderson Drive, Errol Place and King Street) have smaller in- and out-of-sample performances. The models for the other three stations (Market Street 2, Union Street and Wellington Road) above the annual regulation of 40 μg m^{-3} have very similar performance which is expected based on the boxplots. Similarly, as the RMSPE value increases, the 95% bootstrap confidence intervals get wider. For all models, the RMSPE is very low indicating almost perfect predictions.

Table 5.3 provides a summary of the parameter estimates and their respective standard deviations for all six monitoring stations. From the table, it is clear that all parameters are highly significant with very small p-values indicating that all parameters are statistically significant. For all models, the emissions parameter is positive which is logical - as emissions increase so does the NO_2 annual average. On the other hand, all wind speed parameters are negative which is also logical - as wind speed increases, the dispersion of the pollutants is faster and the NO_2 annual average is reduced. The wind direction terms are all positive, therefore as the winds become more western prevailing, the NO_2 annual average is increased, which is logical given Aberdeen's geographical location. Overall,

Station	In	Out
Anderson Drive	0.17 (0.15, 0.19)	0.18 (0.16, 0.21)
Errol Place	0.12 (0.11, 0.14)	0.13 (0.11, 0.15)
King Street	0.28 (0.25, 0.32)	0.30 (0.26, 0.34)
Market Street 2	0.43 (0.38, 0.48)	0.45 (0.40, 0.51)
Union Street	0.43 (0.37, 0.50)	0.46 (0.39, 0.52)
Wellington Road	0.44 (0.39, 0.49)	0.46 (0.41, 0.52)

TABLE 5.2: Comparing the predictive performance of the preferred linear regressions (with the three inputs as covariates) for predicting the ADMS-Urban simulations runs for the NO_2 annual average (μg m^{-3}) for all monitoring stations in Aberdeen. The 95% bootstrapping intervals are provided in brackets.

all models indicate a very good fit with $R^2_{\text{adj.}}$ values of about 99% and no issues on the diagnostic plots.

Station	Coefficient	Estimate	Stand. Error	p-value
Anderson Drive	Intercept	31.36	0.02	< 2e-16
	Emissions	0.08	0.0008	< 2e-16
	Wind Speed	-0.06	0.002	< 2e-16
	Wind Direction	0.02	0.002	< 2e-16
Errol Place	Intercept	28.12	0.01	< 2e-16
	Emissions	0.05	0.0005	< 2e-16
	Wind Speed	-0.05	0.001	< 2e-16
	Wind Direction	0.01	0.001	9e-15
King Street	Intercept	36.04	0.03	< 2e-16
	Emissions	0.12	0.001	< 2e-16
	Wind Speed	-0.10	0.002	< 2e-16
	Wind Direction	0.05	0.003	< 2e-16
Market Street 2	Intercept	47.48	0.05	< 2e-16
	Emissions	0.22	0.002	< 2e-16
	Wind Speed	-0.17	0.004	< 2e-16
	Wind Direction	0.06	0.005	< 2e-16
Union Street	Intercept	49.07	0.05	< 2e-16
	Emissions	0.23	0.002	< 2e-16
	Wind Speed	-0.19	0.004	< 2e-16
	Wind Direction	0.02	0.005	0.0002
Wellington Road	Intercept	43.95	0.05	< 2e-16
	Emissions	0.19	0.002	< 2e-16
	Wind Speed	-0.16	0.004	< 2e-16
	Wind Direction	0.07	0.005	< 2e-16

TABLE 5.3: Summary of the linear regressions for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for the six monitoring stations in Aberdeen with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

5.2.2 GP modelling of the Aberdeen data

In order to improve the quality of the predictions in terms of RMSPE and establish a framework for handling non-deterministic simulation models, the GP models (as recommended by [169]) for the simulated by ADMS-Urban NO_2 annual average concentration for each station individually are fitted using the **DiceKriging** package [165] in R [157]. GP are used to check whether estimating the correlation between the points in the LHC space would improve the predictions. Four different kernels will be compared using the in- and out-of-sample RMSPE on a 10-fold CV. Since the 3D variograms plots in Figure 4.11 did not show evidence of a nugget effect, the nugget effect for these GP models is set to zero. As with the linear regression models, the Anderson Drive case will be presented in full, whereas short summaries about the other stations will be provided in order to avoid repetition.

Anderson Drive

Firstly, the exponential kernel is used for modelling the data. The model with just the three inputs as covariates is presented in Table 5.4. The fixed effect parameter estimates are very similar to those for the linear model shown in Table 5.1 but as expected the standard error estimates are larger than those for the linear model in Table 5.1. The estimated random effects are presented in Table 5.5. The hyperspatial range parameter $\widehat{\boldsymbol{\theta}}$ estimates are quite high values (based on the observed span of the covariates) as expected based on the individual variograms for Aberdeen in Figures 4.8, 4.9 and 4.10. The variance is quite small at just 0.07. Lastly, the diagnostic plots for the model were examined to assess the fit of the model to the data. The plots in Figure 5.3 indicate that the model is a good fit. The residuals vs. fitted values plot in a) are randomly scattered. Furthermore, the residuals vs. each variable plots in e), f) and g) are also randomly scattered. The actual vs. fitted values plot in b) lie on the equivalence line indicating the model predicts quite well. The normal qq-plot in plot c) indicates there are a few outliers (also visible in the residuals plots in a), e), f) and g)) and hence, the tail points are off the normality line. However, looking at the histogram of the residual values in d), it appears that the residuals are normally distributed with just a few outliers above 0.25, which is also confirmed by a Shapiro-Wilk test with a p-value of 0.65. It has to be noted that the overall variance in the residuals is reduced as the main cloud of residuals lies in the interval -0.10 to 0.10 in comparison to the linear regression in Figure 5.2, where the main cloud of the residuals are in the interval -0.25 to 0.25. Therefore, the range of the residuals is too small to indicate a problem with the fit.

Coefficient	Estimate	Stand. Error
Intercept	31.31	0.17
Emissions	0.08	0.003
Wind Speed	-0.06	0.006
Wind Direction	0.02	0.006

TABLE 5.4: Summary of the fixed effect parameters from the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

$\widehat{\theta}_{\text{Emissions}}$	$\widehat{\theta}_{\text{Wind Speed}}$	$\widehat{\theta}_{\text{Wind Direction}}$	$\widehat{\sigma}^2$
98.16	46.16	59.55	0.07

TABLE 5.5: Summary of the hyperspatial range parameters and variance from the GP regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

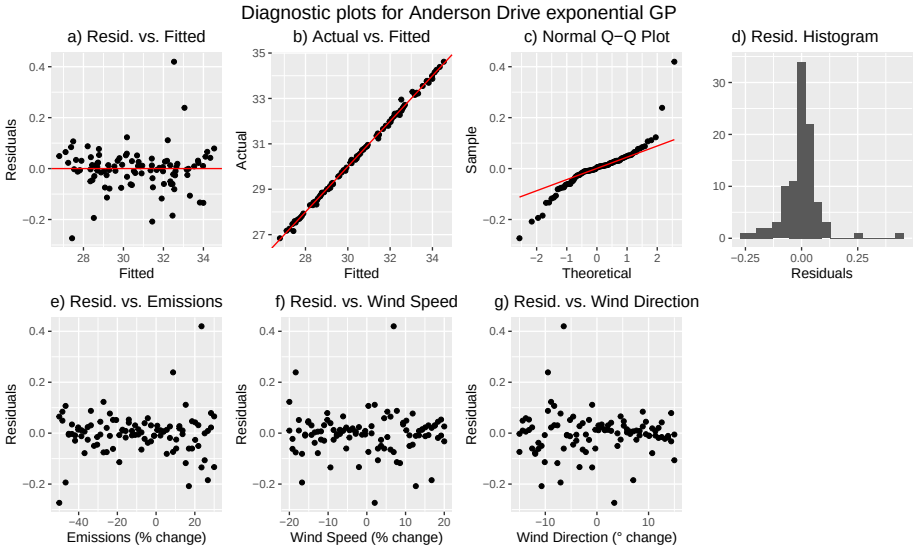

FIGURE 5.3: Diagnostic plots for the GP model with an exponential kernel for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

Next, the Matérn $\frac{3}{2}$, the Matérn $\frac{5}{2}$ and Gaussian kernels were also fitted and the in- and out-of-sample predictive performance of the models is compared to the exponential kernel predictive performance in Table 5.6. The in-sample prediction performance is calculated using a leave-one-out validation, i.e. 98-fold CV. Overall, the four models have very similar performance both in- and out-of-sample. The in-sample and out-of-sample performance as well as the confidence intervals sometimes appear the same due to rounding. The exponential kernel does provide the lowest out-of-sample performance, with almost twice the reduction in comparison to the Matérn $\frac{5}{2}$ which has the highest out-of-sample prediction error. Based on the 95% bootstrap confidence intervals, the difference between the exponential and the Matérn $\frac{3}{2}$ kernels, and the Matérn $\frac{5}{2}$ and

Gaussian kernels are significant but within the pairs there is no statistically significant difference. However, as the exponential kernel has slightly lower RMSPE, it is chosen as the final model.

Model	**E+WS+WD**	
Kernel	**In**	**Out**
Exponential	0.08 (0.06, 0.11)	0.09 (0.06, 0.11)
Matérn $\frac{3}{2}$	0.09 (0.06, 0.11)	0.10 (0.07, 0.12)
Matérn $\frac{5}{2}$	0.17 (0.15, 0.19)	0.17 (0.15, 20)
Gaussian	0.16 (0.14, 0.18)	0.16 (0.14, 0.18)

TABLE 5.6: Comparing the predictive performance of the GP models under different kernels for the ADMS-Urban simulations for the NO_2 annual concentrations (μg m^{-3}) for the Anderson Drive station with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

Errol Place

The predictive performance of the four kernels were compared for Errol Place. It was found that the exponential kernel provides the lowest out-of-sample performance. The hyperspatial range parameters are estimated to be $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 92.64, \widehat{\theta}_{\mathrm{WS}} = 36.39, \widehat{\theta}_{\mathrm{WD}} = 73.05]^{\top}$ which indicates high correlation between the inputs as expected. Furthermore, the variance estimate $\widehat{\sigma}^2 = 0.03$. The estimates are higher than the ones for Anderson Drive but still indicate that the measurement and random errors in the data are quite small.

King Street

At the King Street station, the four different kernels again performed very similarly in- and out-of-sample but the exponential kernel performed best. The correlation measured between the inputs is measured as the hyperspatial range parameter is estimated as $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 82.25, \widehat{\theta}_{\mathrm{WS}} = 29.79, \widehat{\theta}_{\mathrm{WD}} = 83.43]^{\top}$ which indicates high correlation between the inputs, although not as high as for the other stations. The variance estimate is $\widehat{\sigma}^2 = 0.14$ which suggests that the measurement and random errors left after modelling the data are quite small.

Market Street 2

The best kernel for the Market Street 2 station was found to be the exponential kernel based on the out-of-sample prediction power. The model estimates high correlation

between the inputs with the hyperspatial range parameter $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 89.65, \widehat{\theta}_{\mathrm{WS}} = 38.48, \widehat{\theta}_{\mathrm{WD}} = 66.12]^{\top}$. The estimated variance is $\widehat{\sigma}^2 = 0.40$ which is higher than the models for the other stations and it suggests that the model does not perform as well as the other models.

Union Street

For modelling the NO_2 annual average from the ADMS-Urban simulations for the Union Street station, the exponential kernel performed best in terms of prediction. The model estimates high correlation between the inputs with hyperspatial range parameter $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 49.40, \widehat{\theta}_{\mathrm{WS}} = 7.99, \widehat{\theta}_{\mathrm{WD}} = 2934.23]^{\top}$. The variance estimate is $\widehat{\sigma}^2 = 0.23$.

Wellington Road

Lastly, the different kernels were compared on their predictive performance at Wellington Road and the exponential kernel was found to perform best. The model estimates high correlation between the inputs with hyperspatial range parameter estimates $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 89.90, \widehat{\theta}_{\mathrm{WS}} = 39.62, \widehat{\theta}_{\mathrm{WD}} = 67.35]^{\top}$. The variance estimate for the model is $\widehat{\sigma}^2 = 0.39$.

5.2.3 Findings

In Table 5.7, the best models (both linear regression and GP) in terms of prediction power are presented for each station. For all stations, the same three covariates were used - the three inputs, i.e. emissions, wind speed and wind direction. These models are very simplistic but parsimonious and the diagnostic plots suggests the models are a good fit to the data. For all stations, both the in- and out-of-sample prediction performance is better for the GP cases by about two times which indicates that the estimation of the hyperspatial range parameters improves the predictive performance. The stations with higher NO_2 annual averages have larger in- and out-of-sample RMSPEs. It is interesting to note that all GP models choose the exponential kernel. Therefore, it has to be concluded that the different stations require the same level of smoothing. As the exponential kernel is chosen, this suggests that the rougher surfaces seem to work better for the Aberdeen data. Overall, both models have very small RMSPEs indicating almost perfect predictions but the GP models RMSPEs are twice as small as the linear regression ones. Furthermore, there is a clear statistical significant difference between the linear models and the GP models performance based on the 95% RMSPE bootstrap intervals. Therefore, the GP models are preferred.

Station	Model	Covariates	In	Out
Anderson Drive	LR	E+WS+WD	0.17 (0.15, 0.19)	0.18 (0.16, 0.21)
	GP Exponential	E+WS+WD	0.08 (0.06, 0.11)	0.09 (0.06, 0.11)
Errol Place	LR	E+WS+WD	0.12 (0.11, 0.14)	0.13 (0.11, 0.15)
	GP Exponential	E+WS+WD	0.06 (0.04, 0.08)	0.06 (0.04, 0.08)
King Street	LR	E+WS+WD	0.28 (0.25, 0.32)	0.30 (0.26, 0.34)
	GP Exponential	E+WS+WD	0.13 (0.09, 0.18)	0.14 (0.10, 0.18)
Market Street 2	LR	E+WS+WD	0.43 (0.38, 0.48)	0.45 (0.40, 0.51)
	GP Exponential	E+WS+WD	0.22 (0.15, 0.29)	0.23 (0.16, 0.30)
Union Street	LR	E+WS+WD	0.43 (0.37, 0.50)	0.46 (0.39, 0.52)
	GP Exponential	E+WS+WD	0.26 (0.18, 0.35)	0.28 (0.19, 0.37)
Wellington Road	LR	E+WS+WD	0.44 (0.39, 0.49)	0.46 (0.41, 0.52)
	GP Exponential	E+WS+WD	0.21 (0.14, 0.27)	0.21 (0.15, 0.28)

TABLE 5.7: Comparing the predictive performance of the linear regression and the preferred GP model for predicting the ADMS-Urban simulations runs for the NO_2 annual average (μg m^{-3}) for all monitoring stations in Aberdeen with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

Table 5.8 compares the fixed effect parameter estimates and their respective standard errors from the linear regression and GP models. Overall, the parameter estimates for all models are almost identical. The standard errors are larger for the GP exponential models than the linear regressions as it is expected. This confirms that the fixed effects parts of the two models are very similar and the differences in the prediction powers are due to the additional information for the locations of the simulation scenarios within the LHC sample space. The emissions for all models have positive estimates suggesting that the higher the emissions, the higher the simulated NO_2 annual averages. On the other hand, the wind speed estimates are negative, which is logical as the higher the wind speed, the faster the dispersion of pollutants resulting in lower simulated NO_2 annual averages. Lastly, the wind direction parameters are all positive indicating that more western prevailing winds results in higher simulated NO_2 annual averages.

5.3 Modelling the individual stations in Glasgow

In this section, different models for the 2015 Glasgow NO_2 annual average as simulated with ADMS-Urban are presented based on the example presented for Aberdeen in Section 5.2.

Station	Coefficient	Estim. LM	St. Error LM	Estim. GP	St. Error GP
Anderson Drive	Intercept	31.36	0.02	31.31	0.17
	Emissions	0.08	0.0008	0.08	0.003
	Wind Speed	-0.06	0.002	-0.06	0.006
	Wind Direction	0.02	0.002	0.02	0.006
Errol Place	Intercept	28.12	0.01	28.07	0.11
	Emissions	0.05	0.0005	0.05	0.002
	Wind Speed	-0.05	0.001	-0.05	0.004
	Wind Direction	0.01	0.001	0.01	0.004
King Street	Intercept	36.04	0.03	36.01	0.23
	Emissions	0.12	0.001	0.12	0.004
	Wind Speed	-0.10	0.002	-0.11	0.009
	Wind Direction	0.05	0.003	0.05	0.007
Market Street 2	Intercept	47.48	0.05	47.33	0.40
	Emissions	0.22	0.002	0.22	0.007
	Wind Speed	-0.17	0.004	-0.17	0.01
	Wind Direction	0.06	0.005	0.06	0.01
Union Street	Intercept	49.07	0.05	49.45	0.20
	Emissions	0.23	0.002	0.24	0.005
	Wind Speed	-0.19	0.004	-0.20	0.01
	Wind Direction	0.02	0.005	0.02	0.002
Wellington Road	Intercept	43.95	0.05	43.84	0.40
	Emissions	0.19	0.002	0.19	0.007
	Wind Speed	-0.16	0.004	-0.17	0.01
	Wind Direction	0.07	0.005	0.07	0.01

TABLE 5.8: Summary of the linear regressions and GP exponential parameters for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for the six monitoring stations in Aberdeen with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

5.3.1 Linear regression modelling of the Glasgow data

The simplest model that is applied is a linear regression as presented in Subsection 2.2.1. As seen in the boxplots in Figure 4.20, normality can be assumed for the distribution of the NO_2 annual averages for each station and a linear regression was fitted for each. In its simplest version, each model takes the three inputs (emissions, wind speed and wind direction) on the percentage variation scale from ADMS-Urban as covariates, and the response is the NO_2 annual average for each station. The modelling for Burgher Street will be presented in full due to the different geographical location of the station. To avoid repetition other stations modelling will be only be summarised.

Burgher Street

Firstly, the inputs were checked for multicollinearity using a pairs plot in Figure 5.4 and Pearson's correlation coefficient (see Section 2.1.2). Looking at the top row of the pairs plot, the NO_2 annual averages from the ADMS-Urban simulations for Burgher Street appear only correlated with emissions in which the points appear linearly correlated but at closer inspection there is slight curvature. The points on the plots between the response and wind speed and wind direction are randomly scattered. These conclusions are supported by the Pearson correlation coefficient with the response (and their respective 95% CIs), which are 0.98 (0.96, 0.98) for emissions, -0.14 (-0.33, 0.05) for wind speed and 0.03 (-0.17, 0.22) for wind direction. The correlation coefficients for wind speed and wind direction are close to zero and their 95% CIs include zero, so there is no evidence of correlation between the NO_2 annual averages from the ADMS-Urban simulations for Burgher Street and the wind speed and wind direction.

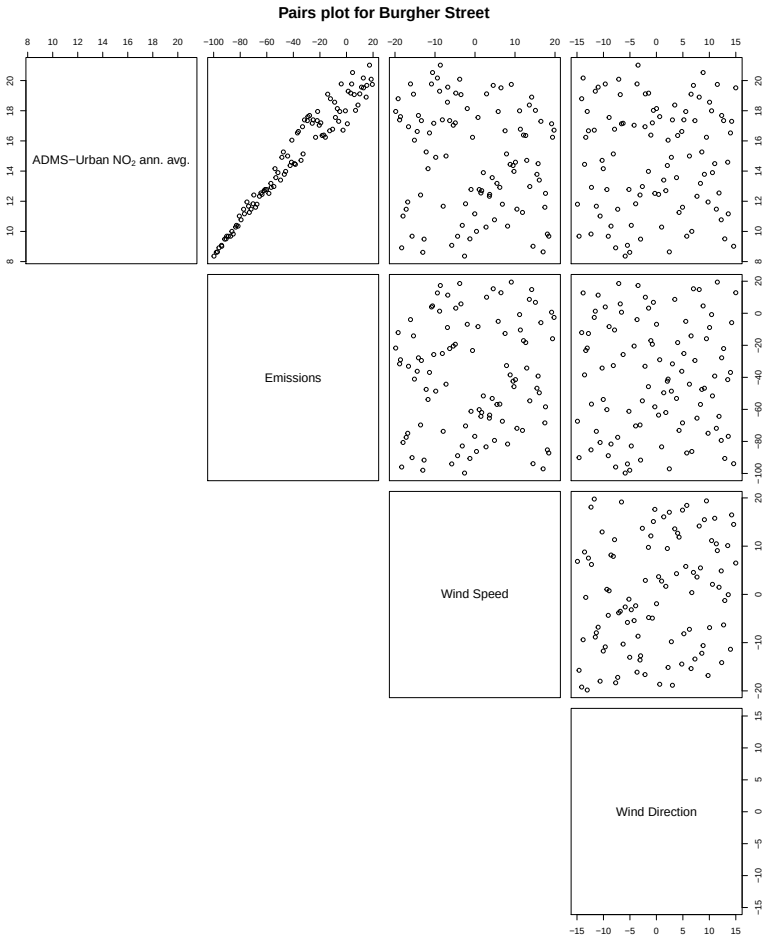

FIGURE 5.4: Pairs plot for the LHC inputs (emissions (% change), wind speed (% change) and wind direction (° change)) and the NO_2 annual average (μg m^{-3}) from the one hundred simulations of ADMS-Urban for the Burgher Street monitoring station in Glasgow.

From the second and third rows of Figure 5.4, it is clear that there is no correlation between the three covariates as the points are randomly scattered. This is further confirmed by the Pearson correlation coefficients (and their respective 95% CIs) which are: 0.03 (-0.17, 0.23) between emissions and wind speed, 0.00 (-0.19, 0.20) between emissions and wind direction, and 0.19 (0.00, 0.37) for wind speed and wind direction. The wind speed and wind direction correlation interval suggests there might be a weak positive relationship as the lower end of the interval is zero which is hard to capture with only one hundred observations, whereas the other two estimated correlation coefficients are close to zero, which is almost the centre of their respective 95% CIs.

Since there is no evidence for multicollinearity, the model with all three covariates is summarised in Table 5.9. All the covariates are significant as their p-values are smaller than 0.05. As is expected, the emissions estimate is positive as the more emissions there are, the higher the NO_2 concentration is in the air. Similarly, the wind speed estimate is expected to be negative as the wind speed increases, the emissions disperse faster and the NO_2 concentration is reduced. It is interesting that as the wind direction increases so does the annual concentration. This is most likely related to the city outline near the Burgher Street station. Lastly, the model has $R^2_{\text{adj.}} = 98.49\%$ which suggests a good fit.

Coefficient	**Estimate**	**Stand. Error**	**p-value**
Intercept	18.80	0.07	< 2e-16
Emissions	0.10	0.001	< 2e-16
Wind Speed	-0.06	0.004	< 2e-16
Wind Direction	0.02	0.005	4.27e-06

TABLE 5.9: Summary of the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

Lastly, the diagnostic plots for the model are presented in Figure 5.5. The residual vs. fitted values in plot a) form a concave parabola. A similar concave shape is seen in the residual values vs. emissions in plot e), which suggests that a new model should be refitted with a square term of the emissions. Moreover, the diagnostic plots for the other seven stations exhibit similar characteristics of bad fit suggesting that a squared term for emissions should be added to all models.

The refitted model with squared emissions for Burgher Street is summarised in Table 5.10. Once again, all the covariates are highly significant as the p-values are much smaller than 0.05. The estimate for emissions is slightly smaller in comparison to the model with just three covariates presented in Table 5.9 but this is expected given that a square term has been added. The squared emissions term has a negative estimate which is to be expected given the concave shape of the parabola in Figure 5.5 plots a) and e). The estimates for wind speed and wind direction have not changed. However, the

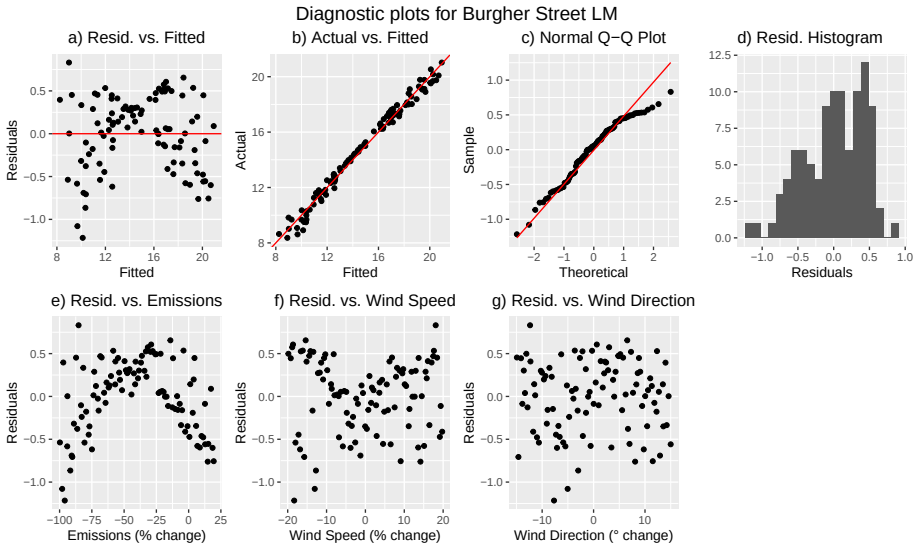

FIGURE 5.5: Diagnostic plots for the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

standard error for both has been reduced in comparison to the ones for the model with three covariates in Table 5.9. The model has $R^2_{adj.} = 99.20\%$, which is higher than the baseline model.

Coefficient	**Estimate**	**Stand. Error**	**p-value**
Intercept	18.69	0.05	< 2e-16
Emissions	0.08	0.003	< 2e-16
Emissions^2	-0.0003	0.00003	6.22e-15
Wind Speed	-0.06	0.003	< 2e-16
Wind Direction	0.02	0.003	1.10e-07

TABLE 5.10: Summary of the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates.

Next, the diagnostic plots for the linear regression with emissions, emissions squared, wind speed and wind direction for covariates are presented in Figure 5.6. The residual vs. fitted values in plot a) exhibit a butterfly pattern which can also be seen in plots e) and f). Since there are issues with both plots e) and f), adding an interaction term is one possible solution.

Hence, a model with the squared term for emissions and an interaction for emissions and wind speed for Burgher Street was fitted. The model has $R^2_{adj.} = 99.80\%$. A summary for that model is provided in Table 5.11. All variables are highly statistically significant. The standard errors for all estimates are smaller in comparison to those in Table 5.10. Additionally, the estimate for wind direction has not changed in comparison to the model presented in Table 5.10. The diagnostic plots for the model with five covariates are examined in Figure 5.7. The plots show the model is a good fit for the data in

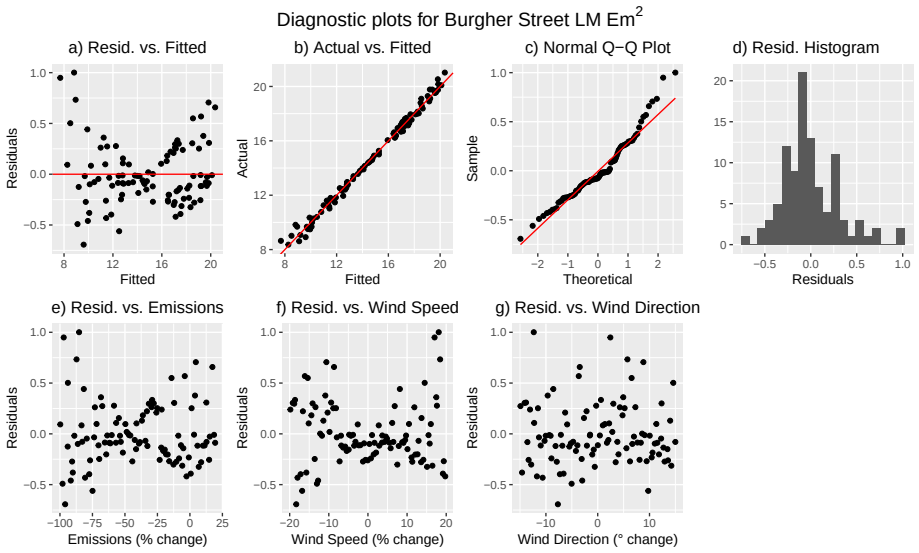

FIGURE 5.6: Diagnostic plots for the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates.

comparison to the previous plots. The residual vs. fitted values in plot a), as well as the residual values vs. the three input variables (emissions, wind speed, wind direction) in plots e), f) and g) respectively, show that the points are randomly scattered indicating that there is no issues with the fit. The actual vs. fitted values in plot b) has all points lying on the equivalence line which indicates the model is a good fit to the data. The normal qq-plot has the points lying on the normality line also suggesting a good fit. Lastly, the histogram of the residual values in plot d) is relatively symmetric. This is further confirmed with a Shapiro-Wilk test, where the p-value is 0.83 and thus, failing to reject the null hypothesis that the residuals are normally distributed.

Coefficient	**Estimate**	**Stand. Error**	**p-value**
Intercept	18.72	0.03	$<$ 2e-16
Emissions	0.08	0.001	$<$ 2e-16
Emissions^2	-0.0002	0.00001	$<$ 2e-16
Wind Speed	-0.08	0.002	$<$ 2e-16
Emissions*Wind Speed	-0.0007	0.00004	$<$ 2e-16
Wind Direction	0.02	0.002	$<$ 2e-16

TABLE 5.11: Summary of the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates.

The three models with different numbers of covariates for the ADMS-Urban simulations for Burgher Street are compared in terms of predictive performance in- and out-of-sample based on RMSPE. The results are summarised in Table 5.12. Both in- and out-of-sample agree that the model with five covariates is the best model as the RMSPE values for

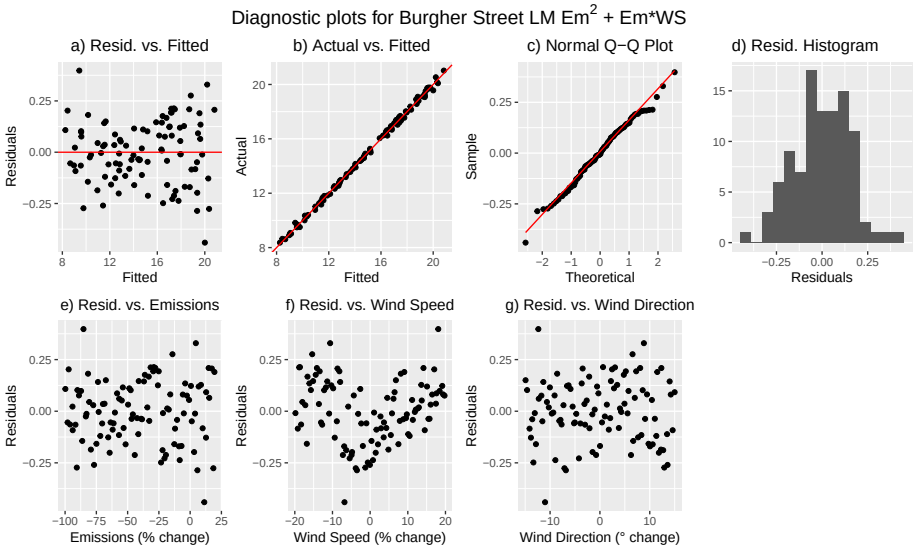

FIGURE 5.7: Diagnostic plots for the linear regression for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates.

both are the lowest. The agreement between the in- and out-of-sample RMSPE indicates that the models do not overfit. Furthermore, the 95% bootstrap intervals indicate that differences between the models are significant as there is no overlap between the intervals. Therefore, the best linear regression for the NO_2 annual average concentration for Burgher Street in terms of prediction performance is the model with the three inputs, emissions squared and the interaction between emissions and wind speed.

Model	**In**	**Out**
Three covariates	0.42 (0.37, 0.48)	0.44 (0.38, 0.50)
Four covariates	0.31 (0.26, 0.36)	0.32 (0.27, 0.38)
Five covariates	0.15 (0.13, 0.17)	0.16 (0.14, 0.18)

TABLE 5.12: Comparing the predictive performance of the different linear regressions for ADMS-Urban simulations runs for the NO_2 annual concentrations (μg m^{-3}) for the Burgher Street station. The 95% bootstrap intervals for the RMSPE are also included.

Byres Road

For all models for the simulated NO_2 annual average at Byres Road, wind direction is not a significant term and was removed from all models. However, the squared emissions and the interaction between emissions and wind speed are significant and improve the model fit. The in- and out-of-sample RMSPE show that adding these additional terms improves the predictions and the best model for predictions has the two additional terms.

The final model has $R^2_{adj.} = 99.96\%$. The final diagnostic plots are very similar to those in Figure 5.7 and hence, omitted to avoid repetition.

Central Station

The Central Station case is quite similar to the Byres Road one as for all models of the simulated NO_2 annual average for Central Station, wind direction is never a statistically significant term. Hence, wind direction is not used for modelling. However, the squared emissions term and the interaction between emissions and wind speed are significant and improve the diagnostic plots. The in- and out-of-sample RMSPE agreed that the best prediction model has emissions squared and an interaction between emissions and wind speed in addition to just emissions and wind speed. The final model has $R^2_{adj.} = 99.91\%$. Again, the final diagnostic plots are very similar to those in Figure 5.7 and hence, omitted to avoid repetition.

Dumbarton Road

When modelling the NO_2 annual average for Dumbarton Road, all inputs are significant. Furthermore, the diagnostic plots suggest that a squared emissions term and an interaction between emissions and wind speed should be added to the model. The in- and out-of-sample RMSPE indicated that the model with the squared term and interaction is the best out of the three models in terms of prediction powers. The final model has $R^2_{adj.} = 99.92\%$ and the its diagnostic plots are very similar to those in Figure 5.7 and hence, omitted to avoid repetition.

Great Western Road

Initially, when modelling the NO_2 annual averages for Great Western Road, wind direction is not significant until the squared term for emissions is added. The in- and out-of-sample RMSPE for the NO_2 annual concentration at Great Western Road showed that the model with five covariates has the best prediction powers in comparison to the others. The final model has $R^2_{adj.} = 99.95\%$ and the diagnostic plots are very similar to those in Figure 5.7 and therefore, omitted to avoid repetition.

High Street

In the models for the NO_2 annual average concentration for High Street, wind direction only becomes significant after the squared emissions and the interaction between

emissions and wind speed are added to improve the fit based on the diagnostic plots. Both the in- and out-of-sample RMSPE agree that the best model has both the squared emissions term and the interaction between emissions and wind speed. The final model has $R^2_{\text{adj.}} = 99.95\%$ and its diagnostic plots are very similar to those in Figure 5.7 and as before are omitted to avoid repetition.

Townhead

Similarly to High Street, wind direction is only significant after the squared emissions term and the interaction between emissions and wind speed are added to the models for the simulated NO_2 annual average for the Townhead monitoring station. The models were compared on their in- and out-of-sample predictive power, which indicated that the model with five covariates is the best one in terms of predictive power. The final model has $R^2_{\text{adj.}} = 99.94\%$. The diagnostic plots of the final model are very similar to those in Figure 5.7 and hence, omitted to avoid repetition.

Waulkmillglen Reservoir

In the case for the Waulkmillglen Reservoir, all of the three ADMS-Urban inputs are significant. As with all other models, a squared emissions and an interaction between emissions and wind speed terms are added to improve the fit. Lastly, the in- and out-of-sample predictive error for the models fitted were compared and the model with five covariates was chosen as the best prediction model as it has the most accurate predictions in comparison to the other models. The final model has $R^2_{\text{adj.}} = 98.94\%$. Its diagnostic plots are very similar to those in Figure 5.7 and therefore, omitted to avoid repetition.

Overall Comparison

In Table 5.13, the in- and out-of-sample RMSPE for the final model for every station are presented. All models have five covariates - emissions, emissions squared, wind speed, an interaction between wind speed, and wind direction. There are two stations (Byres Road and Central Station) for which wind direction is not significant and has not been included in the final model. The models for which wind direction is not significant are marked by an asterisk in Table 5.13. Both the in- and out-of-sample measurements are lowest for Waulkmillglen Reservoir but it has to be noted that simulated NO_2 annual averages for this station are smaller than the simulated values for the other stations as seen in Section 4.2. This is further confirmed by the fact that both the in- and out-of-sample predictive errors are highest for Central Station, where the highest simulated

NO_2 annual concentrations are observed. Furthermore, the 95% bootstrap intervals are much wider for Central Station in comparison to the other stations. The RMSPE for all other stations are all similar as expected. Overall, both the in- and out-of-sample RMSPE values are very close to zero indicating an almost perfect prediction.

Station	**In**	**Out**
Burgher Street	0.15 (0.13, 0.17)	0.16 (0.14, 0.18)
Byres Road *	0.16 (0.14, 0.18)	0.17 (0.15, 0.19)
Central Station *	0.50 (0.43, 0.59)	0.53 (0.45, 0.62)
Dumbarton Road	0.27 (0.23, 0.31)	0.29 (0.24, 0.33)
Great Western Road	0.19 (0.16, 0.21)	0.20 (0.17, 0.20)
High Street	0.19 (0.16, 0.21)	0.20 (0.17, 0.23)
Townhead	0.16 (0.14, 0.18)	0.17 (0.14, 0.19)
Waulkmillglen Reservoir	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)

TABLE 5.13: Comparing the predictive performance of the preferred linear regressions (with all five covariates) for predicting the ADMS-Urban simulations for the NO_2 annual average (μg m^{-3}) for all monitoring stations in Glasgow. An asterisk indicates that wind direction was not included in the model. The 95% bootstrap intervals for RMSPE are also included.

Tables 5.14 and 5.15 summarise the parameter estimates and their respective standard deviations for the eight monitoring stations in Glasgow. All parameters are highly significant as seen from the p-values. For all models, the emissions parameter is positive as expected since the more emissions, the larger the pollutant concentrations. On the other hand, the emissions squared terms are all negative. The wind speed parameters are negative for all monitoring stations which follows the logic that the higher the wind speed, the faster dispersion of pollutants in the air. The interaction between emissions and wind speed has also negative parameter estimates. The wind direction parameter is where there are differences between the stations. For five of the stations (Dumbarton Road, Great Western Road, High Street, Townhead and Waulkmillglen Reservoir), wind direction has negative estimates which means eastern prevailing winds result in the lowering of pollutant concentrations, whereas for Burgher Street the estimate is positive and for Byres Road and Central Station, wind direction was found not be statistically significant. Overall, the models appear to be a very good fit for the data with $R^2_{\text{adj.}}$ values of about 99% and no issues on the diagnostic plots.

Station	Coefficient	Estimate	Stand. Error	p-value
Burgher Street	Intercept	18.72	0.03	< 2e-16
	Emissions	0.08	0.001	< 2e-16
	Em^2	-0.0002	0.00001	< 2e-16
	Wind Speed	-0.08	0.002	< 2e-16
	Em*WS	-0.0007	0.00004	< 2e-16
	Wind Direction	0.02	0.002	< 2e-16
Byres Road	Intercept	34.06	0.03	< 2e-16
	Emissions	0.20	0.001	< 2e-16
	Em^2	-0.001	0.00002	< 2e-16
	Wind Speed	-0.20	0.002	< 2e-16
	Em*WS	-0.002	0.00004	< 2e-16
	Wind Direction			
Central Station	Intercept	63.19	0.08	< 2e-16
	Emissions	0.35	0.004	< 2e-16
	Em^2	-0.002	0.00005	< 2e-16
	Wind Speed	-0.33	0.007	< 2e-16
	Em*WS	-0.003	0.0001	< 2e-16
	Wind Direction			
Dumbarton Road	Intercept	37.65	0.04	< 2e-16
	Emissions	0.23	0.002	< 2e-16
	Em^2	0.0006	0.00003	< 2e-16
	Wind Speed	-0.23	0.004	< 2e-16
	Em*WS	-0.002	0.00007	< 2e-16
	Wind Direction	-0.05	0.003	< 2e-16

TABLE 5.14: Summary of the linear regressions for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for four monitoring stations (Burgher Street, Byres Road, Central Station and Dumbarton Road) in Glasgow with emissions (% change), emissions squared, wind speed (% change), an interaction for emissions and wind speed, and wind direction (° change) as covariates.

5.3.2 GP modelling of the Glasgow data

Applying the **DiceKriging** package [165] in R [157], GP models were fitted (as recommended by [169]) for the simulated by ADMS-Urban NO_2 annual average for each of the eight monitoring stations in Glasgow to check whether estimating the correlations between the points in the LHC improves the predictions in terms of RMSPE. Four different kernels are applied to account for the correlations between the scenarios and are compared using in- and out-of-sample prediction errors. As the 3D variograms in Figure 4.24 show no evidence of a nugget effect, the nugget is set to zero. Based on the Aberdeen models, an upper boundary limit of 1000 is set for the hyperspatial range parameters. The value of the boundary limit is chosen as it is much larger than observed ranges. As with all other subsections, the modelling for one monitoring station, Burgher Street, will be presented in full to avoid repetition.

Station	Coefficient	Estimate	Stand. Error	p-value
Great Western Road	Intercept	33.73	0.03	< 2e-16
	Emissions	0.19	0.002	< 2e-16
	Em^2	-0.0006	0.00002	< 2e-16
	Wind Speed	-0.19	0.003	< 2e-16
	Em*WS	-0.002	0.00005	< 2e-16
	Wind Direction	-0.02	0.002	5.82e-14
High Street	Intercept	35.42	0.03	< 2e-16
	Emissions	0.19	0.002	< 2e-16
	Em^2	-0.0008	0.00002	< 2e-16
	Wind Speed	-0.20	0.003	< 2e-16
	Em*WS	-0.002	0.00005	< 2e-16
	Wind Direction	-0.006	0.002	< 2e-16
Townhead	Intercept	29.37	0.03	< 2e-16
	Emissions	0.14	0.001	< 2e-16
	Em^2	-0.0007	0.00002	< 2e-16
	Wind Speed	-0.16	0.002	< 2e-16
	Em*WS	-0.001	0.00004	< 2e-16
	Wind Direction	-0.006	0.002	0.005
Waulkmillglen Reservoir	Intercept	9.75	0.08	< 2e-16
	Emissions	0.01	0.0004	< 2e-16
	Em^2	-0.00003	0.000005	1.74e-07
	Wind Speed	-0.01	0.0007	< 2e-16
	Em*WS	-0.0002	0.00001	< 2e-16
	Wind Direction	-0.007	0.0006	< 2e-16

TABLE 5.15: Summary of the linear regressions for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for four monitoring stations (Great Western Road, High Street, Townhead and Waulkmillglen Reservoir) in Glasgow with emissions (% change), emissions squared, wind speed (% change), an interaction for emissions and wind speed, and wind direction (° change) as covariates.

Burgher Street

Firstly, the exponential kernel is tested. The first model only takes the three inputs, which are summarised in Table 5.16. The parameter estimates are almost identical to the ones for the linear regression with three inputs as presented in Table 5.9 but the standard error estimates are increased as expected due to the GLS fit. The estimates for the three hyperspatial range parameters are presented in Table 5.17. The high values suggests that the random effects are highly correlated in the input space. Lastly, the diagnostic plots for the model were examined from Figure 5.8. Overall, all plots indicate a good fit except for the histogram and qq-plot of the residuals in plots c) and d), dominated by the deterministic nature of ADMS-Urban. The tails are quite heavy on both ends, although the points only span from -0.25 to 0.3. Additionally, a Shapiro-Wilk test was performed and a p-value of 0.06 indicating that there is no evidence for non-normality of the residuals. However, to check if the fit could be improved, the squared

emissions term is added based on the prior knowledge from the linear regressions.

Coefficient	Estimate	Stand. Error
Intercept	18.59	0.30
Emissions	0.10	0.003
Wind Speed	-0.05	0.009
Wind Direction	0.02	0.004

TABLE 5.16: Summary of the fixed effect parameters from the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

$\widehat{\theta}_{\text{Emissions}}$	$\widehat{\theta}_{\text{Wind Speed}}$	$\widehat{\theta}_{\text{Wind Direction}}$	$\widehat{\sigma}^2$
104.86	50.12	312.01	0.16

TABLE 5.17: Summary of the hyperspatial range parameters and variance from the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

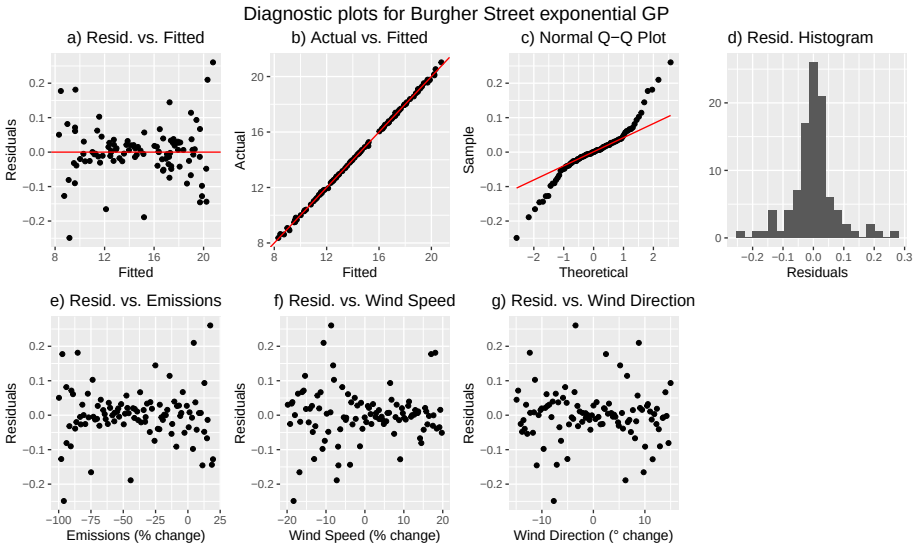

FIGURE 5.8: Diagnostic plots for the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

The model with four covariates (the three inputs from the LHC and the emissions terms squared) has all terms significant as seen in Table 5.18. Interestingly, due to adding the emissions squared term, the range parameters in $\widehat{\boldsymbol{\theta}}$ have changed as seen in Table 5.19 - the value for emissions has remained the same, whereas those for wind speed and wind direction have decreased. The variance estimate $\widehat{\sigma}^2$ has been reduced. The diagnostic plots for the model, in Figure 5.9, are quite similar to the ones for the model with just three covariates in Figure 5.8. Taking into account that in the linear regression case, the diagnostic plots improved the most after adding an interaction term between emissions and wind speed so that term will be added to check if this would improve the fit.

Coefficient	Estimate	Stand. Error
Intercept	18.75	0.23
Emissions	0.08	0.005
Emissions^2	-0.0002	0.00005
Wind Speed	-0.05	0.008
Wind Direction	0.02	0.004

TABLE 5.18: Summary of the fixed effect parameters for the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

$\widehat{\theta}_{\text{Emissions}}$	$\widehat{\theta}_{\text{Wind Speed}}$	$\widehat{\theta}_{\text{Wind Direction}}$	$\widehat{\sigma}^2$
104.88	32.33	203.64	0.10

TABLE 5.19: Summary of the hyperspatial range parameters and variance from the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

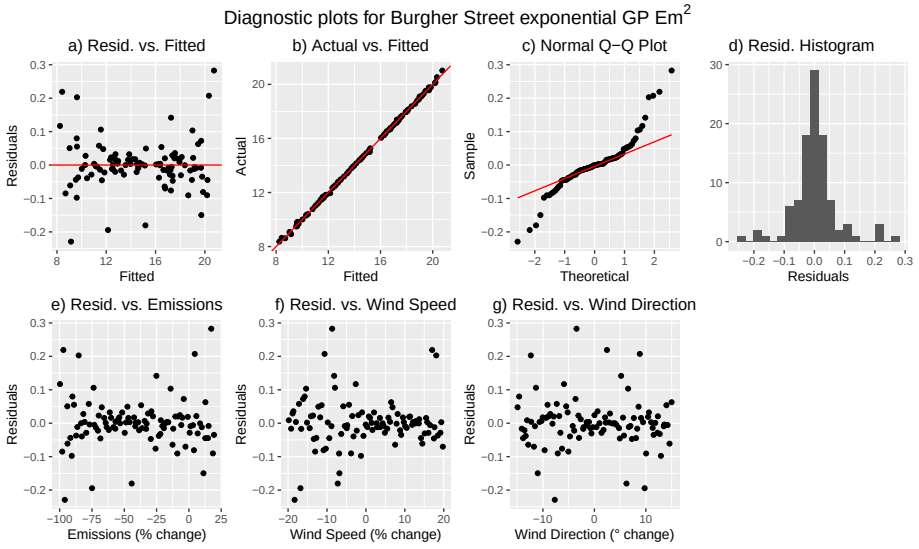

FIGURE 5.9: Diagnostic plots for the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

The model with five covariates (the three LHC inputs with emissions squared and an interaction between emissions and wind speed) has all covariates significant as seen in Table 5.20. The estimates are very similar to the ones from the linear regression presented in Table 5.11. The values for the hyperspatial range parameters are presented in Table 5.21. In this case, the emissions range parameter has increased in value, the wind speed one has remained the same, and the wind direction range parameter has decreased. The variance is reduced in comparison to the model with four covariates. The diagnostic plots in Figure 5.10 show an improvement as the residuals are now

within the range -0.2 to 0.2 and although the qq-plot in part c) looks like the points are off the tails, the range is too small to indicate a problem with the fit. A Shapiro-Wilk test was performed and it was found that the residuals follow the normal distribution.

Coefficient	**Estimate**	**Stand. Error**
Intercept	18.80	0.23
Emissions	0.08	0.005
Emissions^2	-0.0002	0.00005
Wind Speed	-0.08	0.0001
Emissions*Wind Speed	-0.0008	0.00002
Wind Direction	0.02	0.004

TABLE 5.20: Summary of the fixed effect parameters of the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

$\widehat{\theta}_{\textbf{Emissions}}$	$\widehat{\theta}_{\textbf{Wind Speed}}$	$\widehat{\theta}_{\textbf{Wind Direction}}$	$\widehat{\sigma}^2$
154.01	38.13	72.94	0.04

TABLE 5.21: Summary of the hyperspatial range parameters and variance from the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

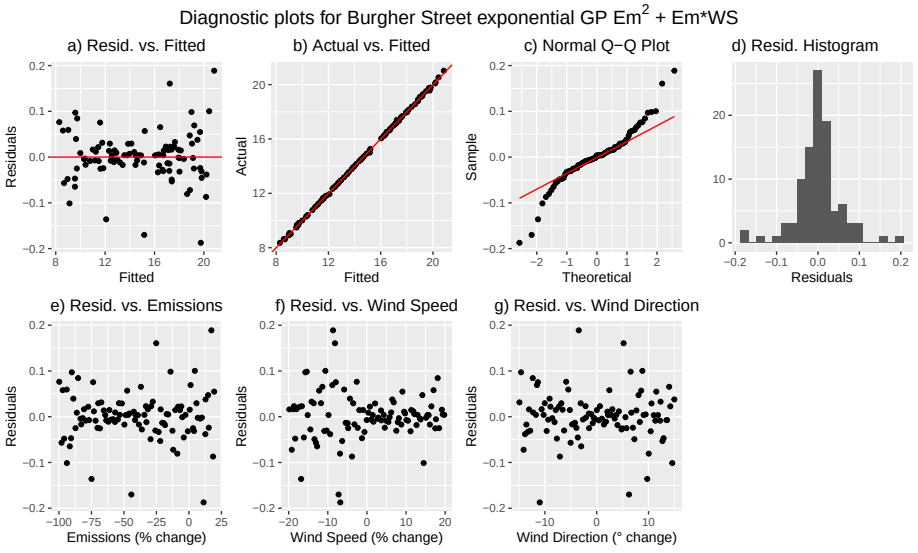

FIGURE 5.10: Diagnostic plots for the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

Lastly, the models were compared based on their prediction power as measured by both in- and out-of-sample RMSPE. In Table 5.22, it appears that the models with three and four covariates have almost the same prediction error estimates both in- and out-of-sample. Based on the 95% bootstrap intervals for the RMSPE, there is a statistically

significant difference between the 3 and 5 covariates models. Therefore, the five covariate model is performing better as the RMSPE values are the smallest. It has to be noted that both the in- and out-of-sample prediction errors for all three GP models are much smaller than the ones from the linear regression in Table 5.12 indicating that the GP models are performing better as accounting for the correlation between the inputs is improving the prediction power of the models.

Model	In	Out
Three covariates	0.07 (0.06, 0.09)	0.09 (0.07, 0.10)
Four covariates	0.07 (0.05, 0.09)	0.08 (0.06, 0.09)
Five covariates	0.05 (0.04, 0.06)	0.06 (0.04, 0.07)

TABLE 5.22: Comparing the predictive performance of the different GP models for ADMS-Urban simulations runs for the NO_2 annual concentrations (μg m^{-3}) for the Burgher Street station under the exponential kernel. 95% bootstrap confidence intervals for the RMSPEs are also provided.

The models under different kernels are very similar so they are omitted for brevity. For all kernels, the in- and out-of-sample RMSPE for all models are presented in Table 5.23. For short, the 3 covariate model has emissions, wind speed and wind direction; the 4 covariate model has emissions, emissions squared, wind speed and wind direction; and the 5 covariate model has emissions, emissions squared, wind speed, an interaction between emissions and wind speed, and wind direction as covariates. The overall best kernel is the exponential one with almost identical performance irrelevant of the number of covariates. Furthermore, the differences between the different kernels in terms of RMSPE are small and not statistically significant based on the overlapping 95% bootstrap confidence intervals. Hence, it appears that the model is invariant to kernel choice. It is interesting that for the Matérn $\frac{5}{2}$ kernels, the model with 4 covariates performs better than the model with 5 covariates suggesting that having 5 covariates might be an overfit. The diagnostic plots for all the models with different kernels are similar to those of the exponential ones in Figures 5.8, 5.9 and 5.10 and therefore, these plots are omitted to avoid repetition. However, the exponential kernel is the most numerically stable as there are singularity issues with the LHC input space for the three other kernels when estimating the hyperspatial range parameters. Therefore, although the exponential kernel does not provide the lowest RMSPE, it is chosen as the final model.

Byres Road

In the Byres Road case, the best kernel in terms of in- and out-of-sample predictive power was again the exponential. Although in the linear model for Byres Road the

	3 covariates		4 covariates		5 covariates	
Kernel	**In**	**Out**	**In**	**Out**	**In**	**Out**
Exponential	0.07 (0.06, 0.09)	0.09 (0.07, 0.10)	0.07 (0.05, 0.09)	0.08 (0.06, 0.09)	0.05 (0.04, 0.06)	0.06 (0.04, 0.07)
Matérn $\frac{3}{2}$	0.08 (0.07, 0.09)	0.10 (0.08, 0.12)	0.08 (0.06, 0.10)	0.09 (0.07, 0.11)	0.07 (0.06, 0.09)	0.08 (0.06, 0.10)
Matérn $\frac{5}{2}$	0.06 (0.04, 0.08)	0.07 (0.06, 0.08)	0.06 (0.05, 0.07)	0.07 (0.06, 0.09)	0.07 (0.06, 0.09)	0.15 (0.11, 0.20)
Gaussian	0.10 (0.07, 0.13)	0.12 (0.09, 0.15)	0.09 (0.07, 0.10)	0.11 (0.09, 0.13)	0.08 (0.06, 0.09)	0.09 (0.08, 0.10)

TABLE 5.23: Comparing the predictive performance of the GP models under different kernels for the ADMS-Urban simulations for the NO_2 annual concentrations (μg m^{-3}) for the Burgher Street station. 95% bootstrap confidence intervals for the RMSPEs are also provided.

wind direction was not significant, it was included in these models as wind direction is important to identify the location of the scenario in the LHC space. It was found that wind direction improves the prediction power of the models. This indicates that the linear regression model has failed to capture the importance of wind direction for the station at Byres Road. The diagnostic plots are very similar to Figure 5.10 and are omitted for brevity. The final model has hyperspatial range parameter estimates $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 111.39, \widehat{\theta}_{\mathrm{WS}} = 27.50, \widehat{\theta}_{\mathrm{WD}} = 1000.00]^{\top}$ which indicates high correlation between the input variables. It has to be noted that the wind direction hyperspatial range parameter is very large and reaches a boundary value which highlights the problem with estimating the effect of wind direction at this monitoring station. Nevertheless, the variance estimate is $\widehat{\sigma}^2 = 0.03$ which is quite low and indicates a good fit.

Central Station

For the Central Station case, wind direction was again included in the models as it is important to identify the LHC location of the different scenarios, even though the parameter was not significant in the linear model. Different kernels were compared based on the predictive performance and the best model has exponential kernel. The diagnostic plots are very similar to Figure 5.10 and hence, omitted. The final model has hyperspatial range parameter estimates $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 60.96, \widehat{\theta}_{\mathrm{WS}} = 69.30, \widehat{\theta}_{\mathrm{WD}} = 1000.00]^{\top}$ which indicates high correlation between the input variables. As with Byres Road, the hyperspatial range parameter for wind speed reaches a boundary value highlighting the problems with estimating the effect of wind direction at this station. The variance estimate is $\widehat{\sigma}^2 = 0.25$ which is higher than the previous monitoring station but not surprising given the much larger simulated NO_2 annual averages by ADMS-Urban for Central Station.

Dumbarton Road

For the Dumbarton Road monitoring station, the different kernels for GP models were compared and the exponential kernel provided the best prediction. The best model has five covariates as all other models for the other stations. The diagnostic plots for the final models are very similar to those in Figure 5.10 and therefore, omitted to avoid repetition. The hyperspatial range parameter estimates are $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 88.19, \widehat{\theta}_{\mathrm{WS}} = 76.76, \widehat{\theta}_{\mathrm{WD}} = 60.04]^{\top}$ indicating high correlation between the inputs and the variance estimate is $\widehat{\sigma}^2 = 0.11$.

Great Western Road

From the predictive performance of the different kernels at Great Western Road, it was clear that the models are very similar and invariant to the choice of kernel but the exponential kernel has the lowest RMSPE and is chosen as the final model. The diagnostic plots are very similar to those in Figure 5.10 and are omitted for brevity. The model has measured high correlation between the inputs as indicated by the hyperspatial range parameter estimates $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 92.72, \widehat{\theta}_{\mathrm{WS}} = 42.70, \widehat{\theta}_{\mathrm{WD}} = 107.66]^{\top}$. The variance estimate is $\widehat{\sigma}^2 = 0.05$.

High Street

From the predictive performance of the GP models with different kernels at High Street, the best model was found to be the one with all five covariates and the exponential kernel. The diagnostic plots for the model are similar to those in Figure 5.10 and hence, omitted. The correlation between the inputs is quite high as the hyperspatial range parameter estimates are $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 85.98, \widehat{\theta}_{\mathrm{WS}} = 32.14, \widehat{\theta}_{\mathrm{WD}} = 290.09]^{\top}$. The variance estimate is $\widehat{\sigma}^2 = 0.04$ which is quite low and indicates a good fit.

Townhead

The predictive performance of the GP models with various kernels were compared and the best prediction model for Townhead is the model with the five covariates and the exponential kernel. The diagnostic plots are similar to those in Figure 5.10 and hence, omitted. The parameter estimates for the hyperspatial range parameter estimates are $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_{\mathrm{E}} = 99.39, \widehat{\theta}_{\mathrm{WS}} = 28.91, \widehat{\theta}_{\mathrm{WD}} = 271.92]^{\top}$. The variance estimate is $\widehat{\sigma}^2 = 0.03$ which is quite low and indicates a good fit.

Waulkmillglen Reservoir

Lastly, the predictive performance of the different kernels for GP models were compared for the Waulkmillglen Reservoir station and it was found that the exponential kernel is best with only 3 covariates. However, all covariates are important in terms of model fit so the model with 5 covariates is preferred. The diagnostic plots are similar to those in Figure 5.10 and are omitted for brevity. The hyperspatial range parameter estimates are $\widehat{\boldsymbol{\theta}} = [\widehat{\theta_{\mathrm{E}}} = 170.53, \widehat{\theta}_{\mathrm{WS}} = 57.05, \widehat{\theta}_{\mathrm{WD}} = 33.68]^{\top}$. The variance estimate is $\widehat{\sigma}^2 = 0.004$ which is lower than for any other station but reasonable given that the ADMS-Urban simulated NO_2 annual averages are the lowest for Waulkmillglen Reservoir.

5.3.3 Findings

In Table 5.24, the best models (both linear regression and GP) based on the prediction power are presented for each station. For all models five covariates are used (emissions, emissions squared, wind speed, an interaction between emissions and wind speed, and wind direction) except for the linear regressions for Byres Road and Central Station where wind direction is not significant and not included in the models. All GP models perform better than the linear regression models with RMSPEs that are between 3 and 5 times smaller. Furthermore, the 95% bootstrap confidence intervals do not overlap each other. All stations have quite similar in- and out-of-sample RMSPE except for Central Station, where the ADMS-Urban simulated NO_2 annual averages are larger than the other stations, and Waulkmillglen Reservoir, where the ADMS-Urban simulated NO_2 annual averages are smaller than the other stations. The GP models with exponential models are preferred to the simple linear regression indicating that the prediction power improves by taking into account the location of the scenarios within the LHC space.

Tables 5.25 and 5.26 compare the fixed effect parameter estimates and their respective standard errors from the linear and GP models for the eight monitoring stations in Glasgow. Overall, the parameter estimates are almost identical. This further confirms that the fixed effects parameters for the two models are very similar and the main difference in their prediction powers come from the additional information of the locations of the ADMS-Urban scenarios within the LHC space. The standard errors for the GP exponential models are larger which is consistent with applying a GLS fit. The parameter estimates for emissions are always positive indicating that the higher the emissions, the higher the simulated NO_2 annual averages. On the other hand, the parameter estimates for emissions squared are always negative. The wind speed parameter estimates are also always negative suggesting that the higher the wind speed, the lower the simulated NO_2 annual averages. The interaction between emissions and wind speed is always negative

Station	Model	Covariates	In	Out
Burgher Street	LR	E+E^2+WS+E*WS+WD	0.15 (0.13, 0.17)	0.16 (0.14, 0.18)
	GP Exponential	E+E^2+WS+E*WS+WD	0.05 (0.04, 0.06)	0.06 (0.04, 0.07)
Byres Road	LR	E+E^2+WS+E*WS	0.16 (0.14, 0.18)	0.17 (0.15, 0.19)
	GP Exponential	E+E^2+WS+E*WS+WD	0.03 (0.03, 0.04)	0.04 (0.03, 0.04)
Central Station	LR	E+E^2+WS+E*WS	0.50 (0.43, 0.59)	0.53 (0.45, 0.62)
	GP Exponential	E+E^2+WS+E*WS+WD	0.09 (0.07, 0.12)	0.10 (0.07, 0.13)
Dumbarton Road	LR	E+E^2+WS+E*WS+WD	0.27 (0.23, 0.31)	0.29 (0.24, 0.33)
	GP Exponential	E+E^2+WS+E*WS+WD	0.08 (0.06, 0.09)	0.08 (0.07, 0.10)
Great Western Road	LR	E+E^2+WS+E*WS+WD	0.19 (0.16, 0.21)	0.20 (0.17, 0.20)
	GP Exponential	E+E^2+WS+E*WS+WD	0.05 (0.04, 0.06)	0.06 (0.05, 0.07)
High Street	LR	E+E^2+WS+E*WS+WD	0.19 (0.16, 0.21)	0.20 (0.17, 0.23)
	GP Exponential	E+E^2+WS+E*WS+WD	0.05 (0.04, 0.05)	0.05 (0.04, 0.05)
Townhead	LR	E+E^2+WS+E*WS+WD	0.16 (0.14, 0.18)	0.17 (0.14, 0.19)
	GP Exponential	E+E^2+WS+E*WS+WD	0.04 (0.03, 0.05)	0.04 (0.04, 0.05)
Waulkmillglen Reservoir	LR	E+E^2+WS+E*WS+WD	0.05 (0.04, 0.06)	0.05 (0.04, 0.06)
	GP Exponential	E+E^2+WS+E*WS+WD	0.02 (0.01, 0.02)	0.02 (0.01, 0.03)

TABLE 5.24: Comparing the predictive performance of the linear regression and the preferred GP model for predicting the ADMS-Urban simulation runs for the NO_2 annual average (μg m^{-3}) for all monitoring stations in Glasgow.

as the dispersion of emissions is dependant on the wind speed. The differences between the models come when looking at the wind direction parameter estimates. For all stations, except for Burgher Street, the estimate is negative. The difference comes from the fact that Burgher Street is the most eastern station in Glasgow. As the wind speed estimate is positive for seven of the monitoring stations, it suggests that the more western prevailing wind, the higher the simulated NO_2 annual average concentrations. For Burgher Street, the more western prevailing wind, the lower the simulated NO_2 annual average concentrations.

5.4 Discussion

Section 5.1 provided a theoretical background for fitting GP models in R using the **DiceKriging** package. The theory was then applied to the Aberdeen case study in

Station	Coef.	Estim. LR	St. Error LR	Estim. GP	St. Error GP
Burgher Street	Intercept	18.72	0.03	18.80	0.23
	Em	0.08	0.001	0.08	0.005
	Em^2	-0.0002	0.00001	-0.0002	0.00005
	WS	-0.08	0.002	-0.08	0.0001
	Em*WS	-0.0007	0.00004	-0.0008	0.00002
	WD	0.02	0.002	0.02	0.004
Byres Road	Intercept	34.06	0.03	34.27	0.19
	Em	0.20	0.001	0.20	0.002
	Em^2	-0.001	0.00002	-0.0006	0.00002
	WS	-0.20	0.002	-0.20	0.005
	Em*WS	-0.002	0.00004	-0.002	0.00006
	WD			-0.0007	0.001
Central Station	Intercept	63.19	0.08	64.04	0.38
	Em	0.35	0.004	0.36	0.01
	Em^2	-0.002	0.00005	-0.002	0.0001
	WS	-0.33	0.007	-0.33	0.01
	Em*WS	-0.003	0.0001	-0.003	0.0002
	WD			-0.002	0.003
Dumbarton Road	Intercept	37.65	0.04	37.82	0.24
	Em	0.23	0.002	0.23	0.005
	Em^2	-0.0006	0.00003	-0.0006	0.00005
	WS	-0.23	0.004	-0.22	0.007
	Em*WS	-0.002	0.00007	-0.002	0.00009
	WD	-0.05	0.003	-0.05	0.007

TABLE 5.25: Summary of the linear regression and the GP exponential parameters for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for four monitoring stations (Burgher Street, Byres Road, Central Station and Dumbarton Road) in Glasgow with emissions (% change), emissions squared, wind speed (% change), an interaction for emissions and wind speed, and wind direction (° change) as covariates.

Section 5.2. The pairs plots for each of the six monitoring stations in Aberdeen suggested a simple linear regression with just the three ADMS-Urban inputs (emissions, wind speed and wind direction) is sufficient as the diagnostic plots did not indicate any issues with the fit. When the linear regression models with just the three covariates were fitted, the models were parsimonious with $R^2_{adj.}$ values for all the models really high ($\sim 99\%$) and the diagnostic plots indicating a good fit. Furthermore, the GP models with just the three inputs were also identified as good fit for the data. The in-sample and 10-fold out-of-sample CV RMSPE (Table 5.7) indicated that the GP exponential kernel models are better than the linear regression ones as there is high correlation between the three covariates.

Based on the study in Section 5.2, the same modelling was applied to the eight monitoring stations in Glasgow in Section 5.3. It was found that the relationship between the simulated NO_2 annual average relationship with emissions is quadratic. Furthermore, all models required an interaction between emissions and wind speed to improve the fit

Station	Coef.	Estim. LR	St. Error LR	Estim. GP	St. Error GP
Great Western Road	Intercept	33.73	0.03	33.95	0.16
	Em	0.19	0.002	0.20	0.003
	Em^2	-0.0006	0.00002	-0.0006	0.00004
	WS	-0.19	0.003	-0.19	0.006
	Em*WS	-0.002	0.00005	-0.002	0.00007
	WD	-0.02	0.002	-0.02	0.004
High Street	Intercept	35.42	0.03	35.63	0.14
	Em	0.19	0.002	0.20	0.003
	Em^2	-0.0008	0.00002	-0.0008	0.00003
	WS	-0.20	0.003	-0.20	0.006
	Em*WS	-0.002	0.00005	-0.002	0.00007
	WD	-0.006	0.002	-0.01	0.002
Townhead	Intercept	29.37	0.03	29.50	0.12
	Em	0.14	0.001	0.15	0.003
	Em^2	-0.0007	0.00002	-0.0007	0.00003
	WS	-0.16	0.002	-0.16	0.005
	Em*WS	-0.001	0.00004	-0.002	0.00007
	WD	-0.006	0.002	-0.008	0.002
Waulkmillglen Reservoir	Intercept	9.75	0.08	9.81	0.04
	Em	0.01	0.0004	0.01	0.0007
	Em^2	-0.00003	0.000005	-0.00003	0.000007
	WS	-0.01	0.0007	-0.01	0.001
	Em*WS	-0.0002	0.00001	-0.0001	0.00001
	WD	-0.007	0.0006	-0.007	0.002

TABLE 5.26: Summary of the linear regression and the GP exponential parameters for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for four monitoring stations (Great Western Road, High Street, Townhead and Waulkmillglen Reservoir) in Glasgow with emissions (% change), emissions squared, wind speed (% change), an interaction for emissions and wind speed, and wind direction (° change) as covariates.

of the models based on the diagnostic plots. However, it was found that for the Byres Road and Central Station monitoring stations, wind direction is not significant. When the GP models were fitted, all stations retained the model with five covariates based on their predictive power. Additionally, an exponential kernel was chosen as best in terms of prediction power. The models for the Glasgow monitoring network were compared in terms of their predictive performance in Table 5.24 and it was found that the GP exponential kernel models are performing much better than the linear regression ones.

Overall, the models for Aberdeen are much simpler and suggest that the modelling is more straightforward in comparison to the Glasgow case. The prediction errors in the Aberdeen case study reduce about two times by applying the GP models, whereas the Glasgow models have more covariates which is the reason for the bigger predictive improvement (more than three times) for the GP models. The difference in the number of covariates for the two models could reflect the combination of two factors. Firstly,

Aberdeen and Glasgow have different geographical locations - eastern and western Scotland, respectively. Secondly, Glasgow has a much larger population and a more complex city outline. However, a relatively simple emulator for each of the monitoring station in Aberdeen and Glasgow was produced which produces predictions which are very precise. However, in order to examine the air pollution movement across each city, a multivariate model for all stations is necessary.

Chapter 6

Multivariate modelling and emulation of NO_2 annual average concentrations using Gaussian Processes

Chapter 5 presented the modelling of the ADMS-Urban simulated NO_2 annual average concentrations across the AURN monitoring stations in Aberdeen and Glasgow. It was found that the univariate (single response) Gaussian Process (GP) models have better prediction power based on in- and out-of-sample Root Mean Squared Prediction Error (RMSPE) in comparison to the linear regression models. In this chapter, the work from Chapter 5 is extended by introducing a multivariate (multiple responses) GP model in order to account for the correlation between the measurements at the different stations in one city. The aim is to investigate whether the prediction performance will be improved by utilising these correlations and whether using a multivariate model will result in time reduction for model fitting (in terms of a smaller number of models fitted). The rest of this chapter is organised as follows: Section 6.1 introduces the methodology of fitting a Bayesian multivariate GP model. A simulation study is conducted in Section 6.2 to check the performance of the multivariate emulator and also compare its ability to estimate the hyperspatial range parameters to the performance of the univariate frequentist emulator implemented via the **DiceKriging** software. Following the simulation study, an application of the proposed Bayesian multivariate GP model to the Aberdeen ADMS-Urban simulations is presented in Section 6.3. Section 6.4 presents the application of the proposed Bayesian multivariate GP emulator to the Glasgow ADMS-Urban simulations. Section 6.5 provides a concluding discussion.

6.1 Multivariate GP process

Chapter 5 presented univariate GP models using frequentist inference, specifically the BFGS algorithm, a quasi-Newton algorithm introduced in Subsection 2.2.4. However, more commonly a Bayesian approach to inference in this context is chosen as it provides a more "comprehensive and natural structure to represent and deal with uncertainty" [199], different information sources can be used, and it is possible to estimate probability distributions on the parameters of interest [140]. A study presented in [89] shows that the Bayesian paradigm provides more accurate prediction results in comparison to a frequentist approach in a limited simulation setting. Therefore, to perform multivariate modelling in this chapter, a multivariate GP emulator will be used with inference in the Bayesian paradigm. A multivariate Bayesian emulator was developed by Conti and O'Hagan in [49], which was then extended by Overstall and Woods in [143]. In this chapter, the work in [143] will be altered by applying the exponential correlation function based on the univariate models from Chapter 5.

6.1.1 Model definition and estimation

The model from Section 5.1 is extended to a multivariate response as follows. The data likelihood model is given by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{B} + \mathbf{Z}\,, \tag{6.1}$$

where:

- $\mathbf{Y}$ is the response matrix $(n \times q)$ containing the results from n scenarios for q locations (in this case stations) with each row $\mathbf{y}_i$ $(i = 1, \ldots, n)$ the output for a set of inputs at all stations;
- $\mathbf{X}$ is the design matrix $(n \times p)$ containing the intercept and the covariates, where each row $\mathbf{x}_i$ contains the intercept and inputs for a single scenario and is the same for all stations;
- $\boldsymbol{B}$ is the parameter matrix $(p \times q)$ containing the fixed effect parameters (different fixed effect parameters for each of the monitoring stations), which need to be estimated; and
- $\mathbf{Z}$ is the error matrix $(n \times q)$ that has a matrix normal distribution (presented in Appendix A) $\mathbf{Z} \sim \mathbf{MN}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{R}(\boldsymbol{\theta}))$.

Thus, the response matrix also follows a matrix normal distribution:

$$\mathbf{Y}|\boldsymbol{B},\boldsymbol{\Sigma},\mathbf{R}(\boldsymbol{\theta}) \sim \mathbf{MN}(\mathbf{X}\boldsymbol{B},\boldsymbol{\Sigma},\mathbf{R}(\boldsymbol{\theta}))\,, \tag{6.2}$$

where $\boldsymbol{\Sigma}$ is a positive definite column scaling matrix $(q \times q)$ which models the correlation between the q different stations. $\boldsymbol{\Sigma}$ is a matrix of free form because ADMS-Urban uses a factor to vary the predictions at different locations in the city, not the actual spatial positions of the stations. $\mathbf{R}(\boldsymbol{\theta})$ is a positive definite row scaling matrix $(n \times n)$, modelling the spatial correlation in the input space. Since it is assumed that the output $\mathbf{y}_i$ for all $i = 1,\ldots,n$ have the same uncertainty around them, $\mathbf{R}(\boldsymbol{\theta})$ is specified as a correlation matrix in [143] and in this work, $\mathbf{R}(\boldsymbol{\theta})$ is modelled using the exponential function because as seen in Chapter 5, all univariate frequentist models for both the Aberdeen and Glasgow monitoring stations use the exponential function for best forecasting results. Therefore, $\mathbf{R}(\boldsymbol{\theta})$ has a structure dependency on the hyperspatial range parameters vector $\boldsymbol{\theta}$ $(d \times 1)$, where d is the dimension of the input space. As previously shown in Chapter 4, the Latin Hypercube (LHC) space for both Aberdeen and Glasgow is based on varying three inputs (emissions (E), wind speed (WS) and wind direction (WD)). Hence, the parameter vector $\boldsymbol{\theta} = [\theta_{\mathrm{E}}, \theta_{\mathrm{WS}}, \theta_{\mathrm{WD}}]^\top$ has dimension $d = 3$. The exponential correlation function between two rows of the input space (designed by the LHC) $\mathbf{u}_i = [u_{i\mathrm{E}}, u_{i\mathrm{WS}}, u_{i\mathrm{WD}}]^\top$ and $\mathbf{u}_j = [u_{j\mathrm{E}}, u_{j\mathrm{WS}}, u_{j\mathrm{WD}}]^\top$ is:

$$\mathbf{R}_{ij}(\boldsymbol{\theta}) = \exp\left(-\left\{\left(\frac{|u_{i\mathrm{E}} - u_{j\mathrm{E}}|}{\theta_{\mathrm{E}}}\right) + \left(\frac{|u_{i\mathrm{WS}} - u_{j\mathrm{WS}}|}{\theta_{\mathrm{WS}}}\right) + \left(\frac{|u_{i\mathrm{WD}} - u_{j\mathrm{WD}}|}{\theta_{\mathrm{WD}}}\right)\right\}\right). \tag{6.3}$$

The distribution of the response matrix can be re-written for a vector of the stacked (by column) responses as:

$$\mathrm{vec}(\mathbf{Y})|\boldsymbol{B},\boldsymbol{\Sigma},\mathbf{R}(\boldsymbol{\theta}) \sim \mathbf{N}(\mathrm{vec}(\mathbf{X}\boldsymbol{B}),\boldsymbol{\Sigma} \otimes \mathbf{R}(\boldsymbol{\theta}))\,, \tag{6.4}$$

where $\mathrm{vec}(\cdot)$ denotes the column stacking of a matrix into a vector and $\otimes$ denotes the Kronecker product (an element by matrix multiplication), hence the covariance matrix $\boldsymbol{\Sigma} \otimes \mathbf{R}(\boldsymbol{\theta})$ is of size $qn \times qn$. The set of parameters $(\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$ are assigned a joint prior distribution, where the pair $(\boldsymbol{B}, \boldsymbol{\Sigma})$ and $\boldsymbol{\theta}$ are assumed to be independent:

$$f(\boldsymbol{B},\boldsymbol{\Sigma},\boldsymbol{\theta}) = f(\boldsymbol{B},\boldsymbol{\Sigma})f(\boldsymbol{\theta}) = f(\boldsymbol{B}|\boldsymbol{\Sigma})f(\boldsymbol{\Sigma})f(\boldsymbol{\theta})\,. \tag{6.5}$$

The prior for $(\boldsymbol{B}, \boldsymbol{\Sigma})$ is specified by a Matrix-Normal distribution and an inverse-Wishart prior distribution [143] respectively, which are given below:

$$\boldsymbol{B}|\boldsymbol{\Sigma},\boldsymbol{\theta} \sim \mathbf{MN}(\mathbf{M},\boldsymbol{\Sigma},\boldsymbol{\Omega})\,, \tag{6.6}$$

$$\boldsymbol{\Sigma}|\boldsymbol{\theta} \sim \mathbf{IW}(\mathbf{S}^{-1},\delta)\,, \tag{6.7}$$

where $\mathbf{M}, \mathbf{\Omega}, \mathbf{S}$ and δ are hyperparameters with non-informative specifications defined as follows:

- for $\mathbf{M}$, the prior is a $p \times q$ matrix of zeros;
- for $\mathbf{\Omega}$, the prior is a diagonal matrix $(p \times p)$ with entries of 100 000 on the diagonal;
- for $\mathbf{S}$, the prior is an identity matrix $\mathbb{I}_q$ $(q \times q)$ in order to have an uninformative prior as recommended in [171]; and
- for δ, the prior is set to the constant q.

The prior for the hyperspatial range parameter $\boldsymbol{\theta}$ is decomposed as:

$$f(\boldsymbol{\theta}) = f(\theta_{\mathrm{E}}) f(\theta_{\mathrm{WS}}) f(\theta_{\mathrm{WD}}) \,, \tag{6.8}$$

where $f(\theta_{\mathrm{E}}) \propto 1$, $f(\theta_{\mathrm{WS}}) \propto 1$, and $f(\theta_{\mathrm{WD}}) \propto 1$ are non-informative improper flat priors on the positive real line. A plug-in numerical approximation (the BFGS algorithm as described in Subsection 2.2.4 and applied in Chapter 5) is used for the estimation of $\boldsymbol{\theta}$ as an MCMC approach is "computationally cumbersome" since it requires inverting $\mathbf{R}(\boldsymbol{\theta})$ and calculating its derivative at each step of the MCMC algorithm. This approach is recommended in both [49] and [143]. The BFGS algorithm is used to maximise the unnormalised marginal posterior density (derived in [144]):

$$f(\boldsymbol{\theta}|\mathbf{Y}) \propto |\mathbf{R}(\boldsymbol{\theta})|^{-\frac{q}{2}} |\mathbf{D}|^{\frac{q}{2}} |\mathbf{Y}^\top \mathbf{R}(\boldsymbol{\theta})^{-1} \mathbf{Y} - \mathbf{E}^\top \mathbf{D}^{-1} \mathbf{E}|^{-\frac{(\delta+n+q-1)}{2}} \,, \tag{6.9}$$

where:

- $\mathbf{D} = (\mathbf{X}^\top \mathbf{R}(\boldsymbol{\theta})^{-1} \mathbf{X})^{-1}$; and
- $\mathbf{E} = \mathbf{D}(\mathbf{X}^\top \mathbf{R}(\boldsymbol{\theta})^{-1} \mathbf{Y})$.

After $\boldsymbol{\theta}$ is estimated by $\widehat{\boldsymbol{\theta}}$, the posterior distributions of $(\boldsymbol{B}, \mathbf{\Sigma})$ have the following closed form solutions:

$$\boldsymbol{B}|\mathbf{Y}, \mathbf{\Sigma}, \widehat{\boldsymbol{\theta}} \sim \mathbf{MN}(\mathbf{M}_{\boldsymbol{B}}, \mathbf{\Sigma}, \mathbf{\Psi}) \,, \tag{6.10}$$

$$\mathbf{\Sigma}|\mathbf{Y}, \widehat{\boldsymbol{\theta}} \sim \mathbf{IW}(\mathbf{\Xi}^{-1}, \upsilon) \,. \tag{6.11}$$

The updated parameters have closed form expressions as follows:

- $\mathbf{M}_{\boldsymbol{B}} = \left(\mathbf{X}^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{Y} + \mathbf{\Omega}^{-1} \mathbf{M}$;

- $\boldsymbol{\Psi} = \left(\mathbf{X}^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X} + \boldsymbol{\Omega}^{-1}\right)^{-1}$;
- $\boldsymbol{\Xi} = \mathbf{Y}^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{Y} + \mathbf{M}^\top \boldsymbol{\Omega}^{-1}\mathbf{M} + \mathbf{S} - \mathbf{M}_{\boldsymbol{B}}^\top \boldsymbol{\Psi}^{-1}\mathbf{M}_{\boldsymbol{B}}$; and
- $\upsilon = \delta + n$.

There are closed form solutions for posterior means of the pair ($\boldsymbol{B}$, $\boldsymbol{\Sigma}$), which are:

$$\widehat{\boldsymbol{B}} = \mathbf{M}_{\boldsymbol{B}}\,, \tag{6.12}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{\upsilon - q - 1}\boldsymbol{\Xi}\,. \tag{6.13}$$

6.1.2 Prediction of new observations

The main aim in this chapter is prediction of the NO_2 annual average concentrations at unobserved (untested) sets of inputs for the ADMS-Urban simulator. Therefore, how to predict using the multivariate model is described. Let $\mathbf{Y}_0$ ($n_0 \times q$) be a matrix of new outputs at the same q stations. Then $\mathbf{Y}_0$ has the following joint distribution with $\mathbf{Y}$:

$$\begin{pmatrix}\mathbf{Y}_0\\ \mathbf{Y}\end{pmatrix} \sim \mathbf{MN}\left(\begin{pmatrix}\mathbf{X}_0\\ \mathbf{X}\end{pmatrix}\boldsymbol{B}, \widehat{\boldsymbol{\Sigma}}, \begin{pmatrix}\mathbf{R}_0(\widehat{\boldsymbol{\theta}}) & \mathbf{T}(\widehat{\boldsymbol{\theta}})^\top\\ \mathbf{T}(\widehat{\boldsymbol{\theta}}) & \mathbf{R}(\widehat{\boldsymbol{\theta}})\end{pmatrix}\right)\,, \tag{6.14}$$

where $\mathbf{X}_0$ is the design matrix ($n_0 \times p$) for the untested sets of inputs (i.e. $\mathbf{X}_0$ contains the new set of covariates, which are based on the locations of interest defined by the untested sets of inputs within the LHC space), $\mathbf{R}_0(\widehat{\boldsymbol{\theta}})$ is the correlation matrix ($n_0 \times n_0$) between the new sets of inputs, and $\mathbf{T}(\widehat{\boldsymbol{\theta}})$ is the correlation matrix ($n \times n_0$) between the observed and unobserved sets of simulation inputs. $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\theta}}$ are the estimates to which $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ are respectively fixed. Therefore, the conditional distribution of a matrix of new observations follows the matrix t-distribution (described in Appendix A) as the matrix $\widehat{\boldsymbol{\Sigma}}$ is estimated. Therefore, the mean and variance of the new observations are estimated in a similar fashion to kriging (Subsection 2.3.1):

$$\mathbf{Y}_0|\mathbf{Y}, \boldsymbol{\theta} \sim \mathbf{MT}(\mathbf{Q}, \widehat{\boldsymbol{\Sigma}}, \mathbf{A}(\widehat{\boldsymbol{\theta}}), \upsilon)\,, \tag{6.15}$$

where:

- $\mathbf{Q}$ is called the location matrix ($n_0 \times q$) and contains the predicted values of the new observations, which are estimated as:

$$\mathbf{Q} = \mathbf{X}_0\boldsymbol{B} + \mathbf{T}(\widehat{\boldsymbol{\theta}})^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{B})\,; \text{and} \tag{6.16}$$

- $\mathbf{A}(\widehat{\boldsymbol{\theta}})$ is a row scale matrix $(n_0 \times n_0)$, and contains the variance at the new observations:

$$\begin{aligned}\mathbf{A}(\widehat{\boldsymbol{\theta}}) = \mathbf{R}_0(\widehat{\boldsymbol{\theta}}) - \mathbf{T}(\widehat{\boldsymbol{\theta}})^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{T}(\widehat{\boldsymbol{\theta}}) + \\ + (\mathbf{X}_0 - \mathbf{T}(\widehat{\boldsymbol{\theta}})^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{X}) \boldsymbol{\Psi} (\mathbf{X}_0 - \mathbf{T}(\widehat{\boldsymbol{\theta}})^\top \mathbf{R}(\widehat{\boldsymbol{\theta}})^{-1} \mathbf{X})^\top .\end{aligned} \tag{6.17}$$

In general, the matrix t-distribution of $\mathbf{Y}_0$ given $\mathbf{Y}, \boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ can be re-written in vector form as follows:

$$\text{vec}(\mathbf{Y}_0) | \mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\theta} \sim \text{t}(\text{vec}(\mathbf{Q}), \boldsymbol{\Sigma} \otimes \mathbf{A}(\boldsymbol{\theta})) . \tag{6.18}$$

Therefore, the estimated variance-covariance matrix of $\text{vec}(\mathbf{Y}_0)$ is the Kronecker product of $\widehat{\boldsymbol{\Sigma}}$ and $\mathbf{A}(\widehat{\boldsymbol{\theta}})$.

6.2 Simulation studies

In Chapter 5, each univariate frequentist model produced for both Aberdeen and Glasgow estimates a set of the hyperspatial range parameters $\boldsymbol{\theta}$ associated with the input space designed using a LHC for each of the monitoring stations. However, there was a large variation between the hyperspatial range parameters estimates for each of the stations in Chapter 5. Most notably, the univariate model for Union Street in Aberdeen provided the hyperspatial range parameter estimates as $\widehat{\boldsymbol{\theta}} = [49.40, 7.99, 2934.23]^\top$. This suggests that the univariate models struggled with estimating at least one (the wind direction) of the hyperspatial range parameters. Since the multivariate emulator imposes the same hyperspatial range parameters for all monitoring stations, it is beneficial to further investigate this issue before applying the multivariate emulator to the Aberdeen and Glasgow ADMS-Urban simulations. Cressie states that hyperspatial range parameters are difficult to estimate but a "sensible" estimate would allow for relatively unaffected forecasting [50]. This is confirmed by Zhang [212], who showed that the hyperspatial range parameters cannot be estimated consistently for input spaces with dimensions $d \geq 3$. In order to test the effect of estimating the hyperspatial range parameters in a similar setting to the data previously presented, two simulation studies will be performed. The data for studies will be closely based on the Aberdeen case study. The simulation studies will:

(i) compare the ability of the models to correctly identify the hyperspatial range parameters in a setting very close to the real-life data.

(ii) assess the loss of predictive power to mis-estimating the hyperspatial range parameters.

6.2.1 Simulation study 1: Estimating the hyperspatial range parameters

Initialising the simulation

The first simulation study is a multiple response example based on two of the monitoring stations in Aberdeen (Market Street 2 and Wellington Road) as the two stations have very similar parameter estimates for the univariate case as seen in Subsection 5.2.3. The simulations are generated using the following parameters:

- the response matrix $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]^\top$ (98×2) is the simulated responses for two monitoring stations with the vector $\mathbf{Y}_1$ (98×1) being based on Market Street 2 and the vector $\mathbf{Y}_2$ (98×1) being based on Wellington Road;
- the input matrix $\mathbf{X}$ (98×3) is the LHC design used for the Aberdeen case study and described in Subsection 4.1.3;
- there are three fixed effect parameters to resemble the models in Chapter 5. The parameters and the intercept are chosen to be the values estimated by applying the model from the **DiceKriging** software. Hence, the fixed effect parameter matrix $\boldsymbol{B}$ (4×2) is $\begin{bmatrix} 47.33 & 43.84 \\ 0.22 & 0.19 \\ -0.17 & -0.17 \\ 0.06 & 0.07 \end{bmatrix}$;
- the variance-covariance matrix between the outputs mimics the real-life situation and is set to be $\boldsymbol{\Sigma} = \begin{bmatrix} 0.40 & 0.35 \\ 0.35 & 0.39 \end{bmatrix}$. The variances are also set to the values estimated by applying the model from the **DiceKriging** software, whereas the covariances were chosen to ensure high correlation between the two responses (Market Street 2 and Wellington Road). In order to assess the effects of larger variances on the estimates, a second variance-covariance matrix $\boldsymbol{\Sigma}_{\text{high}}$ is used, where the values are ten times larger than the ones in $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{\text{high}} = \begin{bmatrix} 4.00 & 3.50 \\ 3.50 & 3.90 \end{bmatrix}$;
- the correlation matrix $\mathbf{R}(\boldsymbol{\theta})$ is estimated based on $\boldsymbol{\theta}$ being set to be a vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^\top$. The values for $\boldsymbol{\theta}$ were chosen based on the individual variograms for each of the inputs in Figures 4.8, 4.9 and 4.10. Since some of the variograms never plateau and hence, never suggest range values, for baseline values of the hyperspatial range parameters is chosen the set $\boldsymbol{\theta} = [30, 15, 10]^\top$. In order to assess how well estimated the different spans of the hyperspatial range parameters are,

four additional sets for the hyperspatial range parameters are examined. Hence, there are five sets of parameter vectors in total:

$$\begin{aligned}
\boldsymbol{\theta}_1 &= [10, 5, 3.3]^\top, \\
\boldsymbol{\theta}_2 &= [15, 7.5, 5]^\top, \\
\boldsymbol{\theta}_3 &= [30, 15, 10]^\top, \\
\boldsymbol{\theta}_4 &= [60, 30, 20]^\top, \\
\boldsymbol{\theta}_5 &= [90, 45, 30]^\top; \text{and}
\end{aligned}$$

- the error matrix $\mathbf{Z}$ (98×2) is randomly drawn from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma} \otimes \mathbf{R}(\boldsymbol{\theta})$.

Therefore, under 10 different sets of parameters (five sets of hyperspatial range parameters $\boldsymbol{\theta}$ and two variance-covariances matrices $\boldsymbol{\Sigma}$), one hundred simulations are generated under each set of parameters resulting in 1000 simulated data sets in total.

Results

For each set of 100 simulations under a given set of hyperspatial range parameters, three different emulator models will be fitted and used to estimate the hyperspatial range parameters. The three models which will be used are: the frequentist univariate model fitted using the **DiceKriging** software in **R**, which was previously presented in Section 5.1; a univariate simplification of the multivariate Bayesian emulator proposed in this chapter; and the full multivariate Bayesian emulator. The Bayesian models are coded in **R** based on the description in Section 6.1. Boxplots of the parameter estimates will be presented. This would allow the comparison not only between the single and multiple response models but also between the frequentist and Bayesian paradigms. The estimated hyperspatial range parameters vary from 0 to 1000. Hence, when necessary, granulated (zoomed in) plots are also provided. The estimates from the frequentist model are point estimates of the hyperspatial range parameters, whereas the estimates from the Bayesian models are the posterior means. This is based on a similar approach comparing frequentist and Bayesian emulators applied in [89]. The estimates from the different models will be abbreviated as follows:

- **dk1** for the estimates from the emulator in the **DiceKriging** software when applied to the $\mathbf{Y}_1$ data;
- **dk2** for the estimates from the emulator in the **DiceKriging** software when applied to the $\mathbf{Y}_2$ data;

- **mult** for the estimates from the **Multivariate** Bayesian method when applied to the $\mathbf{Y}$ data;
- **uni1** for the estimates from the **Univariate** Bayesian method when applied to the $\mathbf{Y}_1$ data; and
- **uni2** for the estimates from the **Univariate** Bayesian method when applied to the $\mathbf{Y}_2$ data.

Firstly, the estimates for the different hyperspatial range parameters under the variance-covariance matrix $\boldsymbol{\Sigma}$ are presented. For the first set of hyperspatial range parameters $\boldsymbol{\theta}_1 = [10, 5, 3.3]^\top$, the boxplots for each hyperspatial range parameters' estimates are presented in Figures 6.1, 6.2 and 6.3 respectively. The medians for the parameter estimates are only close to the true values for the Bayesian models. When the frequentist paradigm is applied, the true values of the first and third hyperspatial range parameters (Figures 6.1 and 6.3) are contained within the interquartile ranges. The boxes for these two hyperspatial range parameters as estimated by the Bayesian models have smaller interquartile ranges than the frequentist model from the **DiceKriging** software which indicates that there is less variability in their estimates. However, for the second hyperspatial range parameter (Figure 6.2), it appears that the parameter is estimated in the frequentist paradigm as zero almost every time. Granulated versions of the boxplots are required for all three hyperspatial range parameters as both univariate models struggle more with the estimation of the hyperspatial range parameters as there are very large outliers present in their estimates. Additionally, the multivariate Bayesian model has smaller interquartile ranges than the univariate Bayesian models. It is important to note that the Bayesian models do not estimate any of the hyperspatial range parameters as zero, although the parameters in this set are very small.

The boxplots for the estimates for the second set of hyperspatial range parameters $\boldsymbol{\theta}_2 = [15, 7.5, 5]^\top$ are very similar to those for the first set of hyperspatial range parameters in Figures 6.1, 6.2 and 6.3. Therefore, these plots are omitted in order to avoid repetition.

The boxplots for the estimates for the third set of hyperspatial range parameters $\boldsymbol{\theta}_3 = [30, 15, 10]^\top$ are presented in Figures 6.4, 6.5 and 6.6 respectively. The medians for the parameter estimates for all the models are only close to the true values for the first hyperspatial range parameter (Figure 6.4), whereas the second (Figure 6.5) and third (Figure 6.6) hyperspatial range parameters are larger than the median estimates for all the models and barely contained within the interquartile ranges. From the fact that granulated versions are required for all the boxplots, it is clear that the univariate models continue to struggle with the estimation of the hyperspatial range parameters as there are very large outliers. The boxes for the Bayesian models have smaller interquartile

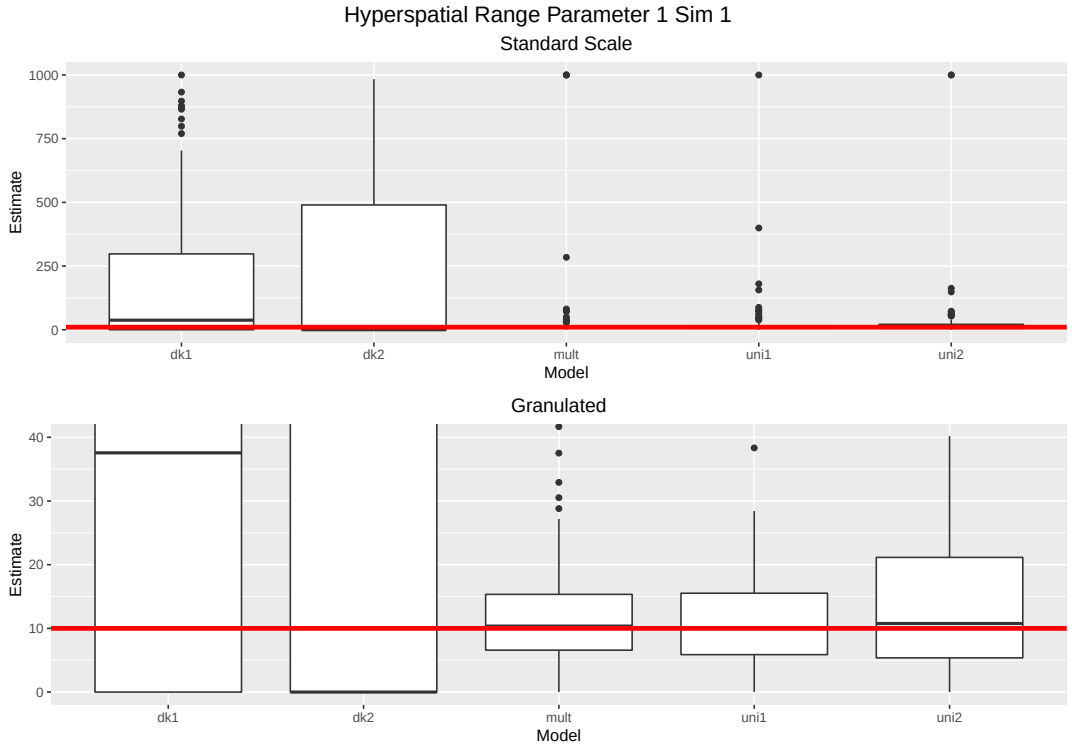

FIGURE 6.1: Boxplots for the first hyperspatial range parameter in the first set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

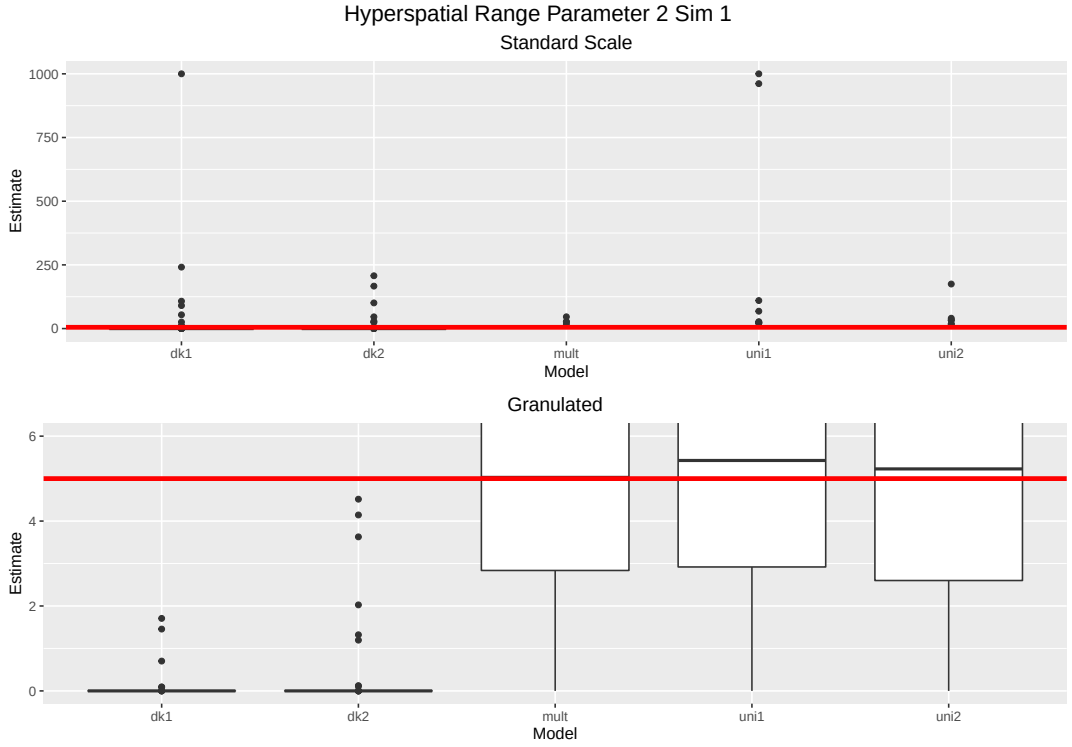

FIGURE 6.2: Boxplots for the second hyperspatial range parameter in the first set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

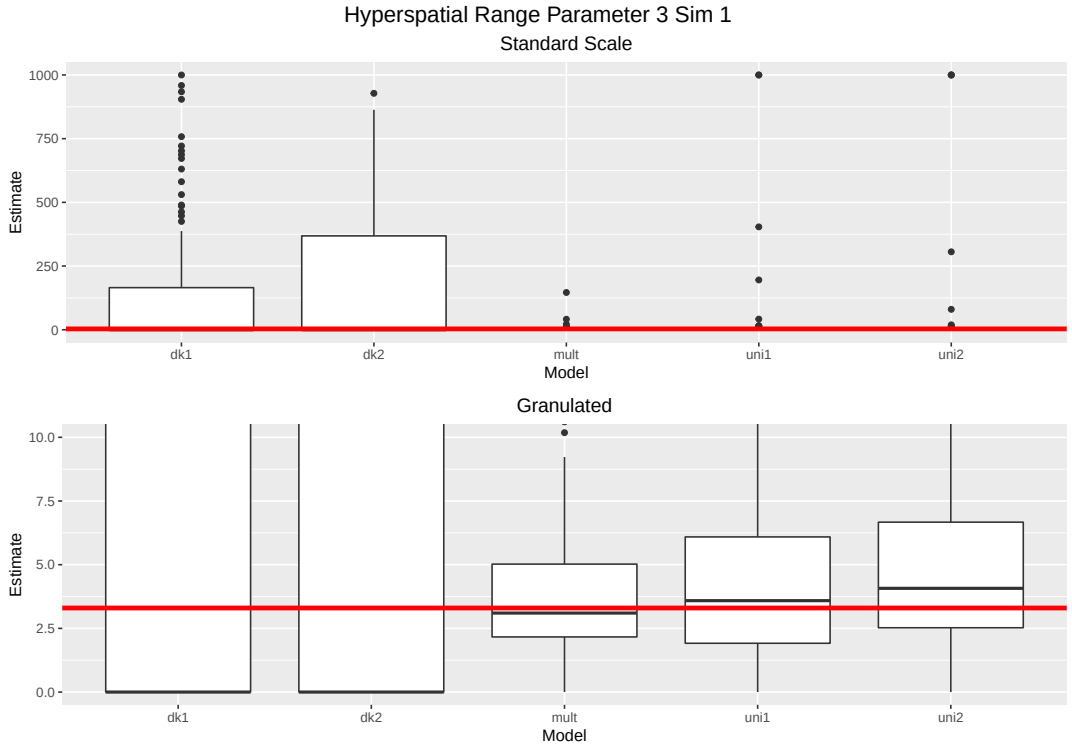

FIGURE 6.3: Boxplots for the third hyperspatial range parameter in the first set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

ranges than the frequentist models from the **DiceKriging** software which indicates that there is less variability in their estimates. Furthermore, the multivariate model has smaller interquartile ranges than the univariate Bayesian models. It is important to note that the multivariate model is the only one that has never estimated the hyperspatial range parameters as zero, although the second and third hyperspatial range parameters are close to zero.

The boxplots for the estimates for the fourth set of hyperspatial range parameters $\boldsymbol{\theta}_4 = [60, 30, 20]^\top$ and the fifth set of hyperspatial range parameters $\boldsymbol{\theta}_5 = [90, 45, 30]^\top$ are also very similar to each other. Therefore, to avoid repetition, only the boxplots for the fifth set of hyperspatial range parameter estimates in Figures 6.7, 6.8 and 6.9 respectively are provided. None of the boxplots requires granulation as the values of the estimated hyperspatial range parameters are larger. For the first hyperspatial range parameter in Figure 6.7, all models appear to underestimate the hyperspatial range parameter as all the medians are below the true value line (in red). The multivariate model is the only one that has no outliers but seems to be struggling most by underestimating the hyperspatial range parameter as the true value is almost as large as the maximum value estimated by the model. It is interesting to note that for both the univariate frequentist and Bayesian approaches, the models for $\mathbf{Y}_1$ are the only ones to include the true value in their interquartile ranges. The situation for the second (Figure 6.8) and third (Figure

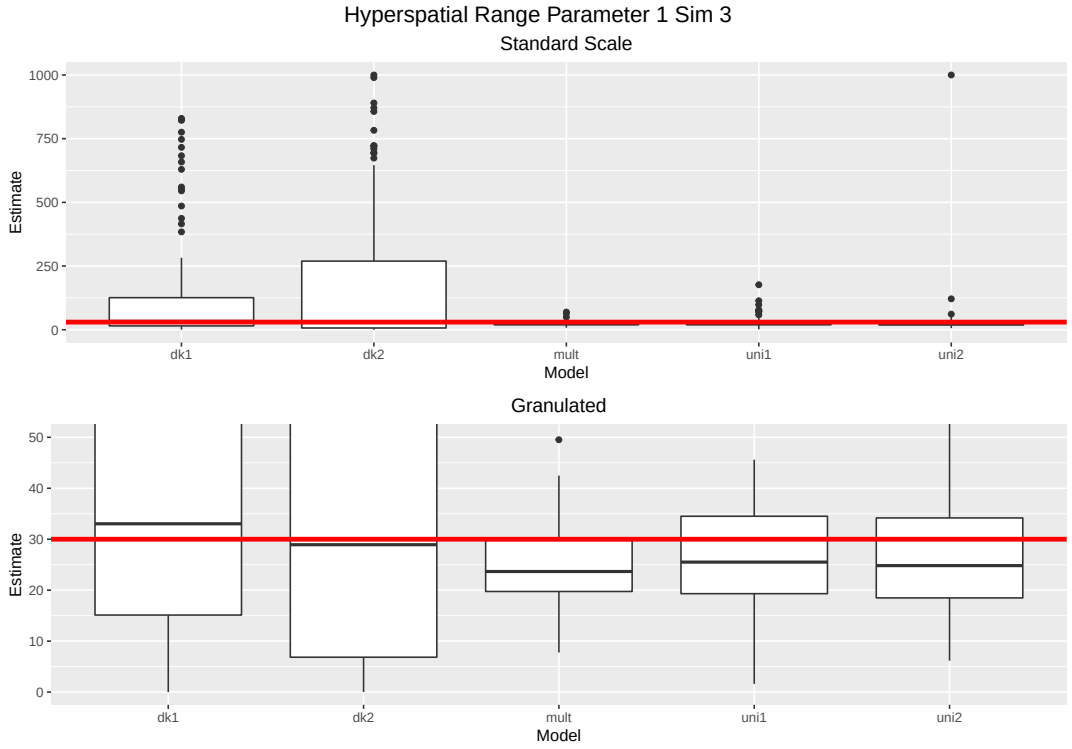

FIGURE 6.4: Boxplots for the first hyperspatial range parameter in the third set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

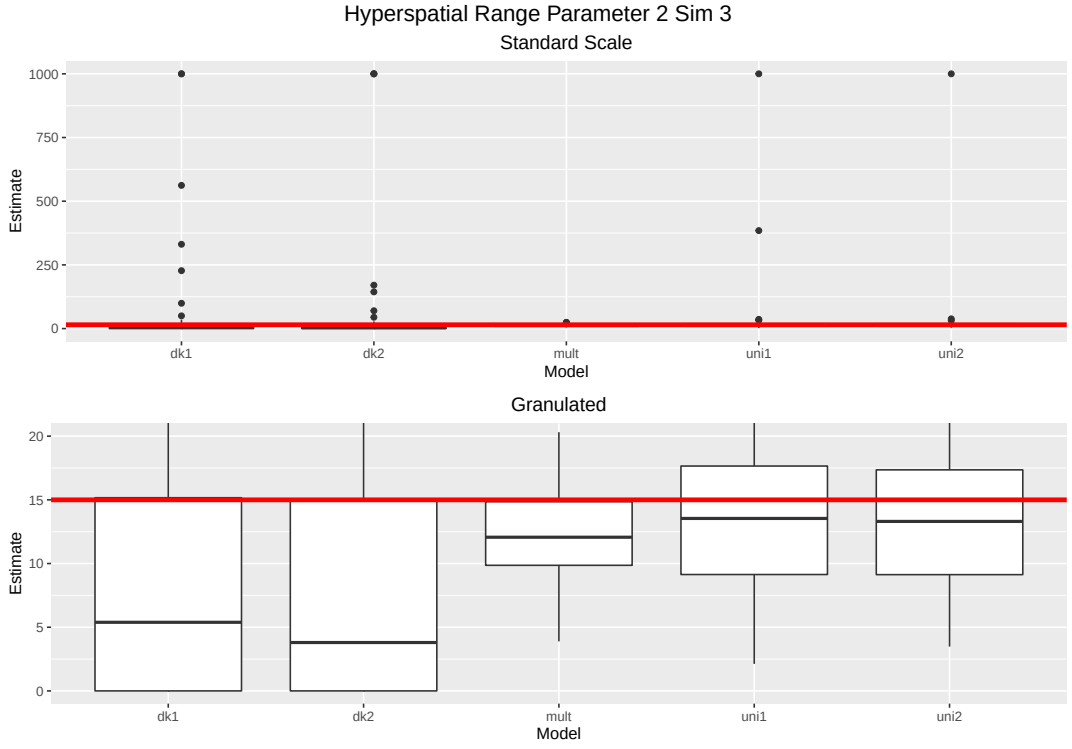

FIGURE 6.5: Boxplots for the second hyperspatial range parameter in the third set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

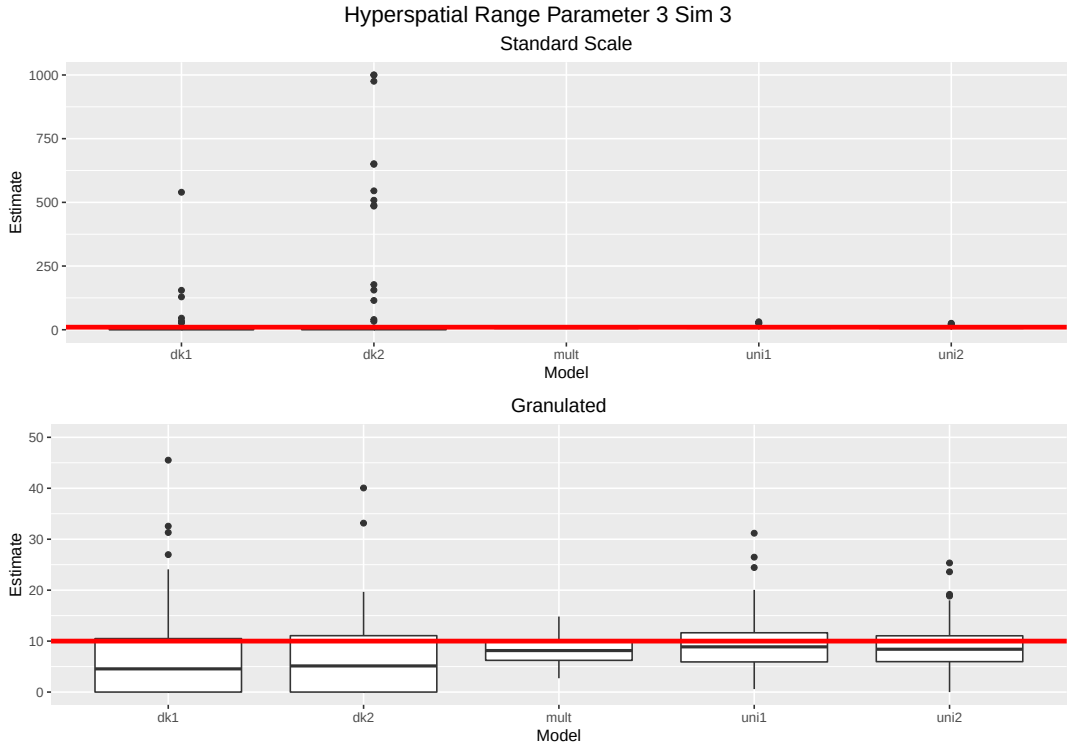

FIGURE 6.6: Boxplots for the third hyperspatial range parameter in the third set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

6.9) hyperspatial range parameters is very similar to the one for the first hyperspatial range parameter with all the models underestimating the true value of the parameter. However, it is apparent that the only model that does not estimate extreme values for the hyperspatial range parameters is the multivariate Bayesian one.

The results from the high variance-covariance matrix $\boldsymbol{\Sigma}_{\text{high}}$ simulations are very similar to the ones from $\boldsymbol{\Sigma}$ and as the size of the hyperspatial range parameter is increased, all models tend to underestimate the true values of the hyperspatial range parameters but the multivariate model remains the only one that does not estimate extreme values. Therefore, the results from the $\boldsymbol{\Sigma}_{\text{high}}$ simulations are omitted for brevity.

Findings

Simulation Study 1 compared the estimated hyperspatial range parameters for three different models (univariate frequentist, univariate Bayesian and multivariate Bayesian). It was found that all models struggle similarly with estimating the hyperspatial range parameters $\boldsymbol{\theta}$ for both variance-covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\text{high}}$. As the values of the hyperspatial range parameters increase, all models tend to underestimate the hyperspatial range parameters as demonstrated by the fact that the median values of estimated hyperspatial range parameters are lower than the true values of the hyperspatial range

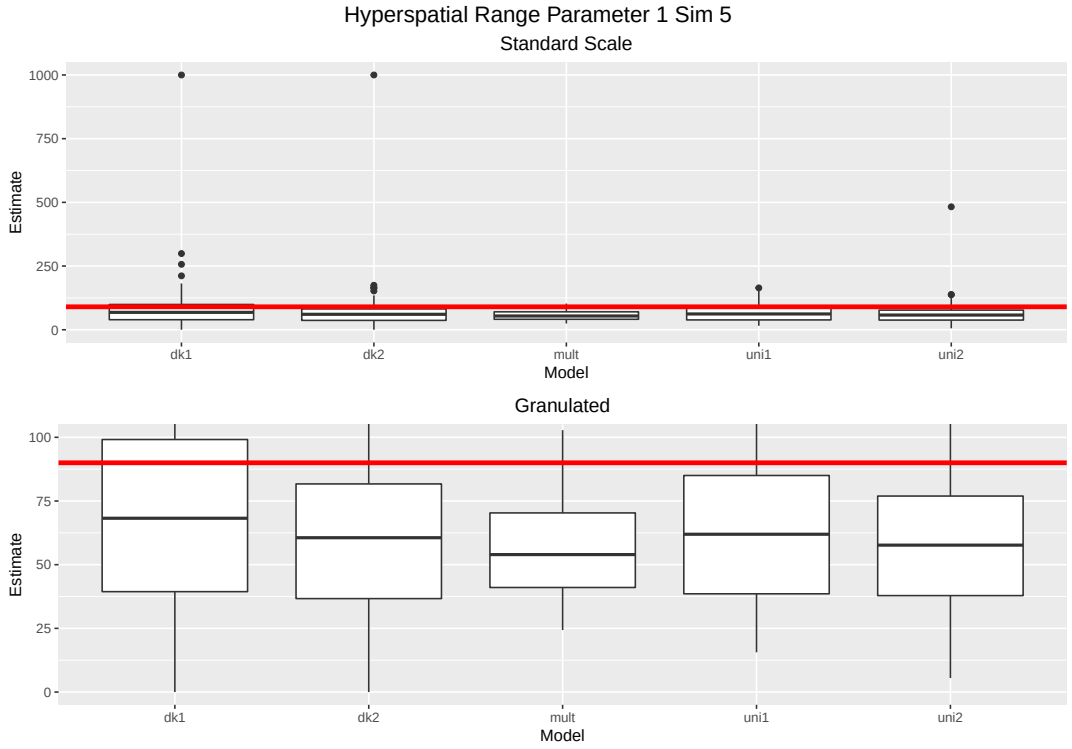

FIGURE 6.7: Boxplots for the first hyperspatial range parameter in the fifth set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

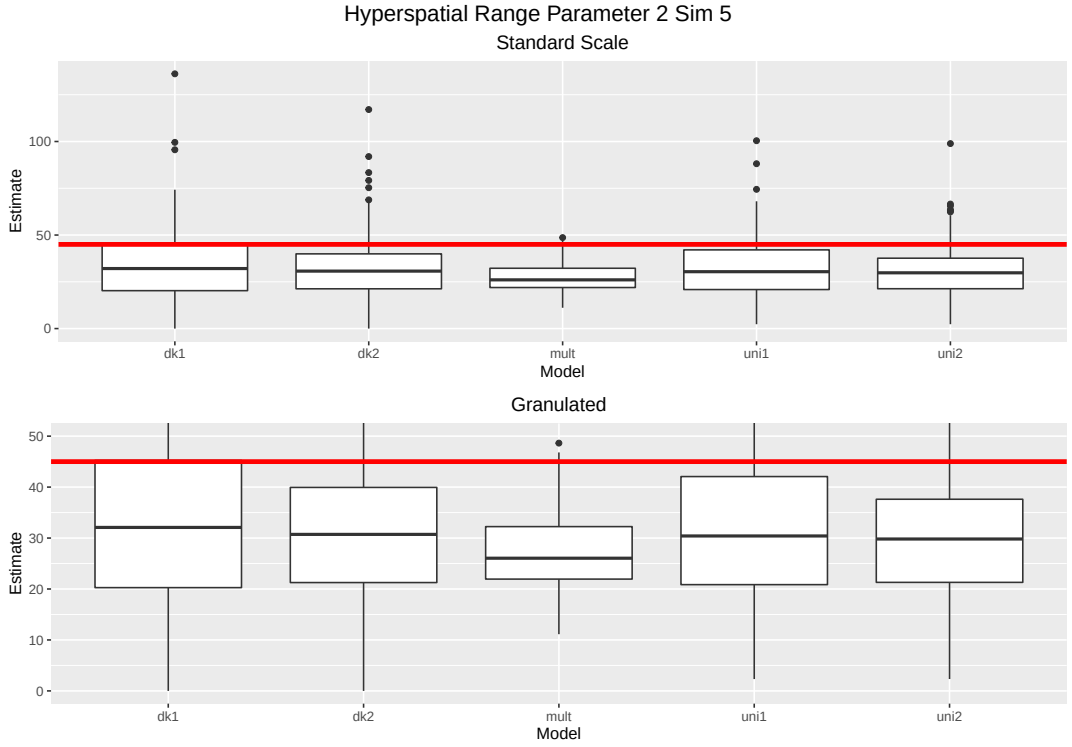

FIGURE 6.8: Boxplots for the second hyperspatial range parameter in the fifth set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

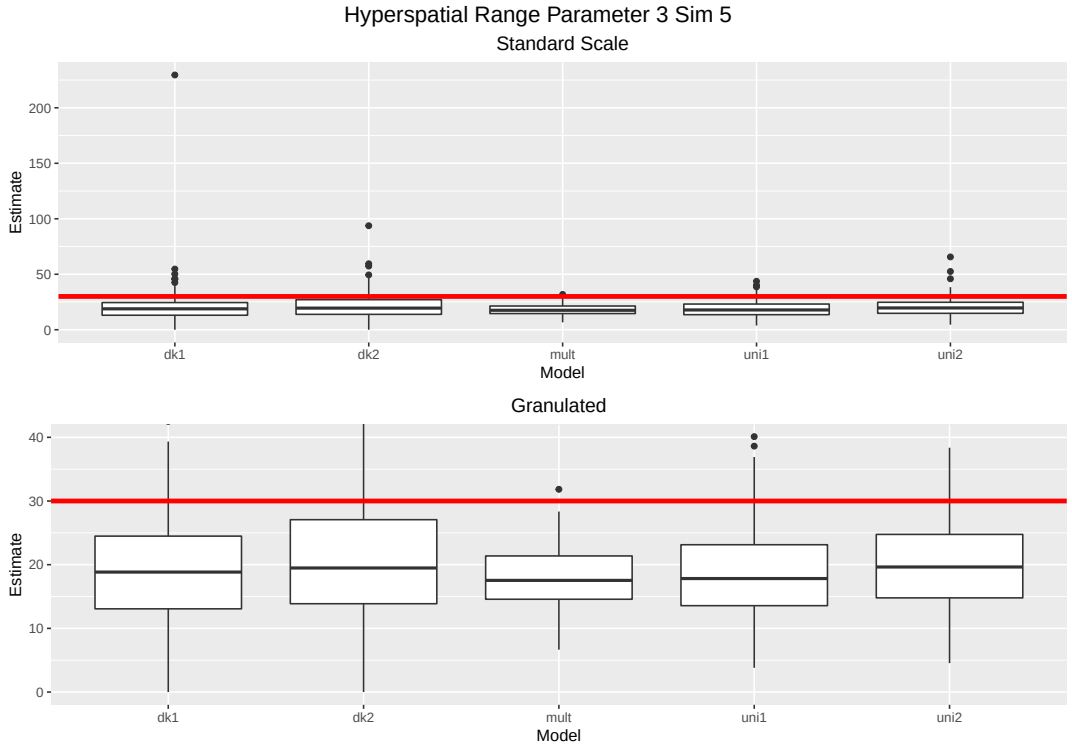

FIGURE 6.9: Boxplots for the third hyperspatial range parameter in the fifth set of simulations under $\boldsymbol{\Sigma}$ as estimated by the different modelling techniques. The red line is the true parameter value.

parameters. The two univariate models have very similar performances with the estimates from the frequentist model having a slightly larger spread than those from the Bayesian one. Although the multivariate Bayesian model has the smallest spread for its estimates of the hyperspatial range parameters, the multivariate model underestimates the true values of the hyperspatial range parameters in comparison to the univariate models.

6.2.2 Simulation study 2: Effect of mis-estimating the hyperspatial range parameters on the prediction quality

Initialising the simulation

The data for the second study were simulated in the same manner as the data for Simulation Study 1 in Subsection 6.2.1. To avoid repetition, the simulation set-up will not be repeated here.

Results

In order to compare the forecasting capabilities of the different models under different sets of hyperspatial range parameter, RMSPE will be computed from a 10-fold cross-validation (CV) analysis in the same way as was done in Chapter 5. As with the hyperspatial range parameters, the results from the three models (the univariate frequentist model from **DiceKriging** software, Univariate Bayesian model and Multivariate Bayesian model) for each of the responses $\mathbf{Y}_1$ and $\mathbf{Y}_2$ will be compared for the two different variance-covariance matrices in Tables 6.1 and 6.2, respectively. The one hundred simulated data sets generated for five different sets of hyperspatial range parameters will be used again. For clarity, these sets will be referred to as:

- Sim 1 for the simulations with parameter vector $\boldsymbol{\theta}_1$;
- Sim 2 for the simulations with parameter vector $\boldsymbol{\theta}_2$;
- Sim 3 for the simulations with parameter vector $\boldsymbol{\theta}_3$;
- Sim 4 for the simulations with parameter vector $\boldsymbol{\theta}_4$; and
- Sim 5 for the simulations with parameter vector $\boldsymbol{\theta}_5$.

In order to signify the different models for the two responses, the following abbreviations are also used:

- **dk1** for the RMSPE estimated by the emulator from the **DiceKriging** software when applied to the $\mathbf{Y}_1$ data;
- **dk2** for the RMSPE estimated by the emulator from the **DiceKriging** software when applied to the $\mathbf{Y}_2$ data;
- **mult1** for the RMSPE estimated by the **Multivariate** Bayesian method when applied to the $\mathbf{Y}_1$ data;
- **mult2** for the RMSPE estimated by the **Multivariate** Bayesian method when applied to the $\mathbf{Y}_2$ data;
- **uni1** for the RMSPE estimated by the **Univariate** Bayesian method when applied to the $\mathbf{Y}_1$ data; and
- **uni2** for the RMSPE estimated by the **Univariate** Bayesian method when applied to the $\mathbf{Y}_2$ data.

Four values for the hyperspatial range parameters are chosen to quantify changes in the prediction quality. The RMSPE when the hyperspatial range parameters are set to their **true** values are used for a benchmark to compare to the predictions for mis-estimating the hyperspatial range parameters at **half** and **double** their true values. In order to assess the prediction performance of the three models, the RMSPE for the models when the hyperspatial range parameters are **estimated** are also provided. For lucidity, these estimations will be referred to as:

- **true** when the hyperspatial range parameters are not estimated by maximising the likelihood but are set to the true values from the simulation;
- **double** when the hyperspatial range parameters are not estimated by maximising the likelihood but are set to double the true values from the simulation;
- **half** when the hyperspatial range parameters are not estimated by maximising the likelihood but are set to half the true values from the simulation; and
- **estim** when the hyperspatial range parameters are estimated by maximising the likelihood.

The mean RMSPE results under the two variance-covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\text{high}}$ are presented in Tables 6.1 and 6.2, respectively. There is one major difference between the two tables. In Table 6.1, presenting the $\mathbf{\Sigma}$ case, the values are more than three times smaller than those in Table 6.2, presenting the $\mathbf{\Sigma}_{\text{high}}$ case. This is to be expected given the increased variance. Otherwise, both tables show very similar trends, which will be discussed in further detail below.

Whenever the three models have been set to the same value of the hyperspatial range parameters, the RMSPEs estimated by the three models are absolutely identical. The largest difference between using the true values of the hyperspatial range parameters and either halving or doubling that value is only 5% for any of the scenarios. This indicates that mis-specification of the hyperspatial range parameters to either double or half their true value has a very small effect on the quality of the predictions.

It is interesting to note that in both Tables 6.1 and 6.2, as the true values of the hyperspatial range parameters increase, the RMSPE decreases and hence, so do the differences between the RMSPE from the different models. This suggests that the larger the hyperspatial range parameters, the less impact they have on the predictions.

The only differences between the RMSPE values from the three different models are observed when the three models have to estimate the hyperspatial range parameters.

At most, the difference between the RMSPE estimated using the **true** hyperspatial range parameters is 14% smaller than the largest RMSPE from the **estimated** cases for each scenario. This indicates that the RMSPE is still relatively unaffected by the mis-estimation of the hyperspatial range parameters.

On the basis of the differences between the RMSPE values for the estimated hyperspatial range parameters, the predictive performance of the three models can be compared with each other. For all cases, the multivariate Bayesian model has the lowest RMSPE, and in some cases the RMSPE for the multivariate Bayesian model is lower than the RMSPE when setting the hyperspatial range parameters to **double** or **half** their values. All of this means that the multivariate model performs best in terms of prediction. In all cases, the second best model in terms of prediction is the univariate Bayesian model. For some cases, the two univariate models have almost identical prediction performance. The differences between the models are not significantly different as the 95% bootstrap confidence intervals for the estimated RMSPE are overlapping.

Simulation	$\boldsymbol{\theta}$	**DK** $\mathbf{Y}_1$	**Uni** $\mathbf{Y}_1$	**Mult** $\mathbf{Y}_1$	**DK** $\mathbf{Y}_2$	**Uni** $\mathbf{Y}_2$	**Mult** $\mathbf{Y}_2$
Sim 1	true	0.61	0.61	0.61	0.61	0.61	0.61
	double	0.63	0.63	0.63	0.62	0.62	0.62
	half	0.62	0.62	0.62	0.61	0.61	0.61
	estim	0.64 (0.55, 0.76)	0.63 (0.55, 0.73)	0.63 (0.54, 0.73)	0.63 (0.54, 0.72)	0.62 (0.52, 0.75)	0.61 (0.51, 0.72)
Sim 2	true	0.57	0.57	0.57	0.56	0.56	0.56
	double	0.59	0.59	0.59	0.57	0.57	0.57
	half	0.59	0.59	0.59	0.58	0.58	0.58
	estim	0.62 (0.55, 0.73)	0.59 (0.50, 0.70)	0.58 (0.49, 0.69)	0.63 (0.52, 0.70)	0.58 (0.49, 0.67)	0.57 (0.49, 0.66)
Sim 3	true	0.45	0.45	0.45	0.44	0.44	0.44
	double	0.46	0.46	0.46	0.45	0.45	0.45
	half	0.46	0.46	0.46	0.46	0.46	0.46
	estim	0.51 (0.40, 0.62)	0.46 (0.38, 0.55)	0.45 (0.38,0.55)	0.50 (0.40, 0.62)	0.46 (0.38, 0.54)	0.45 (0.38,0.53)
Sim 4	true	0.31	0.31	0.31	0.31	0.31	0.31
	double	0.31	0.31	0.31	0.31	0.31	0.31
	half	0.32	0.32	0.32	0.32	0.32	0.32
	estim	0.33 (0.27, 0.40)	0.32 (0.27, 0.38)	0.32 (0.27, 0.37)	0.33 (0.27, 0.40)	0.32 (0.26, 0.38)	0.32 (0.26, 0.37)
Sim 5	true	0.24	0.24	0.24	0.24	0.24	0.24
	double	0.25	0.25	0.25	0.25	0.25	0.25
	half	0.25	0.25	0.25	0.25	0.25	0.25
	estim	0.26 (0.21, 0.31)	0.25 (0.21, 0.31)	0.25 (0.21, 0.30)	0.25 (0.21, 0.31)	0.25 (0.21, 0.31)	0.25 (0.21, 0.29)

TABLE 6.1: Table containing the mean RMSPE for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ under $\boldsymbol{\Sigma}$ using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed Bayesian emulator for the **univariate** case and the **multivariate** emulator) for the four different possible values for the hyperspatial range parameters (true, double, half and estimated). For the estimated RMSPE, there is a 95% bootstrap confidence interval included.

Simulation	$\boldsymbol{\theta}$	DK $\mathbf{Y}_1$	Uni $\mathbf{Y}_1$	Mult $\mathbf{Y}_1$	DK $\mathbf{Y}_2$	Uni $\mathbf{Y}_2$	Mult $\mathbf{Y}_2$
Sim 1	true	1.94	1.94	1.94	1.91	1.91	1.91
	double	1.98	1.98	1.98	1.95	1.95	1.95
	half	1.98	1.98	1.98	1.94	1.94	1.94
	estim	2.01 (1.76, 2.33)	2.01 (1.75, 2.31)	1.98 (1.72, 2.32)	1.98 (1.68, 2.28)	1.98 (1.65, 2.34)	1.95 (1.61, 2.27)
Sim 2	true	1.79	1.79	1.79	1.78	1.78	1.78
	double	1.83	1.83	1.83	1.83	1.83	1.83
	half	1.85	1.85	1.85	1.84	1.84	1.84
	estim	1.95 (1.74, 2.33)	1.86 (1.59, 2.22)	1.84 (1.57, 2.21)	1.94 (1.62, 2.24)	1.85 (1.53, 2.17)	1.82 (1.53, 2.12)
Sim 3	true	1.42	1.42	1.42	1.39	1.39	1.39
	double	1.45	1.45	1.45	1.43	1.43	1.43
	half	1.46	1.46	1.46	1.44	1.44	1.44
	estim	1.62 (1.28, 2.00)	1.46 (1.23, 1.76)	1.44 (1.20, 1.74)	1.59 (1.25, 1.96)	1.43 (1.21, 1.73)	1.42 (1.20, 1.68)
Sim 4	true	0.98	0.98	0.98	0.97	0.97	0.97
	double	0.99	0.99	0.99	0.99	0.99	0.99
	half	1.00	1.00	1.00	1.00	1.00	1.00
	estim	1.04 (0.84, 1.30)	1.00 (0.84, 1.22)	0.99 (0.84, 1.16)	1.03 (0.81, 1.25)	1.00 (0.81, 1.16)	0.99 (0.82, 1.13)
Sim 5	true	0.76	0.76	0.76	0.77	0.77	0.77
	double	0.78	0.78	0.78	0.78	0.78	0.78
	half	0.78	0.78	0.78	0.78	0.78	0.78
	estim	0.80 (0.65, 1.02)	0.79 (0.64, 0.99)	0.79 (0.64, 0.98)	0.80 (0.66, 0.95)	0.79 (0.66, 0.95)	0.78 (0.67, 0.92)

TABLE 6.2: Table containing the mean RMSPE for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ under $\boldsymbol{\Sigma}_{\text{high}}$ using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed Bayesian emulator for the **univariate** case and the **multivariate** emulator) for the four different possible values for the hyperspatial range parameters (true, double, half and estimated). For the estimated RMSPE, there is a 95% bootstrap confidence interval included.

Next, in order to understand better the difference in the RMSPE estimates, two rows from Tables 6.1 and 6.2 will be examined more closely in Table 6.3. Only the RMSPEs for the estimated hyperspatial range parameters will be included in Table 6.3 as it is the only way to assess the prediction power of the three different models when the hyperspatial range parameters have to be estimated. In addition to Table 6.3, in order to provide information on the spread of the RMSPE across the one hundred simulation scenarios, the boxplots for the RMSPEs are examined. Figures 6.10 and 6.11 present the boxplots under $\boldsymbol{\Sigma}$ for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively, whereas Figures 6.12 and 6.13 present the boxplots under $\boldsymbol{\Sigma}_{\text{high}}$ for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively. Firstly, the clear difference between the $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\text{high}}$ cases is observed with the $\boldsymbol{\Sigma}_{\text{high}}$ RMSPEs being more than three times as large as the $\boldsymbol{\Sigma}$ RMSPEs. This affects the spreads of the boxplots in the two scenarios. The boxplots for $\boldsymbol{\Sigma}$ in Figures 6.10 and 6.11 go between 0.49 to 0.82 (spread of 0.33) and 0.49 to 0.78 (spread of 0.29), respectively, whereas the boxplots for the $\boldsymbol{\Sigma}_{\text{high}}$ case in Figures 6.12 and 6.13 go from 1.60 to 2.60 (spread of 1.00) and 1.54 to 2.48 (spread

of 0.94), respectively. It is clear that in the high variance case, the RMSPEs are higher and more varying.

Most importantly, from looking at the RMSPEs for Sim 1 in Table 6.3, it is easiest to order the three models by their performance. In all four cases, the lowest RMSPE is estimated for the multivariate Bayesian model, although this difference is not statistically significant as the 95% bootstrap intervals overlap each other. For the high variance case of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, the univariate frequentist model and the Bayesian univariate model have the same RMSPE values. This is further confirmed by the boxplots in Figure 6.13, where the univariate Bayesian model has the largest spread of RMSPE. In fact, for 36 out of the one hundred simulated data sets, the frequentist univariate model has lower RMSPE than the multivariate Bayesian model and the lowest RMSPE based on estimated hyperspatial range parameters is estimated by the univariate frequentist model. This indicates that in some scenarios, it is possible that the univariate frequentist model will perform better than the multivariate Bayesian one even though the data are simulated as multivariate.

Variance	**DK $\mathbf{Y}_1$**	**Uni $\mathbf{Y}_1$**	**Mult $\mathbf{Y}_1$**	**DK $\mathbf{Y}_2$**	**Uni $\mathbf{Y}_2$**	**Mult $\mathbf{Y}_2$**
$\boldsymbol{\Sigma}$	0.64 (0.55, 0.76)	0.63 (0.55, 0.73)	0.63 (0.54, 0.73)	0.63 (0.54, 0.72)	0.62 (0.52, 0.75)	0.61 (0.51, 0.72)
$\boldsymbol{\Sigma}_{\text{high}}$	2.01 (1.76, 2.33)	2.01 (1.75, 2.31)	1.98 (1.72, 2.32)	1.98 (1.68, 2.28)	1.98 (1.65, 2.34)	1.95 (1.61, 2.27)

TABLE 6.3: Table containing the mean RMSPE for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ under **Sim 1** for the **estimated** hyperspatial range parameters using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed Bayesian emulator for the **univariate** case and the **multivariate** emulator).

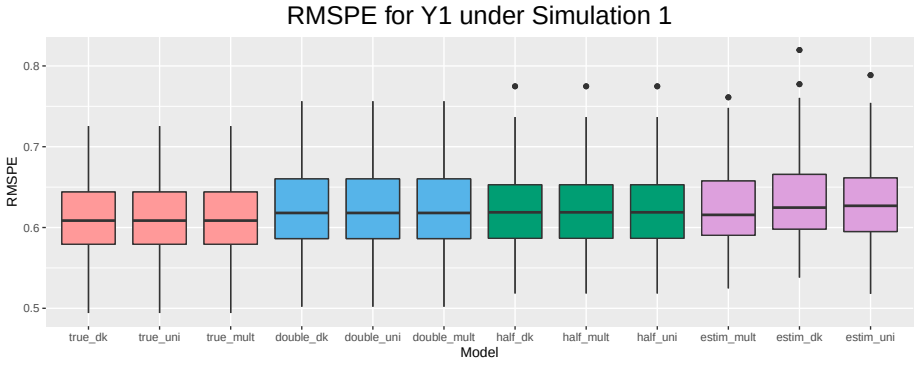

FIGURE 6.10: Boxplots of the RMSPE for $\mathbf{Y}_1$ under $\boldsymbol{\Sigma}$ using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed Bayesian emulator for the **univariate** case and the **multivariate** emulator) for the four different possible values for the hyperspatial range parameters (true, double, half and estimated) for Simulation 1.

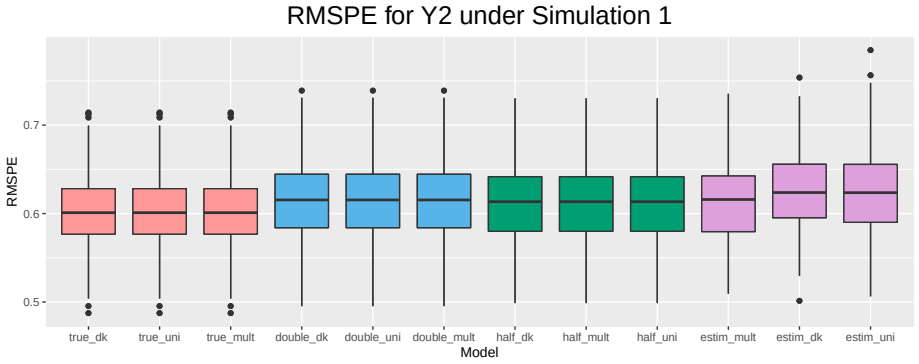

FIGURE 6.11: Boxplots of the RMSPE for $\mathbf{Y}_2$ under $\boldsymbol{\Sigma}$ using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed emulator for the **univariate** case and the **multivariate** emulator) for the four different possible values for the hyperspatial range parameters (true, double, half and estimated) for Simulation 1.

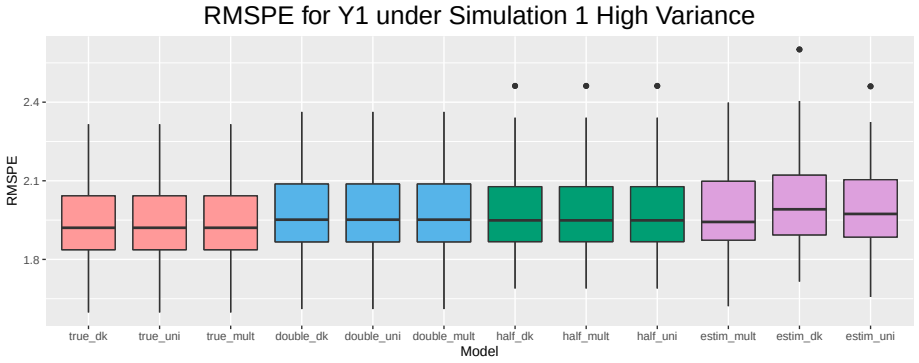

FIGURE 6.12: Boxplots of the RMSPE for $\mathbf{Y}_1$ under $\boldsymbol{\Sigma}_{\text{high}}$ using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed emulator for the **univariate** case and the **multivariate** emulator) for the four different possible values for the hyperspatial range parameters (true, double, half and estimated) for Simulation 1.

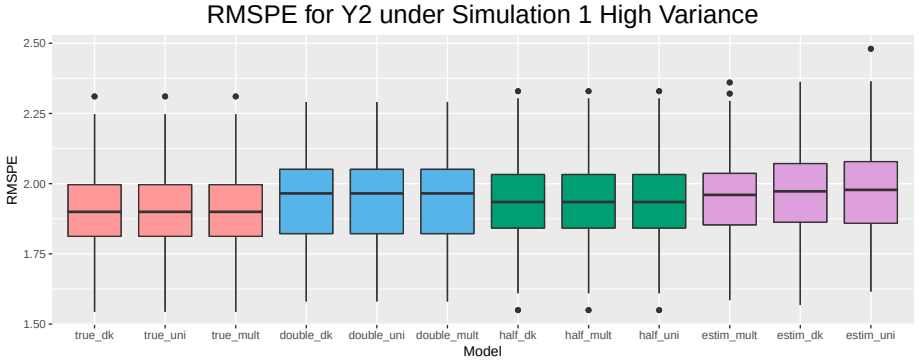

FIGURE 6.13: Boxplots of the RMSPE for $\mathbf{Y}_2$ under $\boldsymbol{\Sigma}_{\text{high}}$ using three different methodologies (the univariate frequentist model from **DiceKriging**, the proposed emulator for the **univariate** case and the **multivariate** emulator) for the four different possible values for the hyperspatial range parameters (true, double, half and estimated) for Simulation 1.

Findings

Simulation Study 2 assessed the loss of predictive power due to mis-estimation of the hyperspatial range parameters by the three different models (univariate frequentist,

univariate Bayesian and multivariate Bayesian). The RMSPEs from both variance-covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\text{high}}$ in the two tables (Tables 6.1 and 6.2 respectively) differ as the $\boldsymbol{\Sigma}_{\text{high}}$ estimates are three times as large as the $\boldsymbol{\Sigma}$ ones. However, it was found that the RMSPE values within each table are very similar to each other regardless of the incorrectly specified values of the hyperspatial range parameters for both variances. For both variance cases, the multivariate Bayesian model outperforms both the univariate Bayesian and the univariate frequentist models. However, it has to be noted that as the true values of the hyperspatial range parameters increased, the RMSPE values decrease and get closer to each other and the effect of mis-estimating the hyperspatial range parameters is reduced. Hence, the difference between the three models is also reduced.

6.2.3 Conclusions

In this section, two simulation studies were performed in order to (i) compare the ability of the three tested models (the univariate frequentist model from the **DiceKriging** software, and the univariate and multivariate versions of the proposed Bayesian model) to correctly identify the hyperspatial range parameters in a setting which mimics the real-life data and (ii) asses the loss of predictive power to mis-estimating the hyperspatial range parameters. The data for both studies were simulated identically. The results for both studies exhibit similar trends for both variance cases.

Simulation Study 1 addressed question (i) and found that all models struggle with correctly identifying the hyperspatial range parameters $\boldsymbol{\theta}$. As the hyperspatial range parameter values were increased, the models tend to underestimate the hyperspatial range parameters. It appeared that the multivariate Bayesian model struggles with estimating the hyperspatial range parameters but it was the only model that did not calculate extreme values for the hyperspatial range parameters.

Simulation Study 2 addressed question (ii) and found that the largest difference by mis-estimating the hyperspatial range parameters from estimating them correctly is 14%. Hence, mis-estimating the hyperspatial range parameters results in relatively unaffected predictions as Cressie suggests in [50]. It was found that the three emulators produce very similar RMSPE with the multivariate Bayesian model slightly outperforming the two univariate models, although the difference is not statistically significant.

The two simulation studies showed that hyperspatial range parameters are hard to estimate correctly, but even with "sensible" [50] estimates for the hyperspatial range parameters the RMSPE is relatively unaffected. For a multivariate setting when the hyperspatial range parameters have similar values, using one set of hyperspatial range

parameter estimates by the multivariate Bayesian model results in lower RMSPE and hence, better prediction performance. In order to test whether using one set of hyper-spatial range parameters for all the monitoring stations in Aberdeen and Glasgow would result in an improved prediction in comparison to the univariate models from Chapter 5, the multivariate Bayesian model is applied to the ADMS-Urban simulations for the NO_2 annual averages in Aberdeen and Glasgow.

6.3 Aberdeen case study

In this section, the multivariate Bayesian model proposed in Section 6.1 will be applied to the ADMS-Urban simulations for the six monitoring stations in Aberdeen and its performance will be compared to the univariate frequentist models in Section 5.2 in order to establish whether fitting a multivariate model results in smaller RMSPE and faster modelling. The univariate version of the proposed Bayesian model will not be fitted as it always performed worse than the multivariate Bayesian model in the simulation studies in Section 6.2. As previously stated in Section 6.1, there is no prior information about any of the parameters and therefore, uninformative priors are used for the multivariate Bayesian model. Furthermore, since in Chapter 5, the univariate frequentist models for all the monitoring stations in Aberdeen have just the three LHC inputs (emissions, wind speed and wind direction) as covariates, the model in this section will also only use the three inputs as covariates. As in Chapter 5, in order to avoid repetition, only the results for the Anderson Drive monitoring station will be presented in full, whereas the results for the other stations will be summarised. Once a final model is chosen based on a 10-fold CV RMSPE, the emulator will be applied to explore the different meteorological conditions under which the annual average regulation for NO_2 will be broken. Lastly, it has to be noted that the work in this section is an alternative model fitting to the one presented in [82] but uses the same data.

6.3.1 Modelling

Anderson Drive

The Anderson Drive fixed effects parameters estimated by the multivariate model are summarised in Table 6.4. The fixed parameters estimated by the multivariate model are almost identical to the estimates from the univariate frequentist model from the **DiceKriging** software presented in the fourth column of the table. The standard errors from the multivariate model (third column) are very similar than those from the univariate model (fifth column) suggesting that there is no difference between the two models. This

indicates that there is a loss of information in the multivariate model in comparison to the univariate one. The variance and hyperspatial range parameter estimates from the two models are compared in Table 6.5. Both models estimate the variance σ^2 to be 0.07, which is quite small. However, the hyperspatial range parameter estimates for the two models are different to each other. The emissions hyperspatial range parameters are quite similar and in both cases the largest. But the multivariate Bayesian model estimates the wind speed to be the second largest, whereas the univariate frequentist model suggests that the wind direction hyperspatial range parameter should be the second largest. Nonetheless, the hyperspatial range parameter estimates for both models indicate high correlation between the inputs in both cases as the large hyperspatial range parameter estimates from both models suggest a residual correlation at high distances.

Coefficient	**Estim. Mult**	**St. Error Mult.**	**Estim. DK**	**St. Error DK**
Intercept	31.33	0.16	31.31	0.17
Emissions	0.08	0.003	0.08	0.003
Wind Speed	-0.06	0.005	-0.06	0.006
Wind Direction	0.02	0.008	0.02	0.006

TABLE 6.4: Summary of fixed effect parameters from the multivariate Bayesian and univariate frequentist GP models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction ($^\circ$ change) as covariates and an exponential kernel.

Model	$\widehat{\theta}_{\textbf{Emissions}}$	$\widehat{\theta}_{\textbf{Wind Speed}}$	$\widehat{\theta}_{\textbf{Wind Direction}}$	$\widehat{\sigma}^2$
Multivariate Bayesian	90.43	61.60	24.61	0.07
Univariate frequentist	98.16	46.16	59.55	0.07

TABLE 6.5: Summary of the hyperspatial range parameters and variance from the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction ($^\circ$ change) as covariates and an exponential kernel.

Then, the diagnostic plots for the multivariate fit at Anderson Drive are examined in Figure 6.14. Overall, the residual plots do not indicate any problems with the fit and are very similar to the diagnostic plots for the univariate frequentist model in Figure 5.3. The residuals vs. fitted values plot in a) shows random scatter, which is further confirmed by the residuals vs. each of the covariates in plots e), f) and g). Although the qq-plot in plot c) appears to have heavy tails, in fact the highest outlier is at 0.40 and the range of the residuals is too small to indicate an issue. Furthermore, the histogram of the residuals in plot d) has a well defined bell-shaped curve indicating no issues. A Shapiro-Wilk test was performed on the residuals and the p-value was found to be 0.05 suggesting that there is no evidence for non-normality of the residuals. Most importantly, the actual vs. fitted values in plot b) shows the points are lying on the equivalence line suggesting an almost perfect fit.

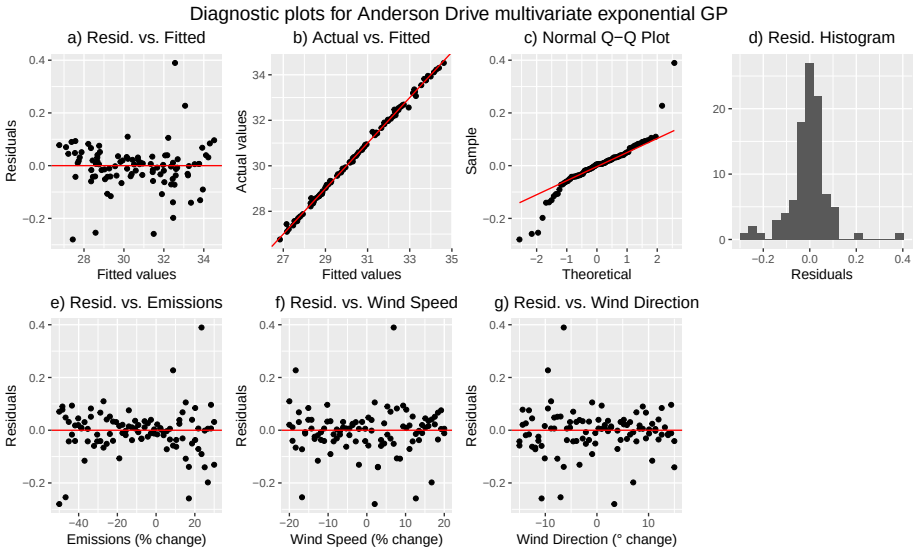

FIGURE 6.14: Diagnostic plots for the multivariate Bayesian GP model with an exponential kernel for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

Lastly, the multivariate Bayesian RMSPE is compared to the one from the univariate frequentist model in Table 6.6. It appears that the two models have identical prediction performance. It is interesting to note that although the two models have different hyperspatial temporal parameters, the RMSPEs are identical, when rounding is applied. Furthermore, the 95% bootstrap confidence intervals indicate that there is no statistically significant difference between the predictive performance of the two models.

Model	RMSPE
Multivariate	0.09 (0.06, 0.11)
Univariate	0.09 (0.06, 0.11)

TABLE 6.6: Comparing the predictive performance of the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Anderson Drive with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

Other stations

Similarly to Anderson Drive, fixed effect estimates by the multivariate Bayesian and the univariate frequentist models are very similar to each other as seen in Table 6.7. For all stations the estimates for the fixed effect parameters are almost identical. The standard errors estimated by the model appear almost identical.

The main differences between the multivariate Bayesian and the univariate frequentist models come from the distinctly estimated hyperspatial range parameters. In order to

Station	Coefficient	Estim. Mult.	St. Error Mult.	Estim. DK	St. Error DK
Anderson Drive	Intercept	31.33	0.16	31.31	0.17
	Emissions	0.08	0.003	0.08	0.003
	Wind Speed	-0.06	0.005	-0.06	0.006
	Wind Direction	0.02	0.008	0.02	0.006
Errol Place	Intercept	28.09	0.12	28.07	0.11
	Emissions	0.05	0.002	0.05	0.002
	Wind Speed	-0.05	0.004	-0.05	0.004
	Wind Direction	0.01	0.006	0.01	0.004
King Street	Intercept	36.04	0.25	36.01	0.23
	Emissions	0.12	0.004	0.12	0.004
	Wind Speed	-0.11	0.008	-0.11	0.009
	Wind Direction	0.04	0.01	0.05	0.007
Market Street 2	Intercept	47.37	0.39	47.33	0.40
	Emissions	0.22	0.007	0.22	0.007
	Wind Speed	-0.17	0.01	-0.17	0.01
	Wind Direction	0.06	0.02	0.06	0.01
Union Street	Intercept	49.52	0.45	49.45	0.20
	Emissions	0.24	0.008	0.24	0.005
	Wind Speed	-0.19	0.01	-0.20	0.01
	Wind Direction	0.02	0.02	0.02	0.002
Wellington Road	Intercept	43.88	0.38	43.84	0.40
	Emissions	0.19	0.007	0.19	0.007
	Wind Speed	-0.16	0.01	-0.17	0.01
	Wind Direction	0.07	0.02	0.07	0.01

TABLE 6.7: Summary of fixed effect parameters from the multivariate Bayesian GP and the univariate frequentist emulator from the **DiceKriging** package for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for all Aberdeen monitoring stations with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

compare the hyperspatial range parameter estimates from the different models, they are summarised in Table 6.8. In terms of the estimates for the hyperspatial range parameter for emissions, the multivariate model and the univariate ones have very similar estimates except for Union Street. The multivariate hyperspatial range parameter estimate for wind speed is higher than the one from all the univariate models. However, the most clear difference comes from the wind direction estimates, where the multivariate Bayesian model estimates a value much smaller than any of the univariate frequentist models.

The differences from the hyperspatial range parameter estimates result in differences in the variance σ^2 and the RMSPE estimates for the multivariate and univariate models. These differences are summarised in Table 6.9. It is clear that the univariate models perform slightly better than the multivariate one in terms of RMSPE for stations with higher NO_2 annual averages, although the difference is not statistically significant as the 95% bootstrap intervals perfectly overlap each other. Furthermore, the variances

Model	$\widehat{\theta}_{\text{EM}}$	$\widehat{\theta}_{\text{WS}}$	$\widehat{\theta}_{\text{WD}}$
Multivariate Bayesian	90.43	61.60	24.61
Anderson Drive Univariate Frequentist	98.16	46.16	59.55
Errol Place Univariate Frequentist	92.64	36.39	73.05
King Street Univariate Frequentist	82.25	29.79	83.43
Market Street 2 Univariate Frenquentist	89.65	38.48	66.12
Union Street Univariate Frenquentist	49.40	7.99	2934.23
Wellington Road Univariate Frenquentist	89.90	39.62	67.35

TABLE 6.8: Summary of the hyperspatial range parameter estimates from the multivariate Bayesian and univariate frequentist models for NO_2 annual average (μg m^{-3}) from ADMS-Urban for the Aberdeen monitoring stations with emissions (% change), wind speed (% change) and wind direction ($^\circ$ change) as covariates and an exponential kernel.

are almost identical. Therefore, it appears that the Aberdeen monitoring stations have very different hyperspatial range parameter estimates and using one set of hyperspatial range parameters for all stations results in a slight loss of information.

Station	Model	$\widehat{\sigma}^2$	RMSPE
Anderson Drive	Multivariate Bayesian	0.07	0.09 (0.06, 0.11)
	Univariate Frequentist	0.07	0.09 (0.06, 0.11)
Errol Place	Multivariate Bayesian	0.04	0.06 (0.04, 0.08)
	Univariate Frequentist	0.03	0.06 (0.04, 0.08)
King Street	Multivariate Bayesian	0.16	0.14 (0.10, 0.18)
	Univariate Frequentist	0.14	0.14 (0.10, 0.18)
Market Street 2	Multivariate Bayesian	0.38	0.23 (0.16, 0.29)
	Univariate Frequentist	0.40	0.23 (0.16, 0.30)
Union Street	Multivariate Bayesian	0.52	0.28 (0.19, 0.38)
	Univariate Frequentist	0.23	0.28 (0.19, 0.37)
Wellington Road	Multivariate Bayesian	0.37	0.22 (0.15, 0.28)
	Univariate Frequentist	0.39	0.21 (0.15, 0.28)

TABLE 6.9: Summary of the variance and RMSPE from the multivariate Bayesian and univariate frequentist models for NO_2 annual average (μg m^{-3}) from ADMS-Urban for the Aberdeen monitoring stations with emissions (% change), wind speed (% change) and wind direction ($^\circ$ change) as covariates and an exponential kernel.

Findings

From Table 6.7 is is clear that both models explain the variation due to the fixed effect terms very similarly. However, applying the same hyperspatial range parameters to all stations results in a negligible loss of information. Both models perform almost perfectly in predicting the NO_2 annual average. This is in accordance with a previous analysis of the ADMS-Urban simulations in [82], where a suboptimal multivariate GP model was applied and it was found that the suboptimal multivariate model is very similar in performance to the univariate linear models for each of the monitoring stations. As the multivariate model requires only fitting one model for all stations, it is chosen as the preferred model.

6.3.2 Emulation

In this subsection, the results from emulating the ADMS-Urban simulations for the multivariate frequentist models for the Aberdeen monitoring stations are examined. The emulated NO_2 annual average concentrations are obtained over a discretised grid of the input space from the LHC. The grid to be evaluated is based on increments of 0.5 along the dimensions for each of the three inputs (emissions, wind speed and wind direction). In addition to calculating the NO_2 annual average and the respective uncertainty around that estimate, the probability of breaking the 40 μg m^{-3} regulation will be provided. The probability of exceedance for each station for the annual average regulation is defined for an observation y^0 with a set of inputs (x^0_{EM}, x^0_{WS}, x^0_{WD}). y^0 follows a normal distribution:

$$y^0 \sim \text{N}(\mu^0, \sigma^{2,0}), \tag{6.19}$$

where μ^0 is the predicted value for the NO_2 annual average and $\sigma^{2,0}$ is the variance for the predicted value, from where $P(y^0 > 40)$ is estimated using the Normal probability density function (pdf). Contour plots for the emulated NO_2 annual average under three variations for wind direction will be provided (similar to the contour plots in [82]) alongside the respective variance for the annual average as well as contour plots for the probability of exceeding the regulation limit of 40 μg m^{-3}. The three variations for wind direction to be examined are:

- -15° variation from the baseline value for 2012, resulting in a more eastern prevailing wind;
- 0° variation from the baseline value for 2012; and

- 15° variation from the baseline value for 2012, resulting in a more western prevailing wind.

On the left of Figure 6.15 presents contour plots for the emulated NO_2 annual averages for each of the stations when the wind direction is set to have a -15° change from the 2012 observed baseline value, whereas on the right are the corresponding standard deviations associated with the observations. From the fact that the standard deviations for each observation are at most 0.40, it has to be concluded that the emulated values have a very small uncertainty bound around them. It is clear that for the stations Anderson Drive and Errol Place there are no combinations of emissions and wind speed for which the annual limit of 40 μg m^{-3} will be exceeded. For King Street, it appears that if the emissions increase and wind speed decreases, the regulation could be breached. For Market Street 2, Union Street and Wellington Road, it is obvious that only a combination of reduction in emissions and increase in wind speed will result in an annual average within the regulations. On the standard deviation plots, it can be seen that the stations with higher NO_2 averages have higher standard deviations associated with them which is expected. Furthermore, there is a trend of increased standard deviations going from left to right on the plots which is associated with the higher NO_2 concentrations as emissions are increased. Also, there are higher standard deviations at the corners of each of the standard deviation plots because these are the edges of the LHC design.

These conclusions are further confirmed by the plots for the probabilities of exceedance in Figure 6.16. The plots for all stations show a sharp transition from not breaching the regulation to breaching the regulation. This is due to the deterministic nature of ADMS-Urban combined with the specific cutoff point from the regulation at 40 μg m^{-3}, which is also confirmed by the results in [82]. For the plots, where there is a high chance of exceedance, the change from probability of zero to probability of one is rapid due to the accuracy of the predictions from the emulator as seen by the low standard deviations for the emulated NO_2 annual average in Figure 6.15. The plots for Anderson Drive and Errol Place show that the probability of exceeding the 40 μg m^{-3} limit is zero for all combinations of emissions and wind speed. For King Street, there is a small region, where exceedances can occur if emissions increase by at least 20%, whereas wind speed decreases by at least 10%. The stations at Market Street 2 and Union Street are the ones that have the highest NO_2 annual averages and the probability plots show that even with reductions in emissions larger than 40% if the wind speed is low, it is very likely that the annual average will not comply with the regulation. The Wellington Street monitoring station probability plot shows that with an emissions reduction by 30%, regardless of wind speed variation, the station will comply with the regulations.

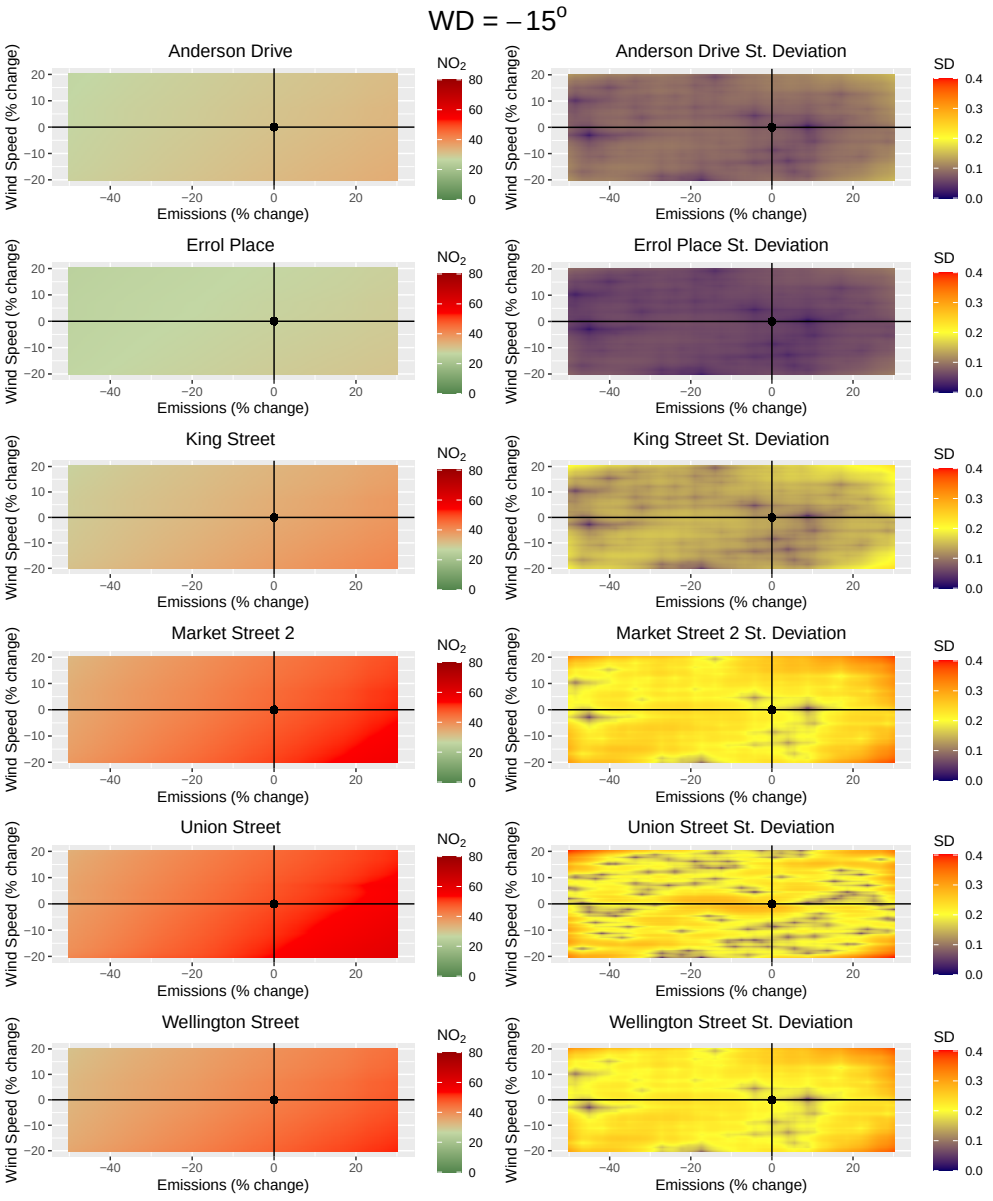

FIGURE 6.15: Contour plots for emulated NO_2 annual average (μg m^{-3}) when wind direction is set to -15° change from the 2012 baseline for the ADMS-Urban simulations for the six monitoring stations in Aberdeen based on the multivariate model with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel are presented on the left side. Contour plots for the standard deviation of the emulated NO_2 annual averages are provided on the right. In each plot, the black circle depicts the baseline realisation.

Next, the contour plots for the emulated NO_2 annual average when wind direction is set to a 0° change from the 2012 observed baseline and its respective standard deviations are presented in Figure 6.17. The plots are very similar to those when wind direction is set to -15° change from the baseline in Figure 6.15. For the monitoring stations at Anderson Drive and Errol Place there does not appear to be any annual averages going above 40 μg m^{-3}. However, for the King Street Station it appears that there is more area covered in red. This means that under the baseline wind direction, King Street

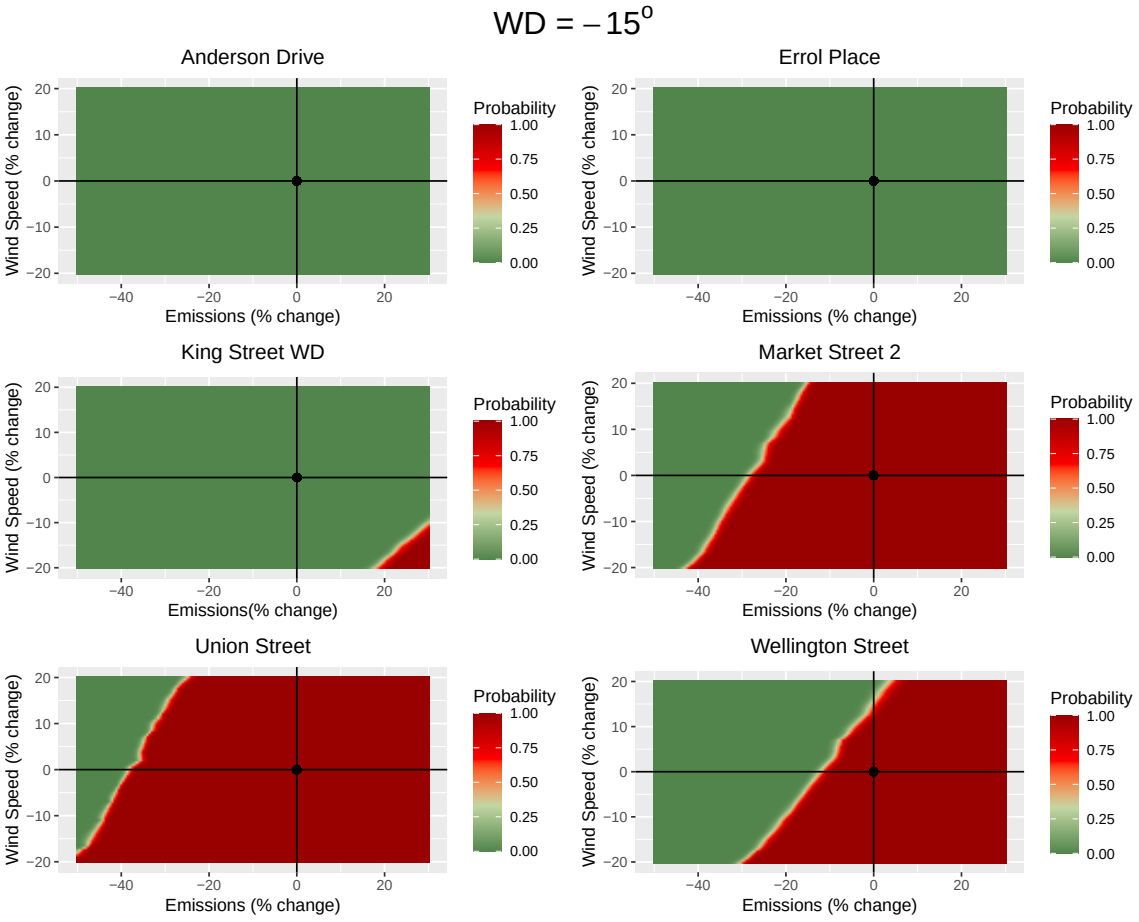

FIGURE 6.16: Contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average (μg m^{-3}) when wind direction is set to -15° change from the 2012 baseline for the ADMS-Urban simulations for the six monitoring stations in Aberdeen based on the multivariate model with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel. In each plot, the black circle depicts the baseline realisation.

has to be considered as an "at-risk" station. Furthermore, Market Street 2, Union Street and Wellington Road are expected to have NO_2 annual average which breaks the regulation. The standard deviations for emulated values for all stations are quite low, at most 0.40. Similar trends to those observed in the standard deviation plots in Figure 6.16 are observed here - stations with higher NO_2 averages have higher standard deviations, standard deviations increase from left to right on the plots and there are higher standard deviations at the corners of each of the standard deviation plots.

The contour plots for the probability of exceedance of 40 μg m^{-3} regulation of the emulated NO_2 annual average when wind direction is set to 0° change from the 2012 observed baseline are presented in Figure 6.18. The change from probability of zero to probability of one is very rapid again because the accuracy of the emulator as demonstrated by the low standard deviation of the emulated values. The probability plots for Anderson Drive, Errol Place and King Street monitoring stations are very similar

to the contour plots for wind direction being set to -15° and that it is very unlikely that the annual average regulation will be broken. However, for the Market Street 2 and Union Street stations, there is high probability that the annual average regulation will be broken regardless of the reduction in emissions, unless wind speed is higher than the baseline value. For the Wellington Road station, it is only certain that the annual average regulation will not be broken only if the emissions are reduced by 40%.

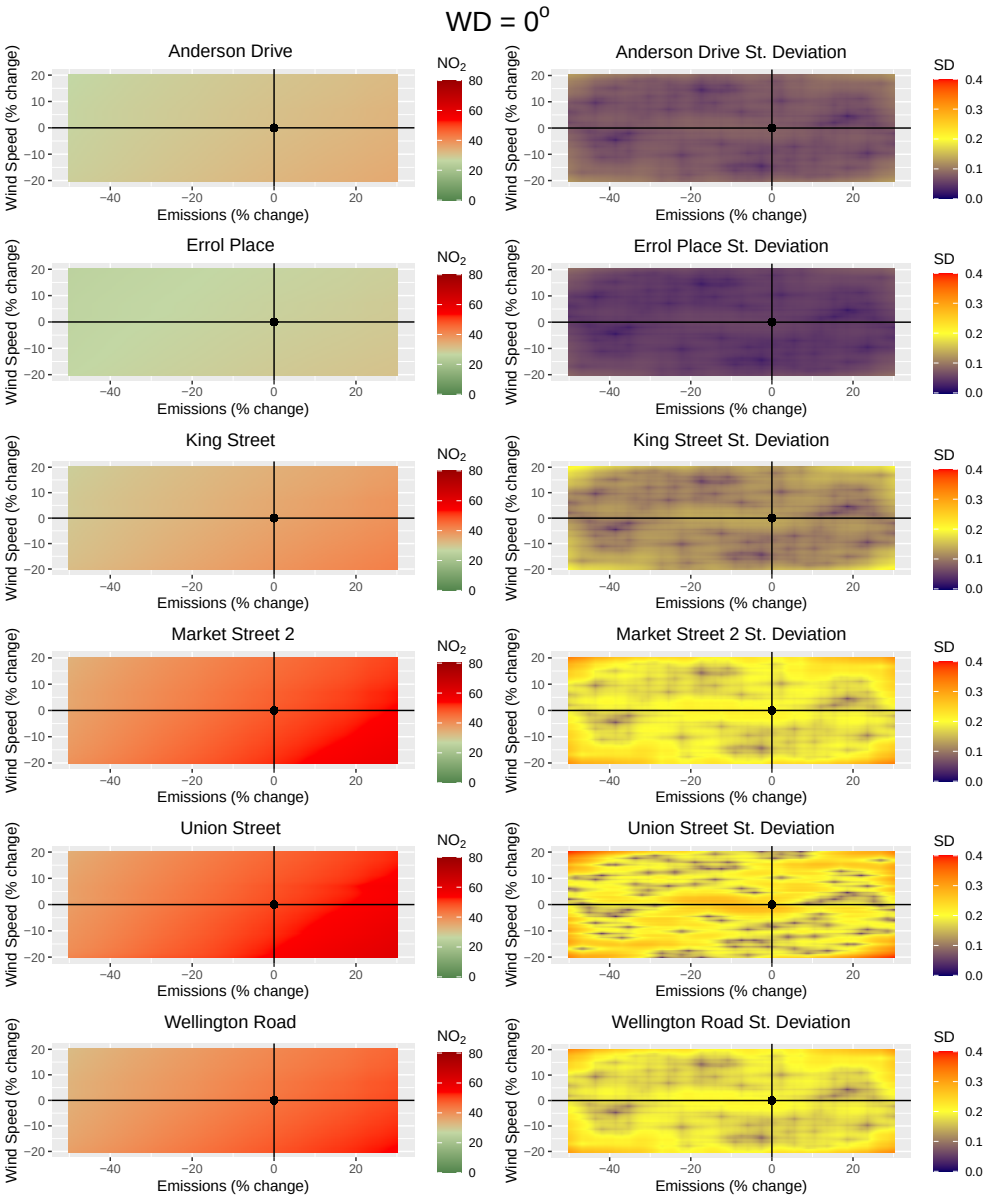

FIGURE 6.17: Contour plots for emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 0° change from the 2012 baseline for the ADMS-Urban simulations for the six monitoring stations in Aberdeen based on the multivariate model with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel are presented on the left side. Contour plots for the standard deviations of the emulated NO_2 annual averages are provided on the right. In each plot, the black circle depicts the baseline realisation.

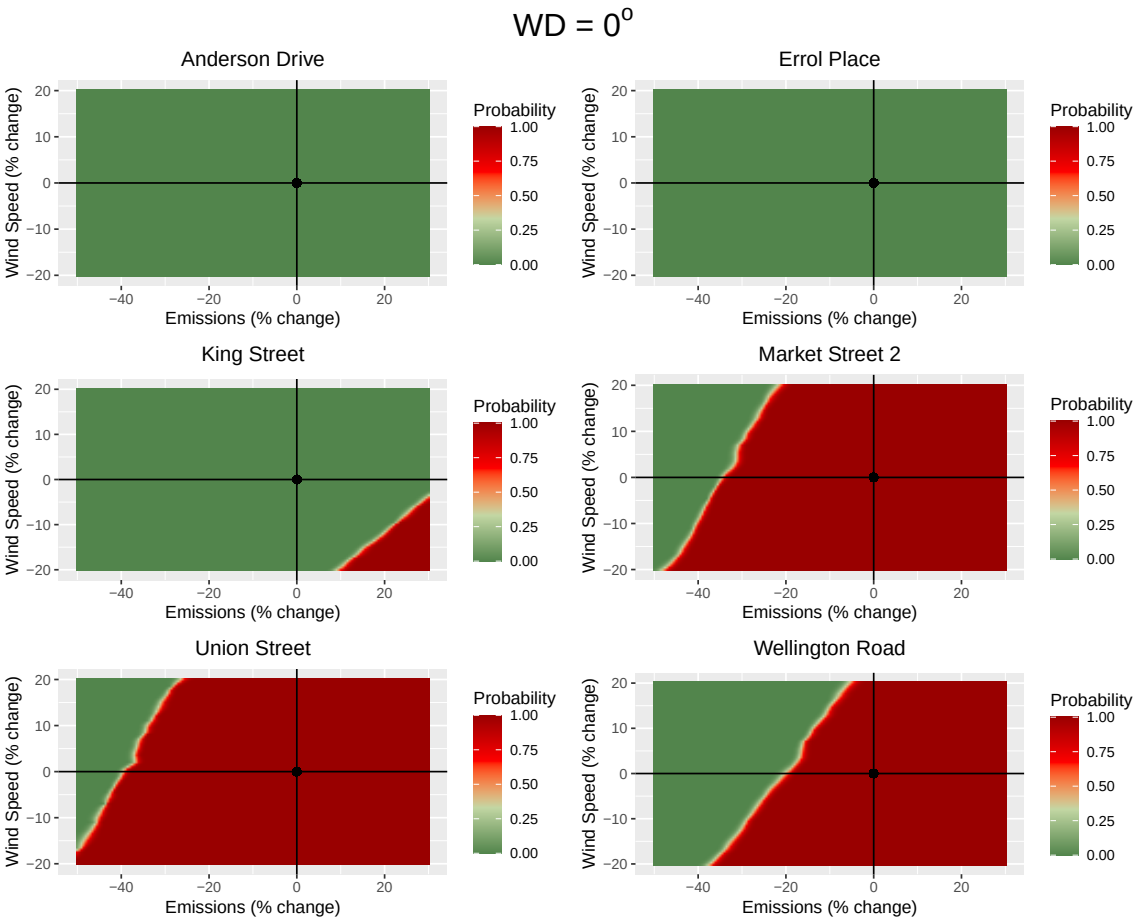

FIGURE 6.18: Contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 0° change from the 2012 baseline for the ADMS-Urban simulations for the six monitoring stations in Aberdeen based on the multivariate model with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel. In each plot, the black circle depicts the baseline realisation.

In Figure 6.19 the contour plots for the emulated NO_2 annual average are examined when the wind direction is increased by 15° from the 2012 observed baseline value. The trend from Figures 6.15 and 6.17 that as the prevailing wind direction becomes more western, the NO_2 annual average increases is confirmed by this plot. The plot for the Anderson Drive monitoring station has a little hint of reddish hue at the bottom right corner of the plot suggesting an increase in the annual average, although the increase will not result in breaking the regulation. For the Errol Place monitoring station there does not appear to be any evidence of NO_2 annual average values which will break the regulation. For King Street the values are getting higher which is worrisome for a station which has been assigned as "at-risk". For the monitoring stations at Market Street 2, Union Street and Wellington Road it appears that most annual averages will be above the 40 μg m^{-3} regulation. The standard deviations contour plots on the right of Figure 6.19 are very similar to those from both Figures 6.15 and 6.17 and the same trends as

before are observed. Once again, the standard deviations are at most 0.40.

The contour plots for the probability of exceedance of 40 μg m^{-3} regulation of the emulated NO_2 annual average when wind direction is set to 15° change from the 2012 observed baseline are presented in Figure 6.20. The change from probability of zero to probability of one is very rapid again because the accuracy of the emulator. The plots confirm what was already seen in Figure 6.19. The Anderson Drive and Errol Place monitoring stations will comply with the regulation regardless of the emissions and wind speed levels. The King Street monitoring station can break the limit in the event of increased emissions and lower wind speed in comparison to the baseline. However, for the monitoring stations at Market Street 2 and Union Street, even if the emissions are reduced by 40%, it is still very likely to break the regulation if the wind speed is lower than the baseline. The Wellington Road monitoring station will only comply with the regulation for any wind speed if the emissions are reduced by 40%.

Findings

The three scenarios for wind direction being set to its low extreme variation from the baseline at -15°, the baseline 0° and high extreme variation from the baseline at 15°, it is seen that wind direction has a small but significant effect on the NO_2 annual average for the six different stations in Aberdeen. It was seen that as the prevailing wind direction is going from eastern to western, the NO_2 annual average concentration increases. This is in accordance with the fixed effect modelling for wind direction, where the parameter estimates are very close to zero but significant. The contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average showed that for two of the monitoring stations (Market Street 2 and Union Street) regardless of the percentage of the reduced emissions, it is still very likely that the regulation will be broken if there is lower than the baseline wind speed. The Wellington Road station can comply with the regulations for all meteorological conditions if the emissions at the monitoring station are lowered by at least 40%. The King Street monitoring station will comply with the regulation as long as the emissions are not increased, whereas the Anderson Drive and Errol Place monitoring stations will never exceed the 40 μg m^{-3} for any set of meteorological conditions.

6.3.3 Conclusion

The multivariate Bayesian model was applied to the ADMS-Urban NO_2 annual average simulations and its performance was compared to the univariate frequentist model from the **DiceKriging** software. It was found that both models estimate very similar

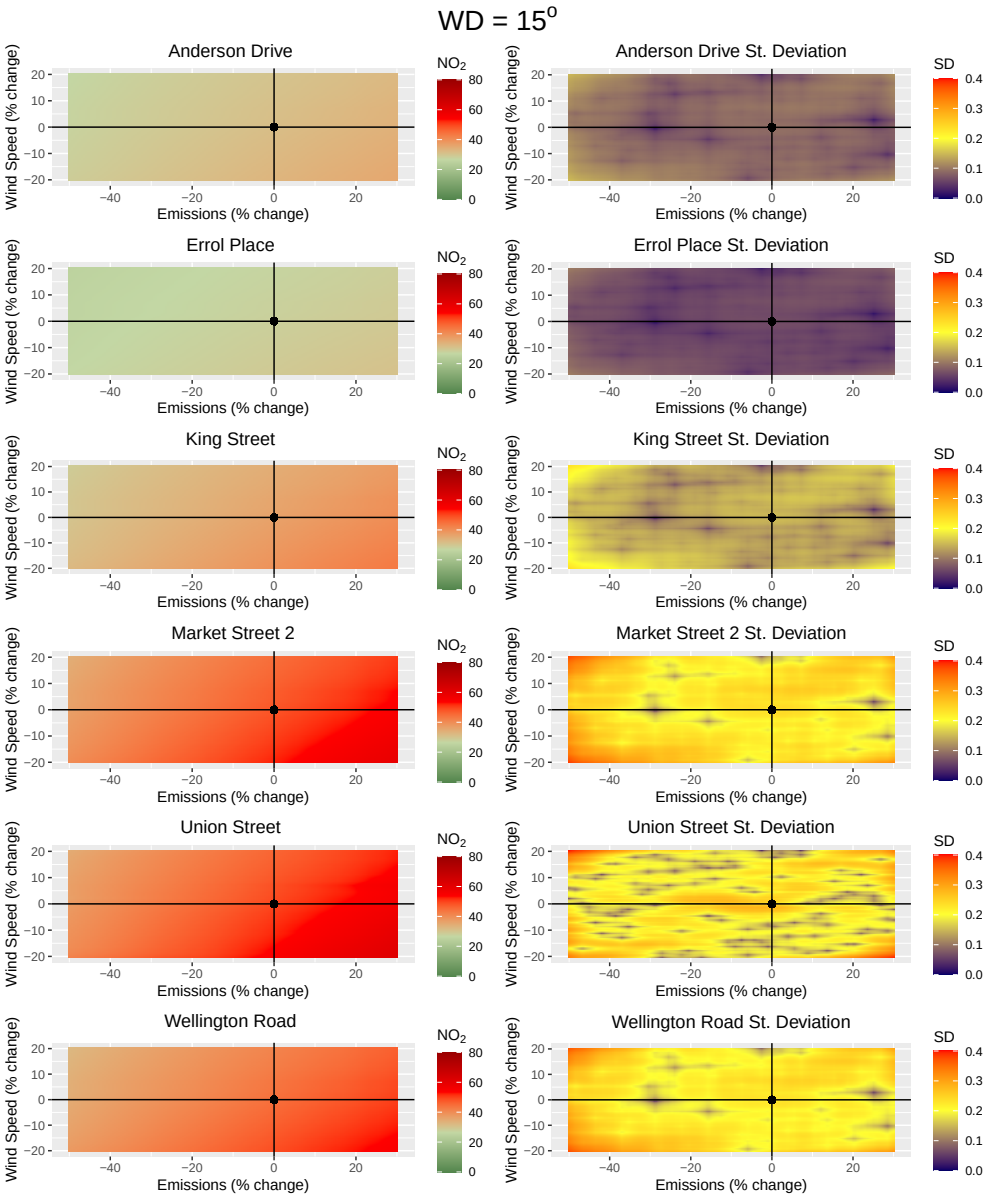

FIGURE 6.19: Contour plots for emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 15° change from the 2012 baseline for the ADMS-Urban simulations for the six monitoring stations in Aberdeen based on the multivariate model with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel are presented on the left side. Contour plots for the standard deviations of the emulated NO_2 annual averages are provided on the right. In each plot, the black circle depicts the baseline realisation.

fixed range effects but the univariate frequentist model estimates have smaller standard error. However, the smoothed set of hyperspatial range parameters estimated by the multivariate Bayesian model is very different from the hyperspatial range parameter values estimated at each station by the univariate frequentist models. Regardless, the multivariate Bayesian and the univariate frequentist models have almost identical performance in terms of prediction for untested values. The multivariate Bayesian model is chosen because it requires running only one model instead of six for each monitoring

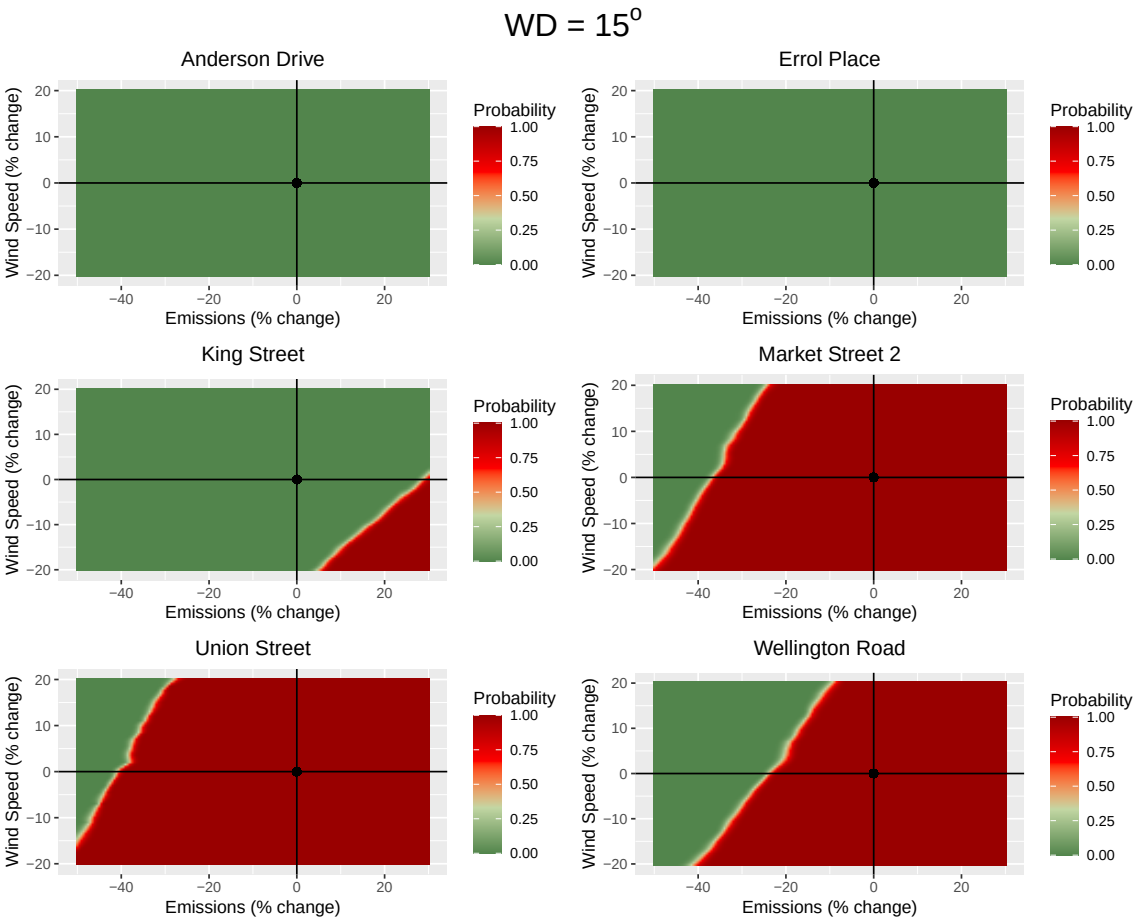

FIGURE 6.20: Contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 15° change from the 2012 baseline for the ADMS-Urban simulations for the six monitoring stations in Aberdeen based on the multivariate model with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel. In each plot, the black circle depicts the baseline realisation.

station.

Hence, the multivariate Bayesian models were used to create emulators for each of the monitoring stations. It was found that as the prevailing wind direction changes to more western, this results in higher NO_2 annual pollutions. This makes sense as Aberdeen is on the North Sea. Overall, it was concluded that for none of the examined scenarios, the NO_2 annual average at Anderson Drive and Errol Place will go above the 40 μg m^{-3}, whereas for King Street some combinations of higher from the baseline emissions and lower from the baseline wind speed will result in breaking the limit. For Market Street 2, Union Street and Wellington Road, emissions reductions (of at least 40%) are required in order to ensure that for some wind speeds (even below the baseline), the regulation limit will not be broken.

6.4 Glasgow case study

In this section, the multivariate Bayesian model proposed in Section 6.1 will be fitted to the ADMS-Urban simulations for the eight monitoring stations in Glasgow and its performance will be compared to the univariate models for Glasgow in Section 5.3. The univariate version of the proposed Bayesian model will not be fitted as it always performed worse than the multivariate Bayesian model in the simulation studies in Section 6.2. Once again, there is no prior information about any of the parameters and therefore, uninformative priors are set as stated in Section 6.1. Based on the linear models for each of the monitoring stations in Glasgow, the univariate modelling for Glasgow tested and compared three different sets of inputs - a model with 3 covariates (emissions, wind speed and wind direction), a model with 4 covariates (emissions, emissions squared, wind speed and wind direction) and a model with 5 covariates (emissions, emissions squared, wind speed, an interaction between emissions and wind speed and wind direction). In order to avoid repetition, only the results for the Burgher Street monitoring station will be presented in full, whereas the results for the other stations will be summarised. Similarly to the univariate GP models for Glasgow in Section 5.3, an upper boundary limit of 1000 is set for the hyperspatial range parameters. Once a final model is chosen based on 10-fold CV RMSPE, the emulator will be applied to explore the different meteorological conditions under which the NO_2 annual average regulation will be broken.

6.4.1 Modelling

Burgher Street

The Burgher Street fixed effect parameters estimated by the multivariate model with three inputs are summarised in Table 6.10 alongside the parameters from the three covariate univariate model from the **DiceKriging** package. Similarly to the Aberdeen case study, the fixed parameter estimates and their standard errors are very similar. The variances and hyperspatial range parameters from the two models are compared in Table 6.11. The variance estimates are very similar with the multivariate model having a slightly smaller variance. However, the hyperspatial range parameters are very different to each other. For the multivariate model, the three inputs have very similar estimates suggesting they contribute quite evenly, whereas the univariate model places the biggest impact on wind direction, followed by emissions and wind speed is last.

The diagnostic plots for the multivariate fit at Burgher Street are then examined in Figure 6.21. The diagnostic plots do not indicate any issues with the fit. The points on the residuals vs. fitted values plot in a) are randomly scattered. The plots in e), f)

Coefficient	**Estim. Mult**	**St. Error Mult.**	**Estim. DK**	**St. Error DK**
Intercept	18.63	0.26	18.59	0.30
Emissions	0.10	0.003	0.10	0.003
Wind Speed	-0.05	0.006	-0.05	0.009
Wind Direction	0.02	0.008	0.02	0.004

TABLE 6.10: Summary of the multivariate Bayesian and univariate frequentist GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

Model	$\widehat{\theta}_{\textbf{Emissions}}$	$\widehat{\theta}_{\textbf{Wind Speed}}$	$\widehat{\theta}_{\textbf{Wind Direction}}$	$\widehat{\sigma}^2$
Multivariate Bayesian	65.47	75.86	57.67	0.14
Univariate Frequentist	104.86	50.12	312.01	0.16

TABLE 6.11: Summary of the hyperspatial range parameters and variance from the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

and g) of the residuals vs. each of the inputs also show random scatter. Similarly to the univariate model, the qq-plot in c) has heavy tails but a symmetric histogram in d) suggests that there is no issue with the fit, especially given that the residuals go from -0.25 to 0.25, which is a very small span and suggest an almost perfect fit. Additionally, a Shapiro-Wilk test was performed and the p-value was estimated to be 0.05 indicating that there is no evidence of non-normality of the residuals. This is further confirmed by the fact that the actual vs. fitted points in plot b) are all lying on the equivalence line.

Lastly, the multivariate Bayesian 10-fold CV RMSPE is compared to the one from the univariate model from the **DiceKriging** package in Table 6.12. The two models have almost identical prediction performance suggesting that although very different set of hyperspatial parameters were used, the models have very similar prediction estimates. This is further confirmed by the fact the 95% bootstrap confidence intervals overlap each other.

Model	RMSPE
Multivariate Bayesian	0.10 (0.08, 0.12)
Univariate Frequentist	0.09 (0.07, 0.10)

TABLE 6.12: RMSPE from a 10-fold CV of the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

However, in the univariate modelling case, the fit was improved by adding additional terms to the model. Hence, the multivariate model was refitted by adding a fourth

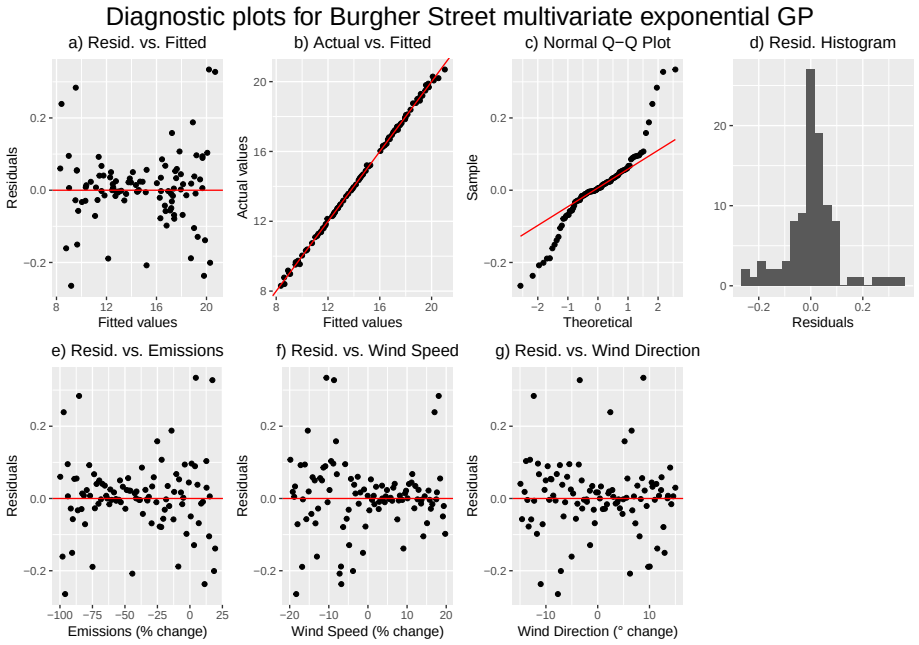

FIGURE 6.21: Diagnostic plots for the multivariate Bayesian GP model with an exponential kernel for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), wind speed (% change) and wind direction (° change) as covariates.

term - the emissions squared. The fixed effect parameters are summarised in Table 6.13. Similarly, to the model with three covariates, the parameter estimates and their standard errors are almost identical. The variances and the hyperspatial range parameters from the two models are compared in Table 6.14. The variance estimates are identical but the hyperspatial range parameters continue to be very different to each other in a similar way to the models with just three covariates. The multivariate Bayesian model estimates that emissions has the strongest effect and wind speed and wind direction are very similar to each other, whereas the univariate frequentist model puts the biggest emphasis on wind direction, followed by emissions and a very small contribution from wind speed.

Coefficient	**Estim. Mult**	**St. Error Mult.**	**Estim. DK**	**St. Error DK**
Intercept	18.76	0.20	18.74	0.23
Emissions	0.08	0.005	0.08	0.005
Emissions^2	-0.0002	0.00006	-0.0002	0.00005
Wind Speed	-0.05	0.006	-0.05	0.008
Wind Direction	0.02	0.007	0.02	0.004

TABLE 6.13: Summary of the GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

The diagnostic plots for the multivariate Bayesian model with four covariates for Burgher Street are examined in Figure 6.22. Overall, the plots indicate a good fit. The problems with heavy tails on the qq-plot in plot c) is continued from the three inputs model and

Model	$\widehat{\theta}_{\text{Emissions}}$	$\widehat{\theta}_{\text{Wind Speed}}$	$\widehat{\theta}_{\text{Wind Direction}}$	$\widehat{\sigma}^2$
Multivariate Bayesian	61.42	48.44	44.05	0.10
Univariate Frequentist	104.88	32.33	203.64	0.10

TABLE 6.14: Summary of the hyperspatial range parameters and variance from the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

the Shapiro-Wilk test has a p-value of 0.04 indicating the residuals are not normally distributed, but the histogram of the residuals in plot d) is symmetric and the range of the residuals (-0.25 to 0.30) is very small to indicate a problem with the fit. Furthermore, the actual vs. fitted points in plot b) are lying just as closely to the equivalence line as for the model with three inputs.

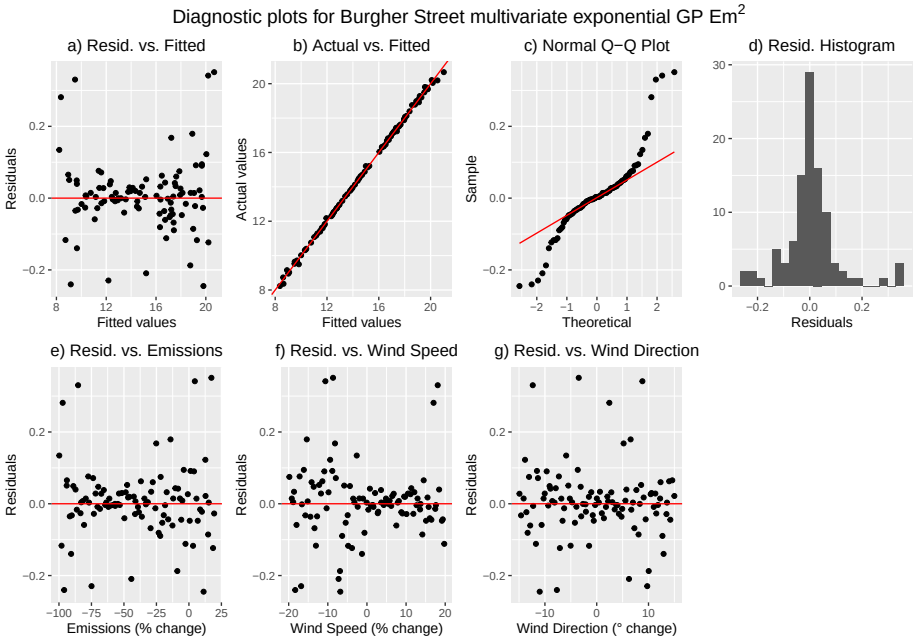

FIGURE 6.22: Diagnostic plots for the multivariate Bayesian GP model with an exponential kernel for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates.

Then, the multivariate Bayesian and the univariate frequentist models with four covariates are compared on the prediction performance using the RMSPE based on a 10-fold CV. Once again, the univariate frequentist model performs better, although not significantly different as the 95% bootstrap confidence intervals are overlapping. However, it is interesting to note that in comparison with the three covariates models (Table 6.12), the multivariate Bayesian model has a slightly lower RMSPE, whereas the frequentist model has the same RMSPE.

Model	RMSPE
Multivariate Bayesian	0.10 (0.07, 0.12)
Univariate Frequentist	0.08 (0.06, 0.09)

TABLE 6.15: RMSPE from a 10-fold CV of the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates and an exponential kernel.

Lastly, the multivariate Bayesian and univariate frequentist models are refitted by adding an additional fifth covariate - the interaction between emissions and wind speed. The fixed effect estimates and their respective standard errors from the two models are compared in Table 6.16. The parameter estimates and their standard errors from the two models are very similar to each other as with the other models. However, it has to be noted that the standard errors for both models are lower than those for the models with three and four covariates. Furthermore, the variances and hyperspatial range parameters were examined in Table 6.17. The two models have identical variances, which are the lowest variances in comparison to the models with three and four covariates. While the hyperspatial range parameters estimated by the multivariate Bayesian model have not changed much with the addition of an extra covariate, the hyperspatial range parameters estimated by the univariate frequentist model are very different. The emissions hyperspatial range parameter is now the largest for both models but the models disagree whether wind speed (multivariate Bayesian) or wind direction (univariate frequentist) has the second largest impact.

Coefficient	**Estim. Mult**	**St. Error Mult.**	**Estim. DK**	**St. Error DK**
Intercept	18.81	0.13	18.80	0.23
Emissions	0.08	0.003	0.08	0.005
Emissions^2	-0.0002	0.00004	-0.0002	0.00005
Wind Speed	-0.09	0.005	-0.08	0.0001
Emissions*Wind Speed	-0.0008	0.00007	-0.0008	0.00002
Wind Direction	0.02	0.005	0.02	0.004

TABLE 6.16: Summary of the multivariate Bayesian and univariate frequentist GP model for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

The diagnostic plots for the multivariate Bayesian model with five covariates for Burgher Street were examined in Figure 6.23. The points for residuals vs. fitted values plot in a) are fanning out to the right suggesting a heteroscedasticity issue. However, the points only fan out from -0.25 to 0.25, which is a very small span. The residuals vs. emissions plot in e) and residuals vs. wind speed in f) shown fanning to the right and

Model	$\widehat{\theta}_{\text{Emissions}}$	$\widehat{\theta}_{\text{Wind Speed}}$	$\widehat{\theta}_{\text{Wind Direction}}$	$\widehat{\sigma}^2$
Multivariate Bayesian	64.51	46.97	37.50	0.04
Univariate Frequentist	154.01	38.13	72.94	0.04

TABLE 6.17: Summary of the hyperspatial range parameters and variance from the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

left, respectively. This suggests that adding the interaction term between emissions and wind speed might be the cause for the fanning out. The qq-plot in c) shows heavy tails similarly to the models with three and four terms but the Shapiro-Wilk test has a p-value of 0.05 indicating that there is no evidence that the residuals are not normally distributed. However, the residuals histogram in d) is symmetric suggesting there is no issue. Lastly, the actual vs. fitted values plot in b) has the points lying perfectly on the equivalence line indicating the model's predictions are very accurate.

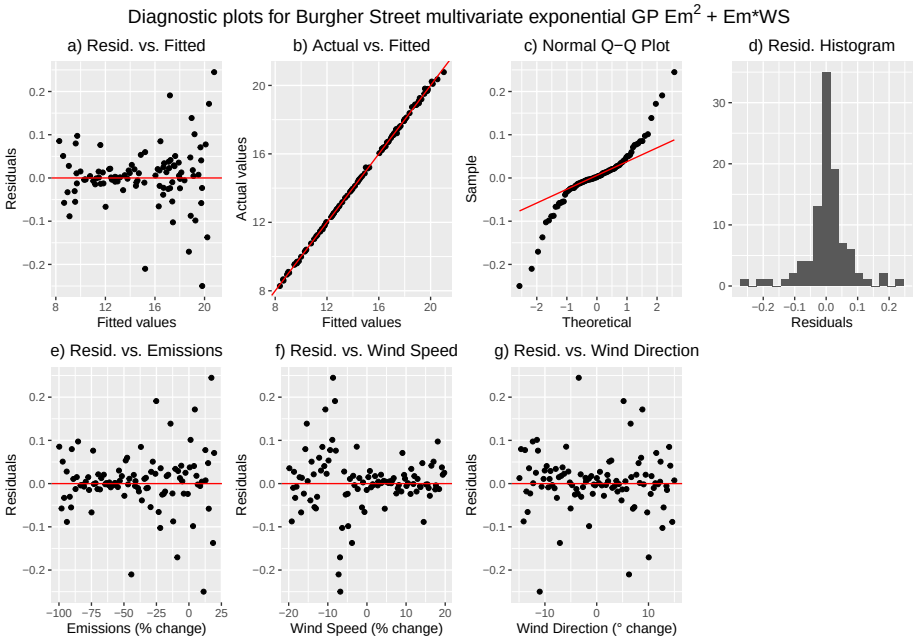

FIGURE 6.23: Diagnostic plots for the multivariate Bayesian GP model with an exponential kernel for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed and wind direction (° change) as covariates.

To compare the predictive performance of the multivariate Bayesian model to the univariate frequentist model, the 10-fold CV RMSPEs from the two models are compared in Table 6.18. The RMSPEs from both models are lower in comparison to the models with fewer covariates proving near perfect predictions and not statistically significantly different as the 95% bootstrap confidence intervals overlap each other. Nevertheless, the

univariate frequentist model has lower RMSPE and therefore, it is better in terms of predicting power.

Model	RMSPE
Multivariate Bayesian	0.07 (0.05, 0.08)
Univariate Frequentist	0.06 (0.04, 0.07)

TABLE 6.18: RMSPE from a 10-fold CV of the multivariate Bayesian and univariate frequentist models for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for Burgher Street with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction ($^\circ$ change) as covariates and an exponential kernel.

Other stations

Similarly to Burgher Street, the models with five covariates had the best RMSPEs and diagnostic plots. Therefore, the fixed effects parameter estimates for the five covariates multivariate Bayesian and univariate frequentist models are compared in Tables 6.19 and 6.20. For all stations, the parameter estimates and their standard errors are the same. The only exception is the parameter estimates for wind direction for the Byres Road and Central Station monitoring stations. The multivariate Bayesian model estimates them as positive, whereas the univariate frequentist model estimates them as negative. However, this is a result from the fact that the univariate modelling of the stations suggested that wind direction is not significant and is not needed for the modelling of these stations. However, wind direction is retained in the models due to the high hyperspatial range parameters estimated for it, which indicates wind direction is important for predictions.

The estimated hyperspatial range parameters by multivariate Bayesian and univariate frequentist models are compared in Table 6.21. For the emissions hyperspatial range parameter, the multivariate model estimates a much lower value than the univariate models with the exception of the Central Station model. The wind speed hyperspatial range parameter from the multivariate model appears to be an average value from those estimated by the univariate models. On the other side, there is a lot of variation in the wind direction estimates from the univariate models, yet the multivariate Bayesian model estimates a value close to the smallest one from the univariate models (for Waulkmillglen Reservoir).

Then the models are compared based on their overall variance σ^2 and the RMSPE based on a 10-fold CV in Table 6.22. The variance estimates are very similar. For Central Station and Dumbarton Road the multivariate Bayesian model has slightly lower variances, whereas the univariate frequentist model performs better for Byres Road, High

Station	Coefficient	Estim. Mult.	St. Error Mult.	Estim. DK	St. Error DK
Burgher Street	Intercept	18.81	0.13	18.80	0.23
	Emissions	0.08	0.003	0.08	0.005
	Emissions^2	-0.0002	0.00004	-0.0002	0.00005
	Wind Speed	-0.09	0.005	-0.08	0.0001
	Emissions*Wind Speed	-0.0008	0.00007	-0.0008	0.00002
	Wind Direction	0.02	0.005	0.02	0.004
Byres Road	Intercept	34.25	0.13	34.27	0.19
	Emissions	0.20	0.003	0.20	0.002
	Emissions^2	-0.0006	0.00003	-0.0006	0.00002
	Wind Speed	-0.20	0.005	-0.20	0.005
	Emissions*Wind Speed	-0.002	0.00006	-0.002	0.00006
	Wind Direction	0.0002	0.005	-0.0007	0.001
Central Station	Intercept	63.88	0.30	64.04	0.38
	Emissions	0.36	0.008	0.36	0.01
	Emissions^2	-0.002	0.00008	-0.002	0.0001
	Wind Speed	-0.33	0.01	-0.33	0.01
	Emissions*Wind Speed	-0.003	0.0002	-0.003	0.0002
	Wind Direction	0.001	0.01	-0.002	0.003
Dumbarton Road	Intercept	37.78	0.18	37.82	0.24
	Emissions	0.23	0.005	0.23	0.005
	Emissions^2	-0.0006	0.00005	-0.0006	0.00005
	Wind Speed	-0.23	0.007	-0.22	0.007
	Emissions*Wind Speed	-0.002	0.0001	-0.002	0.00009
	Wind Direction	-0.05	0.007	-0.05	0.007

TABLE 6.19: Summary of fixed effect parameters from the multivariate GP and the univariate frequentist emulator from the **DiceKriging** package for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for the Glasgow monitoring stations Burgher Street, Byres Road, Central Station and Dumbarton Road with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

Street, Townhead and Wellington Road. However, the RMSPEs are always lower for the univariate frequentist models, although not statistically significantly different as the 95% bootstrap confidence intervals are overlapping each other. This implies that using different hyperspatial range parameters for the different stations is better than using one smoothed set of hyperspatial range parameters for all stations. Nevertheless, the differences are very small and both models have almost perfect prediction.

Findings

Overall, the multivariate Bayesian emulator with five covariates performs better than the emulator with less covariates when modelling the NO_2 annual average from the ADMS-Urban simulator. The multivariate Bayesian model performs almost as well as the univariate frequentist model with five covariates. Applying the same hyperspatial

Station	Coefficient	Estim. Mult.	St. Error Mult.	Estim. DK	St. Error DK
Great Western Road	Intercept	33.92	0.14	33.95	0.16
	Emissions	0.19	0.004	0.20	0.003
	Emissions^2	-0.0006	0.00004	-0.0006	0.00004
	Wind Speed	-0.19	0.005	-0.19	0.006
	Emissions*Wind Speed	-0.002	0.00007	-0.002	0.00007
	Wind Direction	-0.02	0.005	-0.02	0.004
High Street	Intercept	35.60	0.14	35.63	0.14
	Emissions	0.19	0.004	0.20	0.003
	Emissions^2	-0.0008	0.00004	-0.0008	0.00003
	Wind Speed	-0.20	0.005	-0.20	0.006
	Emissions*Wind Speed	-0.002	0.00007	-0.002	0.00007
	Wind Direction	-0.002	0.00007	-0.002	0.002
Townhead	Intercept	29.49	0.13	29.50	0.12
	Emissions	0.15	0.003	0.15	0.003
	Emissions^2	-0.0007	0.00003	-0.0007	0.00003
	Wind Speed	-0.16	0.005	-0.16	0.005
	Emissions*Wind Speed	-0.001	0.00007	-0.002	0.00007
	Wind Direction	-0.01	0.005	-0.01	0.002
Waulkmillglen Reservoir	Intercept	9.81	0.07	9.81	0.04
	Emissions	0.01	0.002	0.01	0.0007
	Emissions^2	-0.00003	0.00002	-0.00004	0.000007
	Wind Speed	-0.01	0.003	-0.01	0.001
	Emissions*Wind Speed	-0.0001	0.00004	-0.0001	0.00001
	Wind Direction	-0.01	0.003	-0.01	0.002

TABLE 6.20: Summary of fixed effect parameters from the multivariate GP and the univariate frequentist emulator from the **DiceKriging** package for the NO_2 annual averages (μg m^{-3}) from ADMS-Urban for the Glasgow monitoring stations Great Western Road, High Street, Townhead and Waulkmillglen Reservoir with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

Model	$\widehat{\theta}_{\text{EM}}$	$\widehat{\theta}_{\text{WS}}$	$\widehat{\theta}_{\text{WD}}$
Multivariate Bayesian	64.51	46.97	37.50
Burgher Street Univariate Frequentist	154.01	38.13	72.94
Byres Road Univariate Frequentist	111.39	27.50	1000.00
Central Station Univariate Frequentist	60.96	69.30	1000.00
Dumbarton Road Univariate Frequentist	88.19	76.76	60.04
Great Western Road Univariate Frequentist	92.72	42.70	107.66
High Street Univariate Frequentist	85.98	32.14	290.09
Townhead Univariate Frequentist	99.39	28.91	271.92
Waulkmillglen Reservoir Univariate Frequentist	170.53	57.05	33.68

TABLE 6.21: Summary of the hyperspatial range parameter estimates from the multivariate Bayesian and univariate frequentist models for NO_2 annual average (μg m^{-3}) from ADMS-Urban for the Glasgow monitoring stations with emissions (% change), emissions squared wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel.

Station	Model	$\widehat{\sigma}^2$	RMSPE
Burgher Street	Multivariate Bayesian	0.04	0.07 (0.05, 0.08)
	Univariate Frequentist	0.04	0.06 (0.04, 0.07)
Byres Road	Multivariate Bayesian	0.04	0.05 (0.04, 0.06)
	Univariate Frequentist	0.03	0.04 (0.03, 0.04)
Central Station	Multivariate Bayesian	0.22	0.14 (0.10, 0.17)
	Univariate Frequentist	0.25	0.10 (0.07, 0.13)
Dumbarton Road	Multivariate Bayesian	0.08	0.09 (0.07, 0.11)
	Univariate Frequentist	0.11	0.08 (0.07, 0.10)
Great Western Road	Multivariate Bayesian	0.05	0.06 (0.05, 0.08)
	Univariate Frequentist	0.05	0.06 (0.05, 0.07)
High Street	Multivariate Bayesian	0.05	0.06 (0.05, 0.07)
	Univariate Frequentist	0.04	0.05 (0.04, 0.05)
Townhead	Multivariate Bayesian	0.04	0.06 (0.05, 0.06)
	Univariate Frequentist	0.03	0.04 (0.04, 0.05)
Wellington Road	Multivariate Bayesian	0.01	0.02 (0.01, 0.03)
	Univariate Frequentist	0.004	0.02 (0.01, 0.03)

TABLE 6.22: Summary of the variance and RMSPE from the multivariate Bayesian and univariate frequentist models for NO_2 annual average (μg m^{-3}) from ADMS-Urban for the Glasgow monitoring stations with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction ($^\circ$ change) as covariates and an exponential kernel.

range parameters to all stations causes a loss of information and therefore, the multivariate Bayesian model is out-performed in terms of predictive power by the univariate frequentist model. The univariate models have better (but not statistically significant) performance in comparison to the multivariate model due to the rapid fluctuation of pollution concentration even for small distances (less than 10 metres). However, using the multivariate Bayesian model is preferred as only one model is fitted for all monitoring stations rather than individual ones. Additionally, although the analysis shows that wind direction is not always significant for some of the monitoring stations in Glasgow (Byres Road and Central Station), it will be retained in the models due to the high hyperspatial range parameter estimated for wind direction and the experts' recommendation.

6.4.2 Emulation

In this subsection, the results from emulating the ADMS-Urban simulations for the multivariate Bayesian models for the Glasgow monitoring stations are examined. The emulated NO_2 annual average concentrations are obtained over a discretised grid of the input space from the LHC. The grid to be evaluated is based on increments of 0.5 along the dimensions for each of the three inputs. In addition to calculating the NO_2 annual average and the respective uncertainty around that estimate, the probability of breaking the 40 μg m^{-3} regulation will be provided in a similar way to the Aberdeen case. Contour plots for the emulated NO_2 annual average under three variations for wind direction will be provided alongside the respective standard deviations for the annual average as well as contour plots for the probability of exceeding the regulation limit of 40 μg m^{-3}. The three variations for wind direction to be examined are:

- -15° variation from the baseline value for 2015, resulting in a more eastern prevailing wind;
- 0° variation from the baseline value for 2015; and
- 15° variation from the baseline value for 2015, resulting in a more western prevailing wind.

Firstly, the contour plots for the emulated NO_2 annual average and the respective standard deviations when the wind direction is set to have -15° change from the 2015 observed baseline values are in Figure 6.24. Three of the stations (Burgher Street, Townhead and Waulkmillglen Reservoir) have values indicating that the simulated annual averages will both breach the annual average regulation of 40 μg m^{-3}. The Byres Road, Dumbarton Road, Great Western Road and High Street stations have a bit of red hue in their plots suggesting that under increased emissions and low wind speed, the annual average will be close to exceeding the regulation. For Central Station, the majority of the NO_2 annual averages are above the regulation. This implies that the station records really high emission values and regardless of the meteorological conditions, the annual regulation will be broken in the majority of the simulated conditions. Furthermore, for Central Station the highest emulated NO_2 annual average is 80 μg m^{-3} which is higher than any other station in both Aberdeen and Glasgow. The variation values for all stations, however, are quite low, indicating good emulation of the simulation scenarios. The overall standard deviation values for Glasgow are much smaller than those for Aberdeen which is due to the more complex fixed effects model, i.e. the larger number of covariates. Central Station and Dumbarton Road have the highest standard deviation values in Glasgow which is in accordance with the fact that those are the two stations with

the highest recorded concentrations. Overall, similar trends as in the Aberdeen plots are seen in the standard deviation plots for Glasgow. Firstly, the stations with higher NO_2 annual averages have higher standard deviations associated with them. Secondly, there is an increase in the standard deviations towards the bottom right corner which are associated with higher emissions and lower wind speeds. Lastly, the corners of each of the plots have higher standard deviations than the mid-points as these are the edges of the LHC, i.e. the standard deviations corresponds to the density of the simulation points.

The contour plots for the probabilities of breaking the 40 μg m^{-3} annual average regulation limit when the wind direction is set to -15° change from the 2015 observed baseline values are in Figure 6.25. As in the Aberdeen case, the change from probability of exceedance of zero to probability of exceedance of one is quite rapid due to the accuracy of the predictions. For Glasgow, there are three stations for which under no conditions the regulation limit is likely to be broken and they are Burgher Street, Townhead and Waulkmillglen Reservoir, which is consistent with the annual averages observed at Figure 6.24. For the Byres Road and Great Western Road stations, increases of emissions over 10% and lower than the baseline wind speed are likely to result in exceedance of the regulation, whereas for High Street any increase of emissions and lower than the baseline wind speed are likely to result in exceedances. For Dumbarton Road, it appears that an increase in emissions, even for higher than the baseline wind speed could result in exceedances. For Central Station the plot suggests that a decrease of almost 60% in emissions will result in high probability of non-exceedance for any meteorological conditions. This is further supported by the almost horizontal line between the probability of zero and probability of one at the contour plot.

Next, the contour plots for the emulated NO_2 annual average and the respective standard deviations when the wind direction is set to 0° change from the 2015 baseline are in Figure 6.26. As in the case when wind direction was set at -15° change from the baseline in Figure 6.24, there are three stations for which values of 40 μg m^{-3} are never observed - Burgher Street, Townhead and Waulkmillglen Reservoir. For Byres Road, Dumbarton Road, Great Western Road, High Street and Townhead it appears that for a number of combinations of emissions and wind speed, there are a few annual averages which are exceeding the regulation limit. For Central Station, the majority of the values are above the regulatory limit and reach values of 80 μg m^{-3}, which is twice the limit. The standard deviations are again quite small with the highest standard deviations at Central Station and Dumbarton Road, where the highest NO_2 concentrations are recorded.

The contour plots for the probabilities of breaking the 40 μg m^{-3} annual average regulation limit when the wind direction is set to 0° change from the 2015 baseline are in

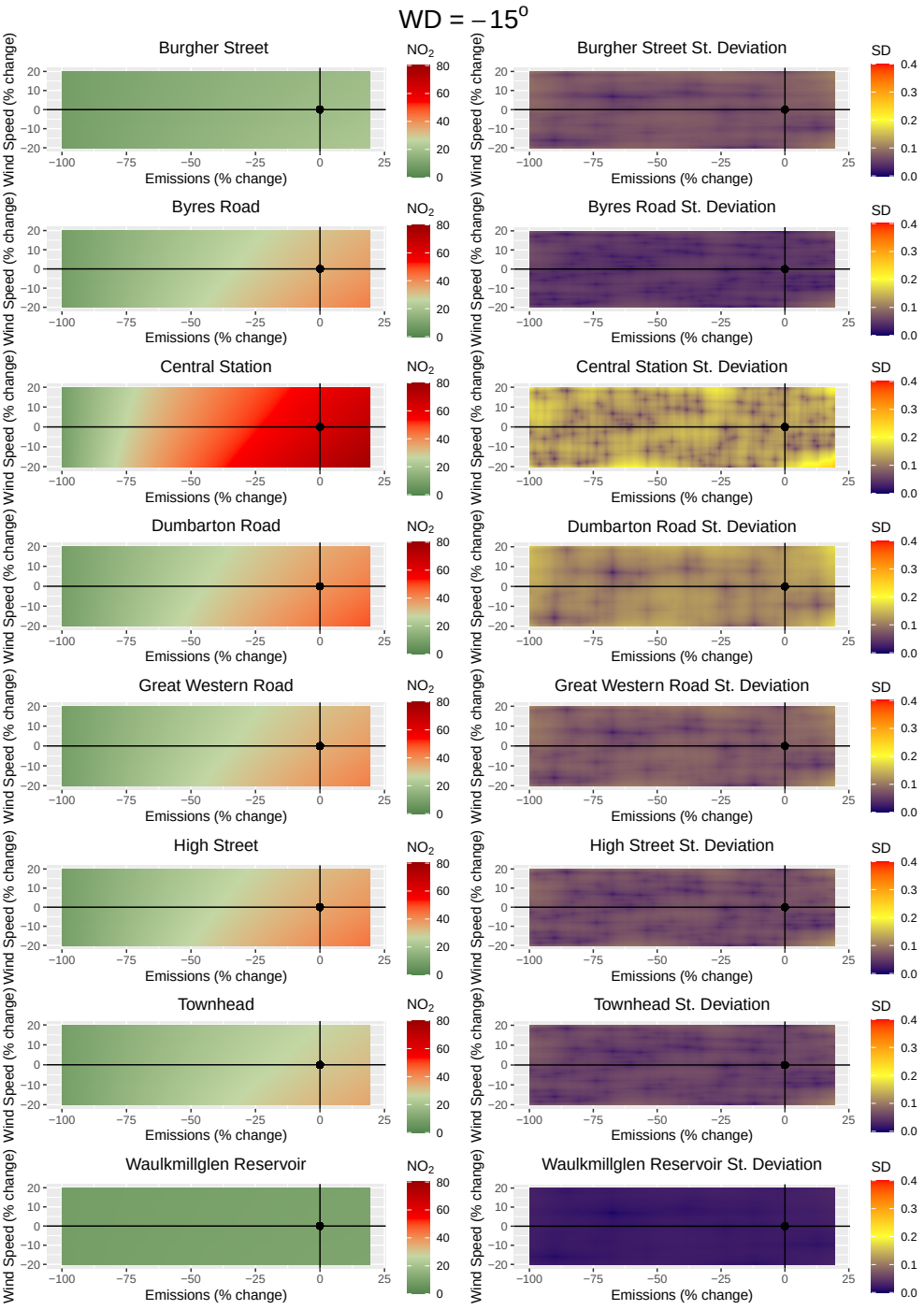

FIGURE 6.24: Contour plots for emulated NO_2 annual average (μg m^{-3}) when wind direction is set to -15° change from the 2015 baseline for the ADMS-Urban simulations for the eight monitoring stations in Glasgow based on the multivariate model with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel are presented on the left side. Contour plots for the standard deviations of the emulated NO_2 annual averages are provided on the right. In each plot, the black circle depicts the baseline realisation.

Figure 6.27. Similarly to the contour plots for the probabilities of exceedance when the wind direction is set to -15° change from the baseline, the change from probability of

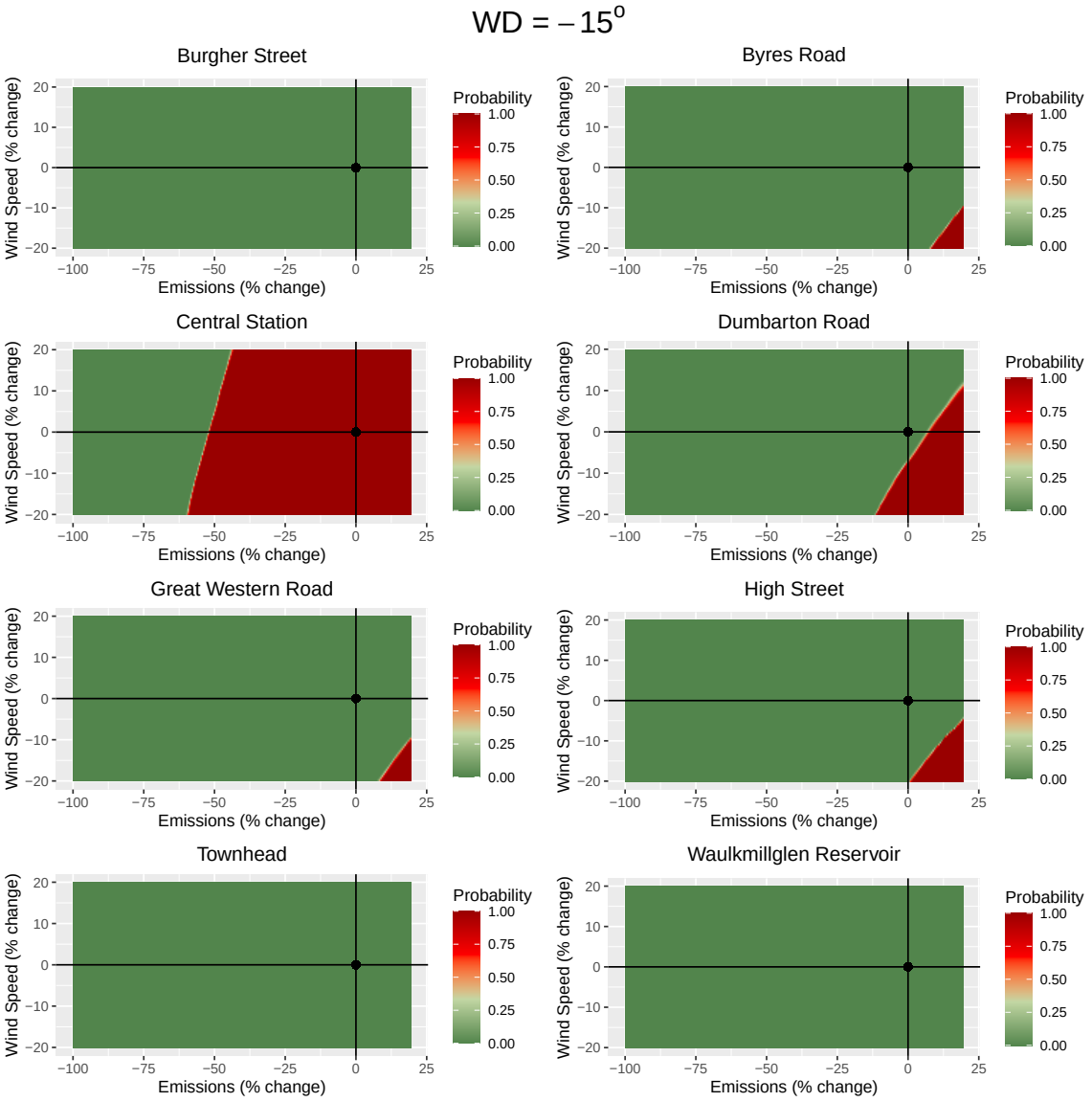

FIGURE 6.25: Contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average (μg m^{-3}) when wind direction is set to -15° change from the 2015 baseline for the ADMS-Urban simulations for the eight monitoring stations in Glasgow based on the multivariate model with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel. In each plot, the black circle depicts the baseline realisation.

exceedance of zero to probability of exceedance of one is quite rapid due to the accuracy of the predictions. Once again, it is unlikely that the regulations will be broken at the Burgher Street, Townhead and Waulkmillglen Reservoir monitoring stations. For Byres Road, Great Western Road, and High Street as long as the emissions are not increased, the annual average regulation will not be exceeded. For Dumbarton Road, it appears that if the emissions are higher than the baseline and the wind speed is lower, it is very likely that the regulation will be exceeded. For Central Station, the only way to ensure that the regulation will not be broken is to reduce the emissions by 60% as seen by the

almost horizontal line between the probability of zero to probability of one. Overall, the prediction plots are very similar to those in Figure 6.25 suggesting that the change in wind direction does not appear to have an effect on the probability that the annual average would exceed the regulatory limit.

Lastly, the contour plots for the emulated NO_2 annual average and the respective standard deviations when the wind direction is set to 15° change from the 2015 baseline are in Figure 6.28. The plot is very similar to those for the wind direction being set to -15° change from the baseline in Figure 6.24 and to 0° change from the baseline in Figure 6.26. As in the previous plots, Burgher Street, Townhead and Waulkmillglen Reservoir do not have annual averages above 40 μg m^{-3}, whereas for Byres Road, Dumbarton Road, Great Western Road and High Street, there are values above the the annual average limit for NO_2. For Central Station, the majority of the plot is red suggesting that the annual average limit is likely to be exceeded under most simulation conditions. Overall, the standard deviations are quite small with the highest standard deviations at Central Station and Dumbarton Road, where the highest NO_2 concentrations are recorded.

The contour plots for the probabilities of breaking the 40 μg m^{-3} annual average regulation limit when the wind direction is set to 15° change from the baseline are in Figure 6.29. Similarly to the previous contour plots for the probabilities of exceedance, the change from probability of exceedance of zero to probability of exceedance of one is quite rapid due to the accuracy of the predictions. For these contour plots, it appears that it is unlikely that the regulations will be broken at the Burgher Street, Townhead and Waulkmillglen Reservoir monitoring stations. For Byres Road, Great Western Road, and High Street as long as the emissions are not higher than the baseline values, the annual average regulation will not be exceeded. For Dumbarton Road, it appears that if the emissions are higher than the baseline and the wind speed is lower than the baseline values, it is very likely that the regulation will be exceeded. For Central Station, the only way to ensure that the regulation will not be broken is to reduce the emissions by more than 60% as demonstrated by the almost horizontal line of separation between the probability of zero and probability of one for breaking the regulation limit. Overall, the prediction plots are very similar to those in Figure 6.25 and Figure 6.27 suggesting that the change in wind direction does not appear to have an effect on the probability that the annual average would exceed the regulatory limit.

Findings

The plots from the three sets of variation of wind direction (-15°, 0° and 15°) are very similar to each other suggesting that wind direction does not have a visible effect

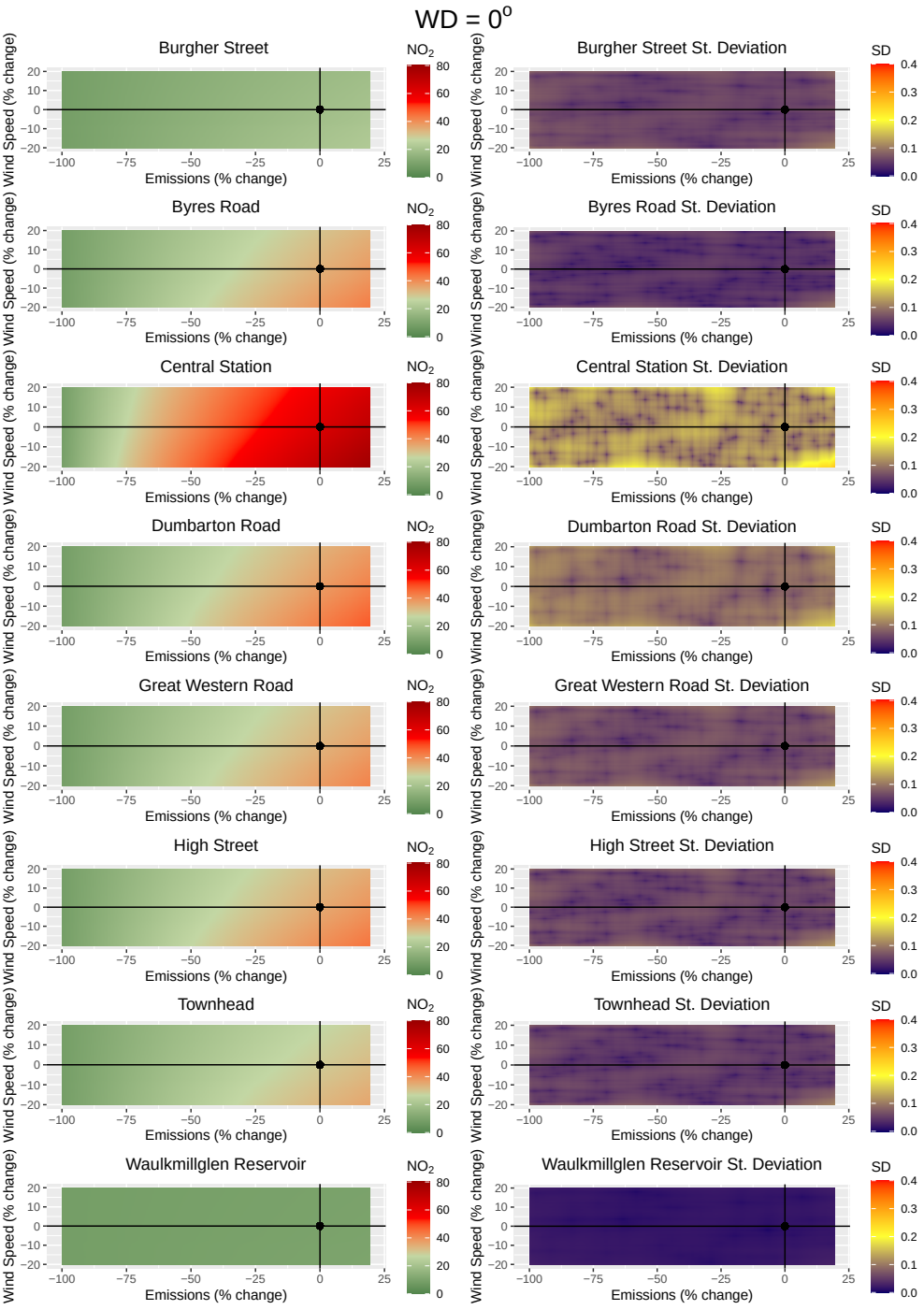

FIGURE 6.26: Contour plots for emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 0° change from the 2015 baseline for the ADMS-Urban simulations for the eight monitoring stations in Glasgow based on the multivariate model with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel are presented on the left side. Contour plots for the standard deviations of the emulated NO_2 annual averages are provided on the right. In each plot, the black circle depicts the baseline realisation.

on the NO_2 annual average concentrations. The plots help identify Burgher Street, Townhead and Waulkmillglen Reservoir as the three stations at which under no set of

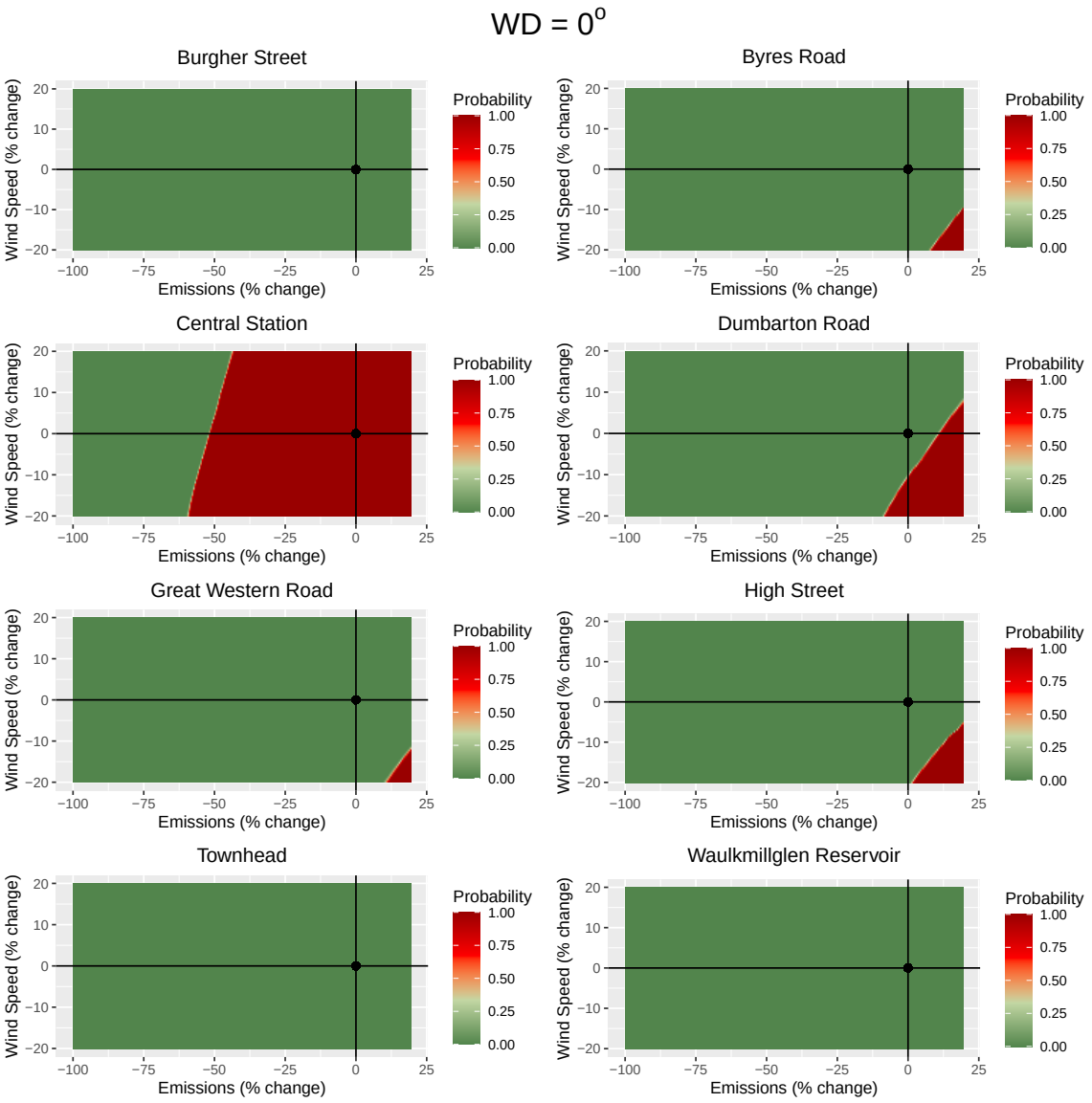

FIGURE 6.27: Contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 0° change from the 2015 baseline for the ADMS-Urban simulations for the eight monitoring stations in Glasgow based on the multivariate model with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel. In each plot, the black circle depicts the baseline realisation.

different emission levels and meteorological conditions, the annual average regulation of 40 μg m^{-3} will be broken. For the Byres Road, Great Western Road and High Street monitoring stations it was shown that as long as the emission levels do not exceed the baseline values, the NO_2 annual average regulation limit will not be broken. For Dumbarton Road, a small reduction of about 5% in emissions will ensure that the regulation is never exceeded. However, Central Station is the monitoring station where the meteorological conditions appear not to have as much of an effect and only a large reduction (of above 60%) in emissions will ensure that monitoring station will comply

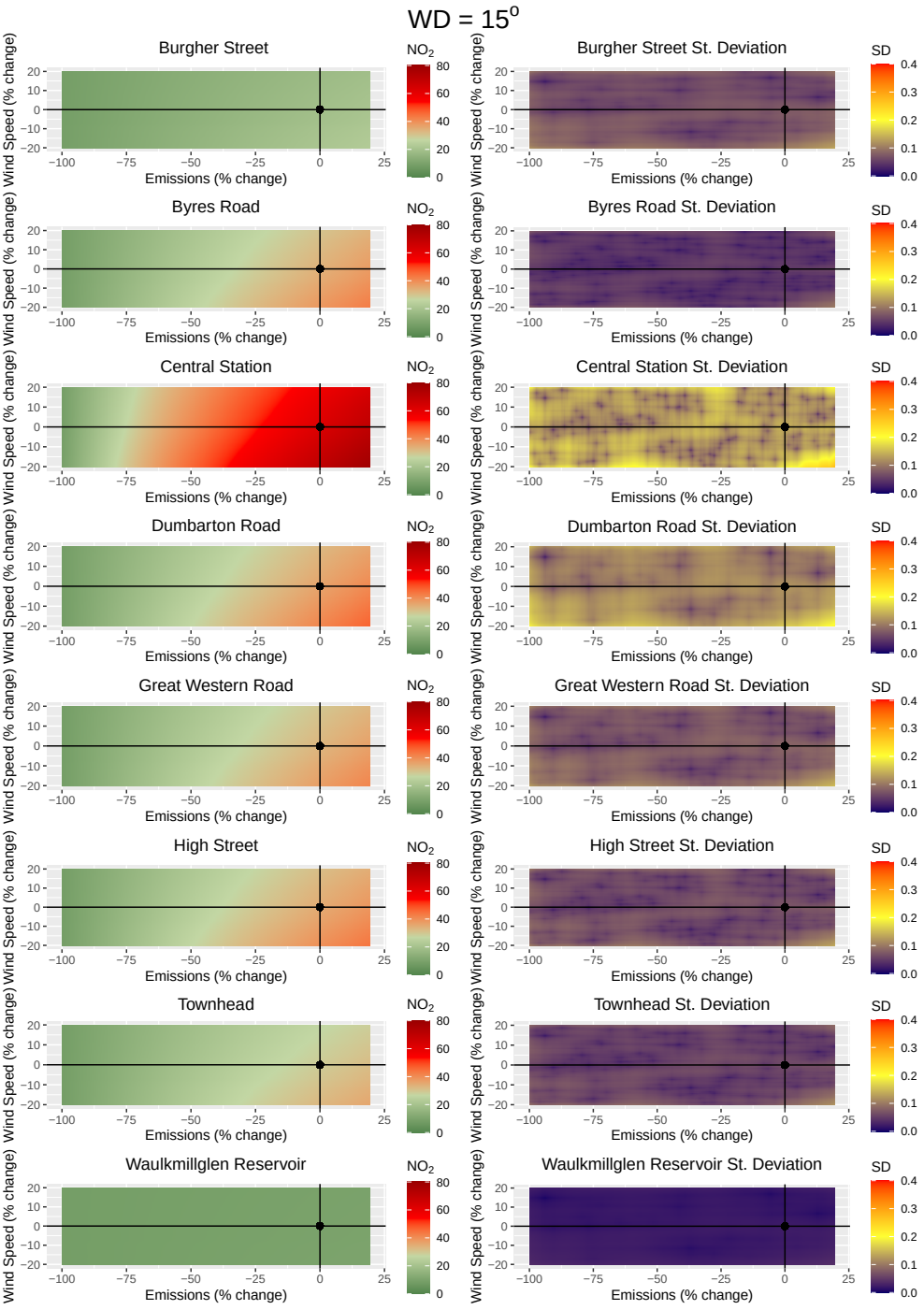

FIGURE 6.28: Contour plots for emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 15° change from the 2015 baseline for the ADMS-Urban simulations for the eight monitoring stations in Glasgow based on the multivariate model with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel are presented on the left side. Contour plots for the standard deviations of the emulated NO_2 annual averages are provided on the right. In each plot, the black circle depicts the baseline realisation.

with the regulation.

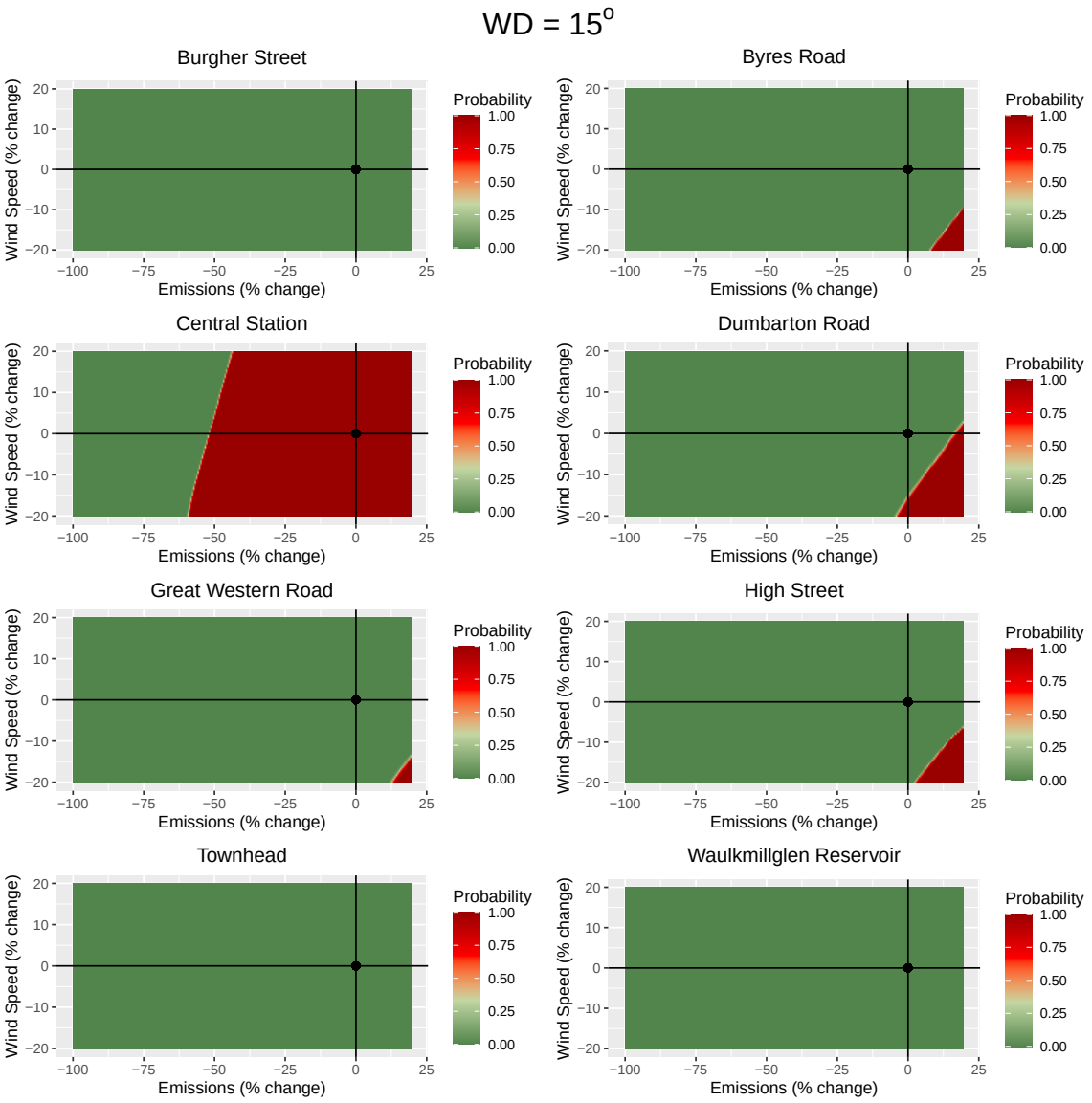

FIGURE 6.29: Contour plots for the probabilities for exceeding the 40 μg m^{-3} for the emulated NO_2 annual average (μg m^{-3}) when wind direction is set to 15° change from the 2015 baseline for the ADMS-Urban simulations for the eight monitoring stations in Glasgow based on the multivariate model with emissions (% change), emissions squared, wind speed (% change), an interaction between emissions and wind speed, and wind direction (° change) as covariates and an exponential kernel. In each plot, the black circle depicts the baseline realisation.

6.4.3 Conclusion

The multivariate Bayesian model was applied to the ADMS-Urban NO_2 annual average simulations and its performance was compared to the univariate frequentist model from the **DiceKriging** software. It was found that a model with five covariates (more complex than Aberdeen) provides the most accurate predictions. The multivariate and the univariate models estimate the parameters for the fixed effects as well as their standard

errors almost identically. The major difference comes from the hyperspatial range parameter estimates. The univariate frequentist models for the eight monitoring stations estimate very different hyperspatial range parameters for each stations and applying one set of parameters as estimated by the Bayesian multivariate model causes a slight loss of information but both models have almost identical and very close to zero RMSPEs.

Hence, the multivariate Bayesian model was used for the emulation of the NO_2 annual average for the eight monitoring stations in Glasgow as it requires fitting only one model instead of eight for each monitoring station. It was found that a change in the prevailing wind direction does not have an effect on the annual average concentrations. For three of the stations (Burgher Street, Townhead and Waulkmillglen Reservoir), the NO_2 annual average will never go above 40 μg m^{-3}. For another three of the stations (Byres Road, Great Western Road and High Street), as long as the emissions do not increase above the baseline, the regulatory limit will also not be broken. For Dumbarton Road, a reduction of 5% from the baseline emissions will result in complying with the regulation limit for all meteorological conditions. For Central Station, the biggest impact for complying with the regulation will be from reducing emissions by 60%.

6.5 Conclusion

In this chapter, a multivariate Bayesian emulator was introduced in Section 6.1 in order to create a multivariate model for the NO_2 annual averages as simulated by ADMS-Urban for Aberdeen and Glasgow, which accounts for the correlation between the observations. The multivariate Bayesian emulator models the responses using a matrix normal distribution, where the correlation matrix was built using an exponential correlation function. The same hyperspatial range parameters are imposed for all the stations in a city. Since there is no prior information available, non-informative priors are set.

In Section 6.2, simulation studies were performed in order to: (i) compare the ability of the multivariate Bayesian model and the univariate frequentist model from Chapter 5 to correctly identify the hyperspatial range parameters in settings close to real life data; and (ii) assess the predictive power of mis-estimating the hyperspatial range parameters by the multivariate Bayesian and univariate frequentist models. Additionally, the univariate simplified version of the multivariate Bayesian model was also applied to the data. The data for both studies were simulated in the same way. It was found that for (i), as the hyperspatial range parameters values increase, all models underestimate the hyperspatial range parameters with the multivariate Bayesian model underestimating the most. For (ii), it was found that RMSPE for setting the hyperspatial range parameters to their true values in comparison to setting them to half, double or letting the models

estimate the hyperspatial range parameters, there is at most 14% difference in the RMSPE. The multivariate Bayesian model appears to perform best although the difference between the models was not statistically significant. Furthermore, as the hyperspatial range parameter values are increased, the difference between the RMSPEs from the different models get smaller. It was found that in a setting with similar hyperspatial range parameters, the multivariate Bayesian model provides the lowest RMSPE. The simulation studies in Section 6.2 proved that hyperspatial range parameters are hard to estimate correctly by all models, but all models provide "sensible" [50] estimates which provide a relatively unaffected RMSPE.

In Section 6.3, the multivariate Bayesian model was applied to ADMS-Urban simulations for Aberdeen and its performance was compared to the univariate frequentist models for each of the monitoring stations from Chapter 5, where the three inputs (emissions, wind speed and wind direction) forming the LHC were used as covariates. The fixed effect parameter estimates and the diagnostic plots for the two models are very similar, and although the hyperspatial parameter estimates were different, the two models have RMSPEs very close to zero and almost identical. Therefore, the multivariate Bayesian model was chosen as the preferred ones and use to emulate over a discretised grid in order to assess the changes in the NO_2 annual average concentration and estimate the probability of exceeding the 40 μg m^{-3} regulation. The results from varying the prevailing wind direction from eastern to western, there is a larger set of combinations of emissions and wind speed for which the regulation will be broken.

In Section 6.4, the multivariate Bayesian model was applied to ADMS-Urban simulations for Glasgow and its performance was compared to the univariate models from Chapter 5, where the final model had five covariates - the three inputs (emissions, wind speed and wind direction) forming the LHC, as well as an emissions squared term and an interaction between emissions and wind speed. As in the Aberdeen case, the fixed effect parameter estimates and the diagnostic plots for the two models are very similar, but the univariate frequentist models have lower (but not statistically significant) RMSPE values than the multivariate Bayesian model. However, the RMSPEs for both models were very close to zero and similar to each other. Therefore, the multivariate Bayesian model was chosen as it requires fitting only one model. It was used to create an emulator for the ADMS-Urban NO_2 annual average predictions by varying the prevailing wind direction. It was found that changing the wind direction does not have a visible effect as there was with Aberdeen. The probability of breaking the regulation is highly dependant on a decrease in emissions in comparison to the baseline value but in the majority of the cases this is also dependent on wind speed being at least the baseline value for 2015.

Overall, applying the multivariate Bayesian model to both Aberdeen and Glasgow proved to be more beneficial than univariate modelling each station as different hyperspatial range parameters are required for each of the monitoring stations as it requires fitting multiple models. The fixed effect model for Glasgow was more complex than the one for Aberdeen, which resulted in lower standard deviations for the emulated Glasgow NO_2 annual averages - with 0.40 and 0.30 respectively for Aberdeen and Glasgow. However, the predictions for the annual average concentrations for both cities come with high accuracy as the largest standard deviation was 0.40. The multivariate Bayesian model provides an almost perfect emulator for the ADMS-Urban NO_2 annual average which is consistent with the fact that ADMS-Urban is a deterministic model. However, the NO_2 regulation is also applied to the hourly concentrations as seen in the following chapters.

Chapter 7

Modelling and emulation of the number of NO_2 hourly exceedances in Glasgow

In Scotland, there are regulations (as stipulated in European Directive 2008/50/EC [76]) on both the NO_2 annual average concentration and the NO_2 hourly concentration limits. In Table 1.1, it was stated that the annual average concentration cannot exceed 40 μg m^{-3}, whereas the hourly concentration could not go over 200 μg m^{-3} more than 18 times a year. In 2019, SEPA (in coordination with other governmental agencies) began a phased introduction of a Low Emission Zone (LEZ) in Glasgow City Centre to try and reduce the NO_2 concentrations. As a result of this, it is crucial to quantify the change in NO_2 hourly concentrations based on reduced emissions for the different monitoring stations in Glasgow. In Chapter 4 (more specifically, Table 4.4), it was shown that for 2015 there were four breaches of the 200 μg m^{-3} regulation limit, all of which occurred at the Central Station monitor. Therefore, it is of interest to examine the changes in the number of exceedances of the hourly limit of 200 μg m^{-3} for varying emissions, wind speed and wind direction using ADMS-Urban simulations. The chapter is organised as follows: Section 7.1 explores the number of NO_2 hourly concentrations exceeding the 200 μg m^{-3} for one hundred ADMS-Urban simulations for each of the eight monitoring stations in Glasgow. Section 7.2 presents the methodology of applying a Poisson generalised linear model (GLM) and quasi-Poisson GLM. Section 7.3 models the number of hourly concentrations above 200 μg m^{-3} for Central Station in Glasgow. Section 7.4 presents an emulator for the number of hourly concentrations above 200 μg m^{-3} for Central Station in Glasgow. Lastly, Section 7.5 provides a concluding discussion.

7.1 Site choice

The ADMS-Urban simulations can be used to examine under which scenarios the EU hourly regulation have been met and when the EU hourly regulation have been broken. Hence, the simulation scenarios for each of the monitoring stations based on the LHC design (varying emissions, wind speed and wind direction) can be used again but this time to model the number of exceedances of the hourly regulation. The number of exceedances above 200 μg m^{-3} for every scenario for every station are visualised in Figure 7.1. From the plots, it is clear that for only three monitoring stations (Central Station, Dumbarton Road and Great Western Road), there are simulations for which the hourly limit of 200 μg m^{-3} has been breached. For Central Station, there are 66 scenarios where at least one hourly observation is above 200 μg m^{-3}, for Dumbarton Road there are 42 such scenarios and 11 for Great Western Road. However, it is only at Central Station that the limit of 18 occurrences above 200 μg m^{-3} has been broken for 11 of the simulations.

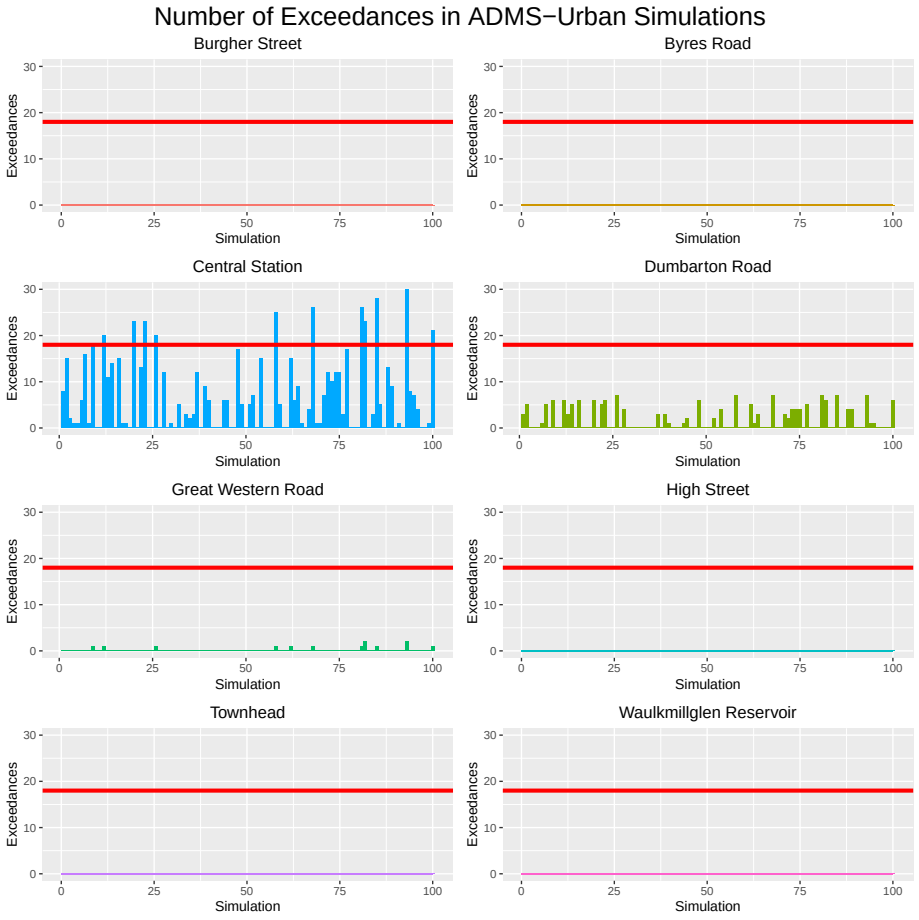

FIGURE 7.1: A barchart of the number of exceedances of the hourly 200 μg m^{-3} regulation in each scenario for each of the eight monitoring stations in Glasgow. A red line indicates the limit of 18 exceedances a year.

Since Central Station is the only monitoring station for which some of the simulation scenarios breach the hourly regulation (200 μg m^{-3} is exceeded more than 18 times), it is identified as the only station of interest. The following sections aim to answer what is the expected number of hourly concentrations above 200 μg m^{-3} given the input emissions, wind speed and wind direction by modelling the Central Station ADMS-Urban simulations.

7.2 Poisson generalised linear model

A common approach for modelling count data (such as the number of hourly concentrations above 200 μg m^{-3}) is a Poisson GLM. In Subsection 2.2.7, GLMs were introduced. This section presents the specific cases of Poisson and quasi-Poisson GLMs as described in [78]. A Poisson GLM assumes that the response y_i is a count of the number of events occurring in a fixed amount of space or time, and depends on a given $\mathbf{x}_i$. The link function $g(\cdot)$ is set to the log function. Hence, the model is:

$$y_i \sim \text{Po}(\mu_i)\,, \tag{7.1}$$

$$\log(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}\,. \tag{7.2}$$

Therefore, the Poisson log-likelihood is:

$$l(\mathbf{y};\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i(\mathbf{x}_i\boldsymbol{\beta}) - e^{\mathbf{x}_i\boldsymbol{\beta}} - \log\prod_{i=1}^{n} y_i!\,, \tag{7.3}$$

where $\boldsymbol{\beta}$ is estimated by minimising the likelihood using an iteratively reweighed least squares (IRWLS) algorithm. In terms of the parametrisation of the exponential family (Equation 2.46), the Poisson distribution has $a(y) = y_i$, $b(\mu) = \log(\mu_i)$, $c(\mu) = -\mu_i$ and $d(y,\phi) = -\log(y_i!)\phi$, where the dispersion parameter is assumed to be $\phi \equiv 1$. Hence, the score function for the Poisson model is:

$$U(\mathbf{y};\boldsymbol{\mu}) = \sum_{i=1}^{n}\left(\frac{y_i}{\mu_i} - 1\right)\,. \tag{7.4}$$

The deviance residuals (Equation 2.53) are used in two specific diagnostic plots for Poisson GLMs. The first plot is the halfnorm quantile plot, where the absolute values of the deviance residuals are plotted against the normal quantiles to check for outliers. The points are expected to lie in a straight line.

Additionally, the variances of the fitted values (approximated by $(\mathbf{y} - \widehat{\boldsymbol{\mu}})^2$) are plotted against the squared difference between the observed and fitted values (as an approximation to the variance of a given value). The points are expected to lie close to the equality line. If this is not the case, this indicates that the Poisson mean-variance assumption $\text{Var}(y_i) = \mathbb{E}(y_i)$ is not appropriate. If the points lie above the equivalence line, that indicates an overdispersion issue ($\text{Var}(y_i) > \mathbb{E}(y_i)$), whereas if the points lie below the equivalence line, that indicates an underdispersion issue ($\text{Var}(y_i) < \mathbb{E}(y_i)$).

Thus, the model can be extended to a quasi-likelihood Poisson log-linear model which assumes:

$$\mathbb{E}(y_i) = \exp(\mathbf{x}_i\boldsymbol{\beta})\,, \tag{7.5}$$

$$\text{Var}(y_i) = \phi\mathbb{E}(y_i)\,, \tag{7.6}$$

where ϕ is an estimated dispersion parameter. In the case of underdispersion $\phi < 1$ and in the case of overdispersion $\phi > 1$. Since the reason causing the dispersion is unknown, the dispersion parameter is estimated as:

$$\widehat{\phi} = \frac{1}{n-p}\sum_i \frac{(y_i - \widehat{y}_i)^2}{\widehat{y}_i}\,. \tag{7.7}$$

Then the quasi-Poisson model has a score function:

$$U(\mathbf{y}, \boldsymbol{\mu}) = \sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{\widehat{\phi}\text{Var}(\mu_i)}\right)\,. \tag{7.8}$$

Furthermore, $\widehat{\phi}$ is used to adjust the standard errors for the parameters for the quasi-Poisson model fit as $\text{Var}(\widehat{\boldsymbol{\beta}}) = \text{diag}\left((\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\widehat{\phi}\right)$.

7.3 Regression modelling of the number of exceedances over 200 μg m^{-3}

In order to estimate the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year, the number of exceedances at Central Station per ADMS-Urban scenario are plotted against each of the inputs in the LHC (emissions, wind speed and wind direction) in Figure 7.2. For emissions and wind speed, the percentage change is used, whereas for wind direction the degree change in direction is noted. Both the wind speed and wind direction plots (plots b) and c) in Figure 7.2) indicate random scatter, suggesting there

is no clear trend on how hourly NO_2 exceedances above 200 μg m^{-3} occur based on the percentage changes in wind speed and direction. However, from plot a) in Figure 7.2, it is clear that as emissions are increased, the number of exceedances above 200 μg m^{-3} increases. There is a clear pattern that for emissions variability below -60%, there are no occurrences of hourly NO_2 concentrations above 200 μg m^{-3}, whereas for emissions variability above -60%, there appears to be a quadratic trend. This suggests that there might be a breakpoint in the relationship.

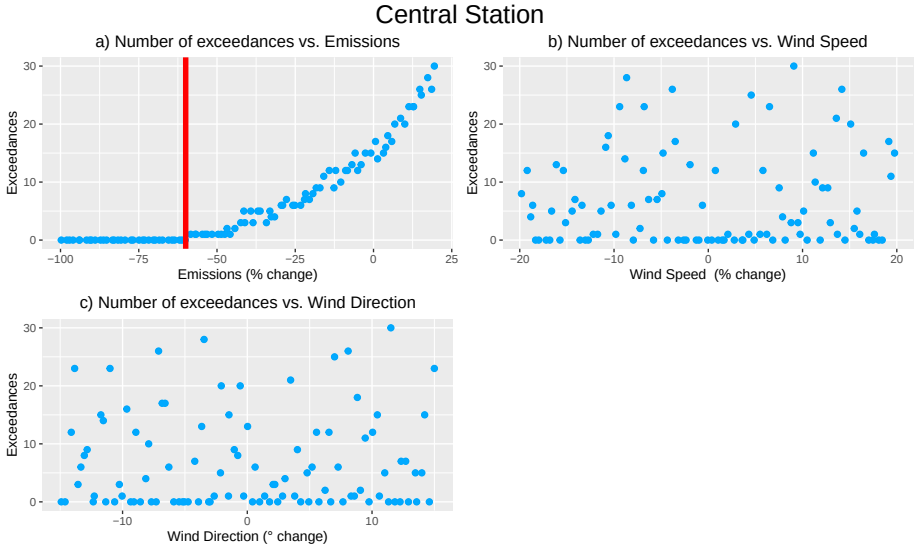

FIGURE 7.2: Scatterplots for the number of exceedances of the NO_2 hourly concentrations above 200 μg m^{-3} in a year per ADMS-Urban scenario against the LHC inputs (emissions (% change), wind speed (% change), and wind direction (° change)).

Additionally, in order to examine any interactions between the three LHC inputs, each input was plotted against the other two and the diameters of the points in the plot are proportional to the number of hourly NO_2 exceedances above 200 μg m^{-3} in Figure 7.3. Plots a) and b) in Figure 7.3 indicate a clear pattern that as the percentage change in emissions is increased, the number of exceedances above 200 μg m^{-3} are also increased. However, in the last plot c), the points are randomly scattered by size indicating no clear pattern. The proportional points plots further highlight that the increase in emissions is the main reason for increased number of NO_2 hourly concentrations exceeding 200 μg m^{-3}. Overall, the plots in Figure 7.3 suggest that there is no need for interaction terms between the three inputs.

Based on this exploratory analysis, there are two possible models that can be fitted - a Poisson generalised linear model (Subsection 7.2) or a segmented (broken-stick) Poisson generalised linear model. Firstly, the Poisson generalised linear model (GLM) is fitted with emissions, emissions squared, wind speed and wind direction as covariates. Although there does not appear to be a relationship between the number of exceedances and wind speed and wind direction, the two covariates are included in the model to

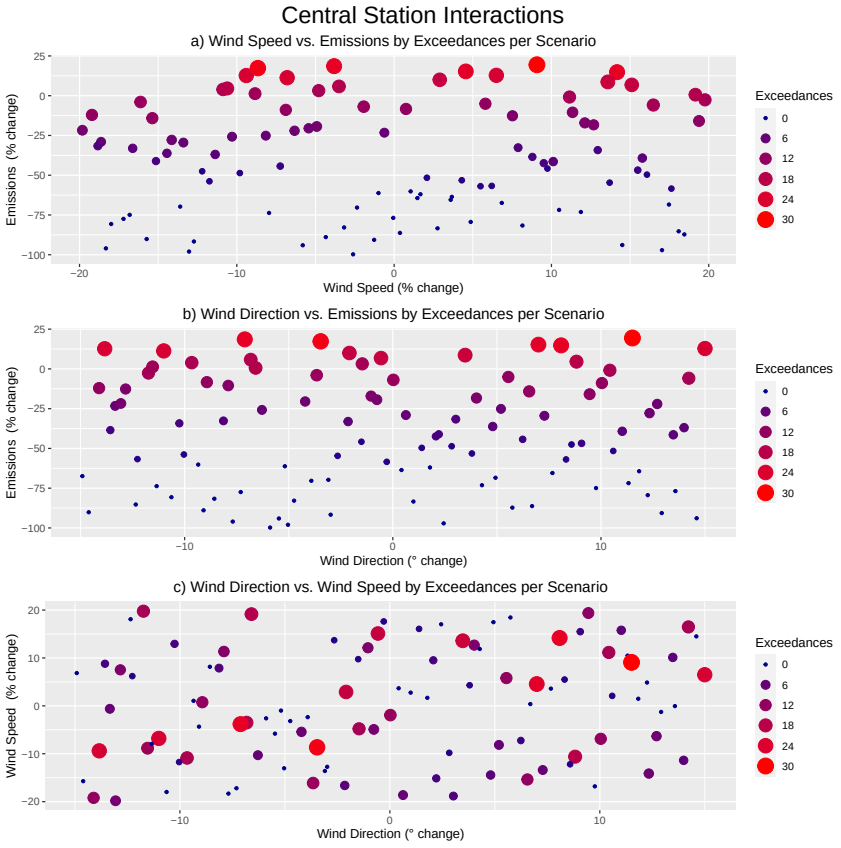

FIGURE 7.3: Scatterplots for the LHC inputs (emissions (% change), wind speed (% change) and wind direction (° change)) at Central Station against each other. The points are proportional to the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year per ADMS-Urban scenario.

formally assess their effect on the number of exceedances. Additionally, it is important when presenting the results from the modelling in front of legislative bodies (e.g. Scottish government) to include wind speed and wind direction to demonstrate their impact or lack of such impact. Let y_i $(i = 1, \ldots, 100)$ be the number of NO_2 hourly exceedances above 200 μg m^{-3}, then $y_i \sim \text{Po}(\mu_i)$. Then, the fitted log-link Poisson GLM (Subsection 7.2) is:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{x}_{\text{EM i}} + \beta_2 \text{x}^2_{\text{EM i}} + \beta_3 \text{x}_{\text{WS i}} + \beta_4 \text{x}_{\text{WD i}}\,, \tag{7.9}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^\top$ is the set of parameters to be estimated, $\text{x}_{\text{EM i}}$ and $\text{x}_{\text{WS i}}$ are the percentage changes in emissions and wind speed respectively, and $\text{x}_{\text{WD i}}$ is the degree change in wind direction for scenario i. The model was fitted using the *glm* function in R.

The diagnostic plots for the model are presented in Figure 7.4. The deviance residuals against wind speed (plot f)) suggest a quadratic curve. Therefore, the model is refitted

to check whether a squared term for wind speed should also be included. The new model fitted is:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{\text{EM i}} + \beta_2 x^2_{\text{EM i}} + \beta_3 x_{\text{WS i}} + \beta_4 x^2_{\text{WS i}} + \beta_5 x_{\text{WD i}} \,. \tag{7.10}$$

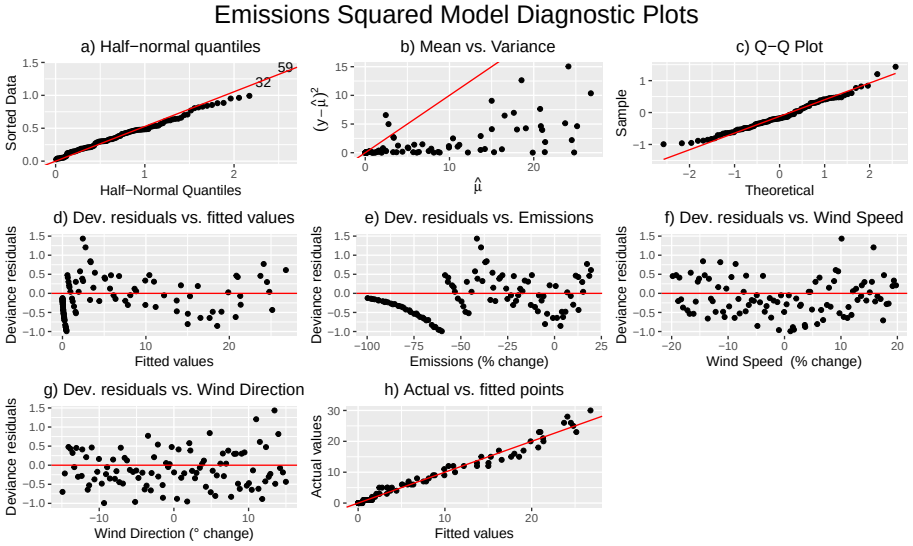

FIGURE 7.4: Diagnostic plots for the Poisson GLM for the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year (as simulated by ADMS-Urban for Central Station) with emissions (% change), emissions squared, wind speed (% change) and wind direction (° change) as covariates.

The diagnostic plots for the model with emissions squared term are presented in Figure 7.5. The half-normal quantiles (plot a), see Subsection 7.2) shows no outliers and indicates that the structural form of the model is explaining the variation well. Furthermore, the proportional deviance explained by the model is 97.68%. The points on the qq-plot (plot c)) lie on the equivalence line suggesting linearity can be assumed and the structural form of the model is appropriate. The deviance residuals vs. the fitted values are randomly scattered and there does not appear to be a problem with the fit. The deviance residuals vs. emissions (plot e)) has a number of points in a curved line for emissions values between -100 and -60. This is not surprising given that the observed exceedances for these emissions are all zero. However, given the small spread of the deviance residuals (between -1.0 and 1.5), there is no indication of an issue with the model fit. The deviance residuals against wind speed (plot f)) no longer exhibit a quadratic curve, although the wind speed squared term is not significant. Similarly, the deviance residuals against wind direction (plot g)) show no patterns. The actual vs. fitted values points (plot h)) lie on the equivalence line, thus indicating the model predictions are similar to the actual values. The only plot indicating an issue is the mean vs. variance (plot b)), where a strong underdispersion is observed as the majority

of the points lie below the equivalence line, which suggests that the assumption for equal mean and variance is broken. As a result of that, the model has to be refitted with a dispersion parameter different from 1 in order to produce realistic standard errors (see Subsection 7.2).

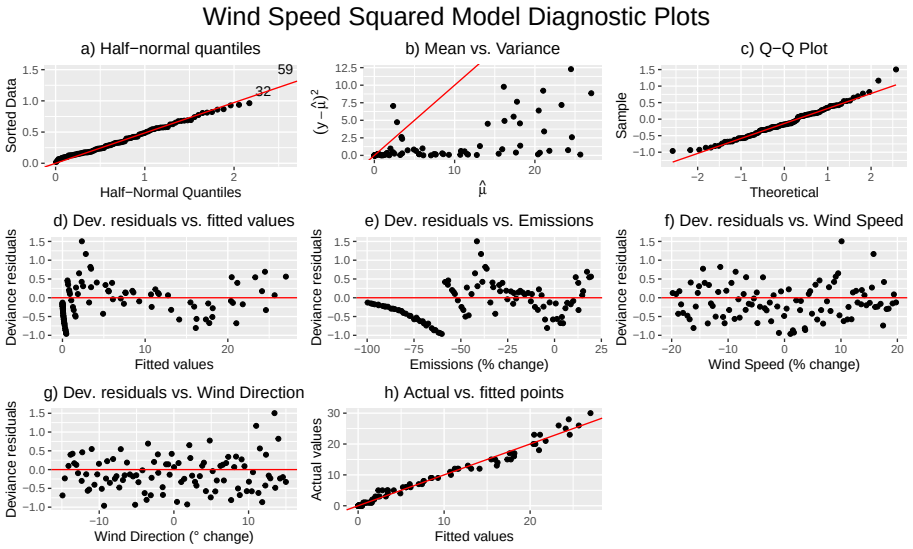

FIGURE 7.5: Diagnostic plots for the Poisson GLM for the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year (as simulated by ADMS-Urban for Central Station) with emissions (% change), emissions squared, wind speed (% change), wind speed squared and wind direction (° change) as covariates.

Table 7.1 presents the summary for the quasi-Poisson GLM where the dispersion parameter is estimated to be 0.20. The parameter estimates for all the covariates are quite small. The emissions term is positive meaning that the number of NO_2 hourly concentrations above 200 μg m^{-3} increases as emissions increase, which is logical. The emissions squared term is significant with a negative estimate. It is interesting to note that after the dispersion parameter was adjusted, the wind direction term became significant. The estimate is positive indicating that the more western prevailing the wind, the higher the NO_2 concentrations, which is consistent with the findings in Chapters 5 and 6. The best estimate for wind speed is negative, which is expected due to the fact that the higher the wind speed, the quicker the pollution disperses. Similarly to the wind direction estimate, the wind speed squared term becomes significant after the dispersion parameter adjustment.

Alternatively, a segmented (broken-stick) quasi-Poisson GLM was also fitted to the data set with the segmentation occurring at emissions $= -60\%$ as suggested by Figure 7.2. This is done in order to check if a segmented model would explain better the variability in the data than the non-segmented model. Let y_i $(i = 1, \ldots, 100)$ be the number of NO_2 hourly exceedances above 200 μg m^{-3}, then $y_i \sim \text{Po}(\mu_i)$. Then, the fitted log-link Poisson GLM for the segmented regression is:

Covariate	Estimate	95% CI	p-value
Intercept	2.77	(2.71, 2.83)	$< 2.00 \times 10^{-16}$
Emissions	0.031	(0.028, 0.033)	$< 2.00 \times 10^{-16}$
Emissions sq.	-0.00045	(-0.00053, -0.00037)	$< 2.00 \times 10^{-16}$
Wind Speed	-0.0025	(-0.0058, 0.0008)	0.1431
Wind Speed sq.	0.0005	(0.0001, 0.0008)	0.0055
Wind Direction	0.0070	(0.0030, 0.0113)	0.0007

TABLE 7.1: Summary of the quasi-Poisson GLM for the number of NO_2 hourly concentrations above 200 μg m^{-3} for the ADMS-Urban scenarios at Central Station. The corresponding estimate, 95% CI and p-value for each of the covariates (emissions (% change), emission squared, wind speed (% change), wind speed squared and wind direction (° change)) are presented.

$$\begin{aligned}\log(\mu_i) = \beta_0 + \beta_1(\mathrm{x_{EM\ i}} + 60)_+ + \beta_2(\mathrm{x_{EM\ i}} + 60)_+^2 \\ + \beta_3 \mathrm{x_{WS\ i}} + \beta_4 \mathrm{x_{WS\ i}^2} + \beta_5 \mathrm{x_{WD\ i}}\,,\end{aligned} \tag{7.11}$$

where:

- $$(\mathrm{x_{EM\ i}} + 60)_+ = \begin{cases} \mathrm{x_{EM\ i}} + 60 & \text{if } \mathrm{x_{EM\ i}} + 60 > 0\,, \\ 0 & \text{otherwise}\,; \text{and} \end{cases} \tag{7.12}$$
- $$(\mathrm{x_{EM\ i}} + 60)_+^2 = \begin{cases} (\mathrm{x_{EM\ i}} + 60)^2 & \text{if } \mathrm{x_{EM\ i}} + 60 > 0\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{7.13}$$

The diagnostic plots for the segmented quasi-Poisson GLM are presented in Figure 7.6. In comparison to the non-segmented quadratic model diagnostic plots in Figure 7.4, the half-normal quantiles plot (plot a)) and the qq-plot (c)) indicate that there are problems with the outliers and indicate the structural form of the model is not explaining the variation well. However, based on the plot e), the issues might be caused by the many zeros in the response. The proportion of deviance explained by the segmented model is 95.47% which is two percent lower than the non-segmented model. Furthermore, the mean vs. variance plot (plot b)) indicates strong underdispersion. Hence, the model was refitted as a segmented quasi-Poisson GLM.

The summary for the segmented quasi-Poisson GLM with an estimated dispersion parameter of 0.39 is presented in Table 7.2. Similarly to the non-segmented model, after estimating the dispersion parameter, the wind direction term becomes significant. The wind speed squared term is significant with p-value of 0.05 and the 95% CI is entirely positive. Overall, the parameter estimates from the two models are very similar except

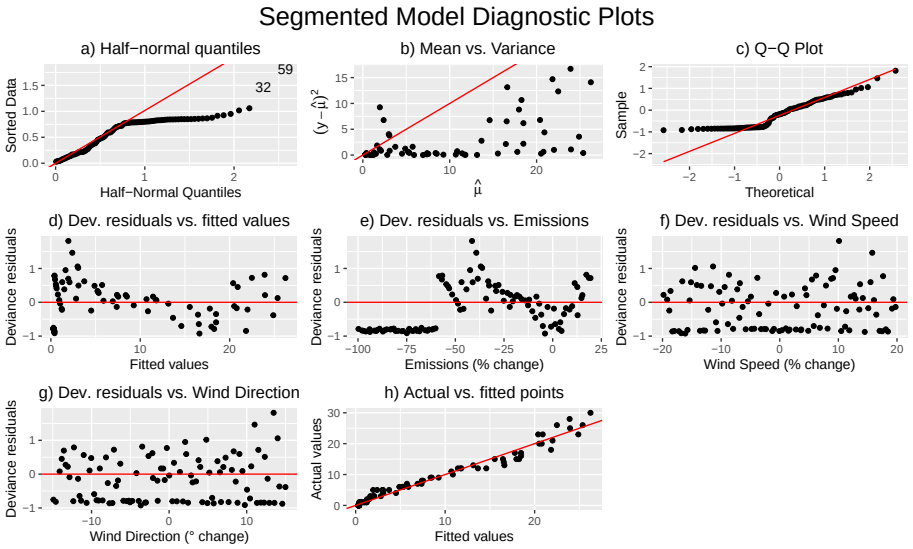

FIGURE 7.6: Diagnostic plots for the segmented quasi-Poisson GLM for the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year (as simulated by ADMS-Urban for Central Station) with emissions (% change), emissions squared, wind speed (% change), wind speed squared and wind direction (° change) as covariates.

for the intercept which the segmented model estimates as negative whereas the non-segmented model estimates as positive. The other difference is that the 95% CIs for the the segmented model are wider than the ones for the non-segmented model.

Covariate	**Estimate**	**95% CI**	**p-value**
Intercept	-1.13	(-1.40, -0.86)	1.43×10^{-12}
$(\mathbf{x}_{\text{EM}} + 60)_+$	0.10	(0.09, 0.11)	$< 2.00 \times 10^{-16}$
$(\mathbf{x}_{\text{EM}} + 60)_+^2$	-0.0006	(-0.0007, -0.0005)	$< 2.00 \times 10^{-16}$
Wind Speed	-0.003	(-0.008, 0.002)	0.21
Wind Speed sq.	0.000456	(0.000005, 0.000908)	0.05
Wind Direction	0.008	(0.003, 0.014)	0.005

TABLE 7.2: Summary of the segmented quasi-Poisson GLM for the number of NO_2 hourly concentrations above 200 μg m^{-3} for the ADMS-Urban scenarios at Central Station. The corresponding estimate, 95% CI and p-value for each of the covariates (emissions (% change), emission squared, wind speed (% change) and wind direction (° change)) are presented.

The diagnostic plots for both the quadratic quasi-Poisson GLM and the segmented quasi-Poisson GLM show issues with the fit. Negative-Binomial GLMs were also fitted but they were found to be unstable because of the many zeros in the data and therefore, these models are not presented. However, the main aim of these models is to be used for out-of-sample prediction. Therefore, the two models are compared using Deviance (see Subsection 2.2.7) and the Root Mean Squared Prediction Error (RMSPE) results from 10-fold cross-validation (CV) (see Subsection 2.2.6) in Table 7.3. Additionally, the deviance degrees of freedom (abridged to Dev. df) are also provided. The Dev. df are estimated as the difference between the total number of observations, n, and the number

of parameters p. Both the deviance and the CV indicate that the preferred model is the non-segmented one. The residual deviance for the non-segmented model is almost twice as small indicating that the non-segmented model captures better the variability in the data but both deviances are really small indicating a good fit. The 10-fold CV is also lower for the non-segmented model, which indicates that the non-segmented model is better at predicting the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year as simulated by ADMS-Urban for Central Station. However, this difference is not statistically significant as the 95% bootstrap intervals for the RMSPEs from the two models are overlapping.

Model	Dev. df	Res. deviance	RMSPE
Non-segmented	94	22.63	1.22 (0.93, 1.48)
Segmented	94	44.25	1.50 (1.14, 1.81)

TABLE 7.3: Comparing the non-segmented and segmented quasi-Poisson models for the number of NO_2 hourly concentrations above 200 μg m^{-3} in a year (as simulated by ADMS-Urban for Central Station) with emissions (% change), emissions squared, wind speed (% change), wind speed squared and wind direction (° change) as covariates based on residual deviance (the corresponding degrees of freedom) and RMSPE (and its 95% bootstrap confidence intervals) based on 10-fold CV.

Based on the diagnostic plots (Figure 7.4 for the non-segmented model and Figure 7.6 for the segmented model) and the comparative statistics in Table 7.3, the non-segmented quasi-Poisson GLM is then used to create an emulator for the number of exceedances for untested scenarios of the ADMS-Urban simulator. The 10-fold CV RMSPE score of 1.22 (Table 7.3) indicates that for a year, the quasi-Poisson non-segmented model on average will mispredict the NO_2 hourly concentrations above 200 μg m^{-3} by only 1 occurrence. Hence, the model performs well in mimicking the ADMS-Urban simulation. Therefore, in Section 7.4 the non-segmented quasi-Poisson model will be used to predict the number of NO_2 hourly concentrations above 200 μg m^{-3} alongside the standard deviations and the probability that the hourly regulation (18 occurrences above 200 μg m^{-3}) will be broken.

7.4 Emulation of the number of exceedances over 200 μg m^{-3}

Contour plots with one of the three LHC inputs set to be fixed will be presented to visualise the changes in the number of NO_2 hourly concentrations above 200 μg m^{-3} under different sets of untested inputs of the ADMS-Urban simulator. The emulated number of NO_2 hourly concentrations above 200 μg m^{-3} is obtained over a discretised

grid of the input space from the LHC. The grid to be evaluated is based on increments of 0.5 along the dimensions for each of the three inputs (emissions, wind speed and wind direction) similar to the grid set-up in Chapter 6. For a new set of inputs $\mathbf{x}_{\text{new}} = [x_{\text{new EM}}, x_{\text{new WS}}, x_{\text{new WD}}]^\top$, the output y_{new} is estimated as:

$$\mathbb{E}(y_{\text{new}}) = \exp(\mathbf{x}_{\text{new}}\widehat{\boldsymbol{\beta}})\,, \tag{7.14}$$

with standard deviation estimated as:

$$\text{SD}(y_{\text{new}}) = \sqrt{\text{Var}(y_{\text{new}})} = \sqrt{\phi\mathbb{E}(y_{\text{new}})}\,. \tag{7.15}$$

The probability of more than 18 NO_2 hourly concentrations above 200 μg m^{-3} can be estimated as $P(y_{\text{new}} > 18)$ using the Poisson probability mass function (pmf). Contour plots for the emulated number of NO_2 hourly concentrations above 200 μg m^{-3} under three variations for wind direction (as in Chapter 6) will be provided alongside the respective standard deviation for the number of hourly concentrations above 200 μg m^{-3} as well as contour plots for the probability of exceeding the regulation limit of 18 concentrations. The three variations for wind direction to be examined are:

- -15° variation from the baseline value for 2015, resulting in a more eastern prevailing wind;
- 0° variation from the baseline value for 2015; and
- 15° variation from the baseline value for 2015, resulting in a more western prevailing wind.

As the expected count is estimated by taking an exponent of the link function, an approximate 95% CI is estimated for y_{new} by:

$$\left(\exp\left(\mathbf{x}_{\text{new}}\widehat{\boldsymbol{\beta}} - 1.96\text{SD}(\mathbf{x}_{\text{new}}\widehat{\boldsymbol{\beta}})\right), \exp\left(\mathbf{x}_{\text{new}}\widehat{\boldsymbol{\beta}} + 1.96\text{SD}(\mathbf{x}_{\text{new}}\widehat{\boldsymbol{\beta}})\right)\right)\,. \tag{7.16}$$

Firstly, the predictions for the number of exceedances for each of three possible wind direction values are examined in Figure 7.7. The plots are almost identical to each other which is expected given that wind speed is not a significant predictor. In all the scenarios, the $(0, 0)$ coordinate for emissions and wind speed is in the red zone, which means that for any wind speed variation it is expected that there will be NO_2 hourly concentration over 200 μg m^{-3} close to or above the 18 exceedances regulation. For WD

= -15°, the expected number of occurrences is 15 with a 95% CI of (14, 17), for WD = 0°, there are expected to be 17 occurrences with a 95% CI of (16, 18), and for WD = +15°, there are expected to be 19 occurrences with a 95% CI of (17, 20). The plots look very similar to each other but there is a clear increase in the number of exceedances as the wind becomes more western prevailing which is in agreement with the conclusions for the NO_2 annual averages in Chapter 6. There is also a clear trend indicating that as emissions increase, so do the number of exceedances. There is a slight vertical curve as to when the predictions become red as a result of the squared wind speed term. The curve is more pronounced at the bottom of each plot which is to be expected - for lower wind speed, there is a larger number of NO_2 hourly concentration over 200 μg m^{-3}. The curvature in wind speed is caused by the fact that high wind speed is more likely to occur during the winter when there are lower temperatures which would slow down the chemical reactions between pollutants and hence, the dispersion of the pollutants.

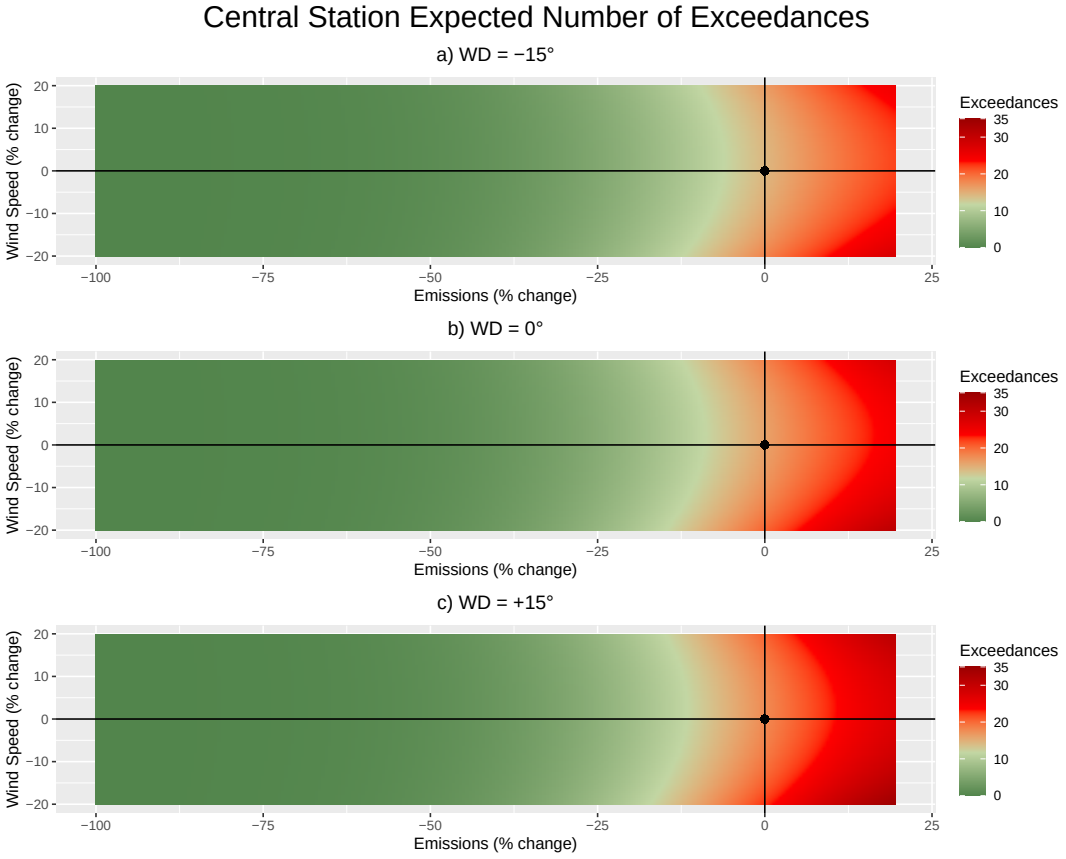

FIGURE 7.7: Contour plots for the emulated number of exceedances of the hourly NO_2 concentrations above 200 μg m^{-3} over a year when wind speed is set to -15° variation from the baseline (a)), 0° variation from the baseline (b)) and +15° from the baseline (c)) for the ADMS-Urban simulations for Central Station based on the quasi-Poisson GLM. In each plot, the black circle depicts $(0, 0)$ coordinate for emissions (% change) and wind speed (% change).

To further investigate the differences between the three examined fixed values of wind direction, the standard deviations plots are presented in Figure 7.8. The standard

deviations increase as the emissions increase to the right of the plots. As expected, the standard deviations also increase as the wind direction becomes more positive, i.e. the wind gets more western prevailing. It is interesting to note that there is curvature in the standard deviation plots similar to the one in the number of exceedances contour plots. The curvature is a result of the squared wind speed term.

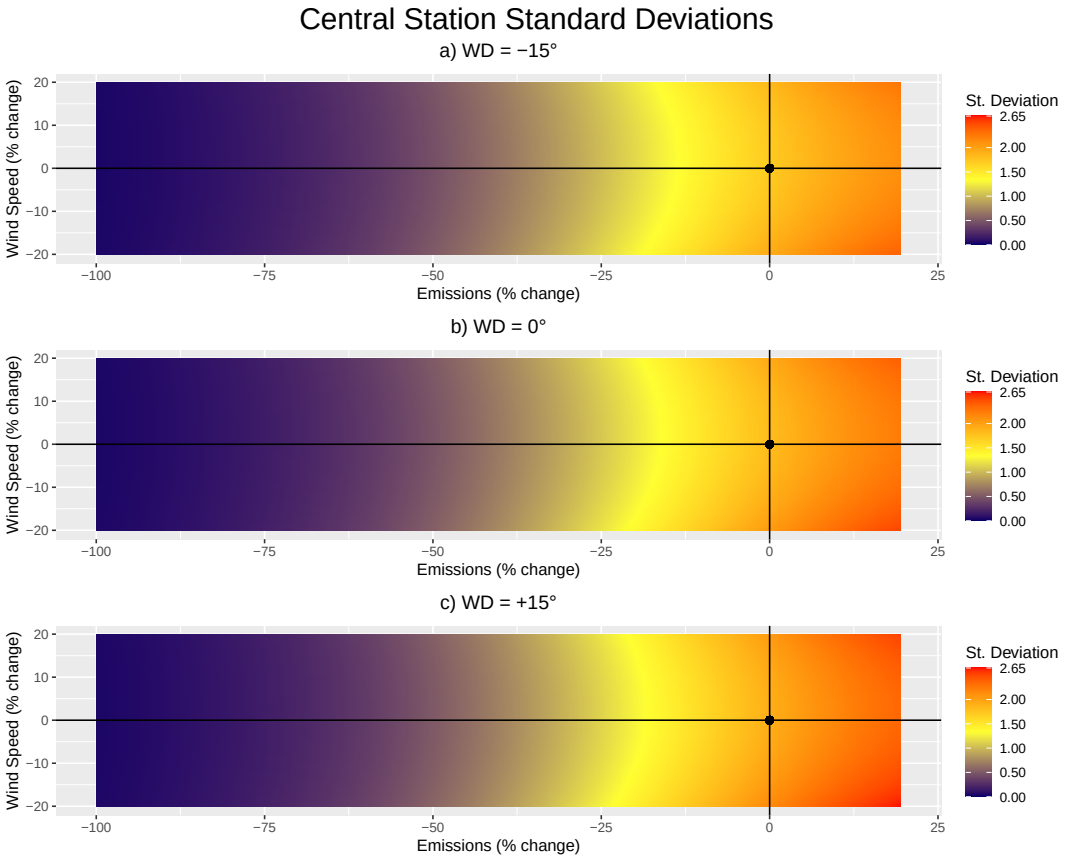

FIGURE 7.8: Contour plots for the standard error of emulated number of exceedances of the hourly NO_2 concentrations above 200 μg m^{-3} over a year when wind direction is set to -15° variation from the baseline (a)), 0° variation from the baseline (b)) and +15° from the baseline (c)) for the ADMS-Urban simulations for Central Station based on the quasi-Poisson GLM. In each plot, the black circle depicts $(0, 0)$ coordinate for emissions (% change) and wind speed (% change).

Lastly, the contour plots for the probability of breaching the 18 occurrences over 200 μg m^{-3} regulation over a year for the three examined fixed values of wind direction are presented in Figure 7.9. As with the previous plots in Figures 7.7 and 7.8, there is clear curvature due to the wind speed squared term. As with the contour plots for the expected numbers of exceedances in Figure 7.7, in Figure 7.9 there is a difference between the three plots. The size of the dark red area increases as the wind direction degrees change from east to west. Examining the $(0, 0)$ for the three plots, the probabilities of exceedances are 13.76%, 25.70% and 42.25%, respectively. It is interesting to compare the result from plot b) as that reflects baseline value for what has occurred in 2015. Although the emulator for ADMS-Urban estimates 17 occurrences (with a 95% CI of

(16, 18)), the probability for that happening is quite low which is logical as there were only four hourly breaches observed in 2015.

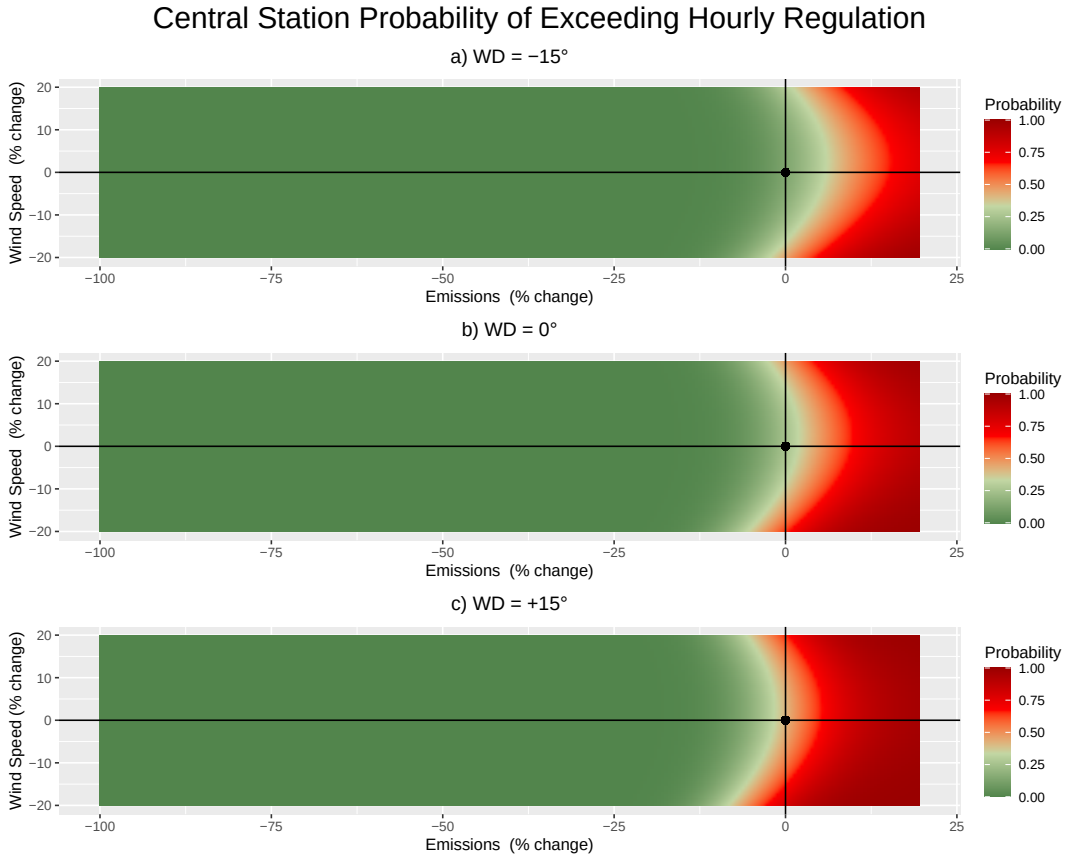

FIGURE 7.9: Contour plots for the probabilities for exceeding the 18 occurrences over 200 μg m^{-3} regulation over a year when wind direction is set to -15° variation from the baseline (a)), 0° variation from the baseline (b)) and +15° from the baseline (c)) for the ADMS-Urban simulations for Central Station based on the quasi-Poisson GLM. In each plot, the black circle depicts $(0, 0)$ coordinate for emissions (% change) and wind speed (% change).

7.5 Conclusion

In this chapter, it was aimed to address the NO_2 hourly concentrations regulation from the European Directive 2008/50/EC [76], where more than 18 occurrences over 200 μg m^{-3} constitute a breach. In order to explore how changes in emissions, wind speed and wind direction affect the hourly regulation, the one hundred ADMS-Urban simulations were used. From Figure 7.1, it became clear that Central Station is the only monitoring station in Glasgow, where the hourly regulation was breached. Therefore, the rest of the chapter focuses only on this specific station.

In Figure 7.2, it was observed that there will be at least one exceedance if the percentage change in emissions is higher than -60% regardless of the wind speed and wind direction

values. This exploratory plot suggested that either a Poisson GLM or a segmented Poisson GLM can be used to model the number of occurrences over 200 μg m^{-3}. Both models struggled with underdispersion, which required refitting them as quasi-Poissons. The final quasi-Poisson GLM was fitted with emissions (% change), emission squared, wind speed (% change), wind speed squared and wind direction (° change) as covariates, whereas the segmented quasi-Poisson GLM had an emissions term (% change), for which $x_{EM\ i} + 60 = 1$ if $x_{EM\ i} + 60 > 0$ or 0 otherwise, similar emissions squared term, wind speed (% change), wind speed squared and wind direction (° change) as covariates. The diagnostic plots for both models (Figures 7.5 and 7.6) showed issues with the fit. However, the main aim of the models was prediction and they were compared using 10-fold CV RMSPE and the non-segmented quasi-Poisson model was chosen as the preferred model for prediction as on average it will mispredict the NO_2 hourly concentrations above 200 μg m^{-3} by only 1 occurrence and performs well in mimicking the ADMS-Urban simulation results.

The quasi-Poisson GLM was then used to create an emulator for the number of exceedances of the NO_2 hourly concentrations above 200 μg m^{-3}. For scenarios with emissions percentage change smaller than or equal to -60% of variation, there are zero expected exceedances. However, in cases where the change in emissions percentage are larger than -60% of the baseline, all three inputs (emissions, wind speed and wind direction) have a significant effect. The emulator was then used to explore the expected number of exceedances for three set values of wind direction: -15° from the baseline, 0° from the baseline and +15° from the baseline. It was found that the more western prevailing the wind, the more hourly NO_2 exceedances above 200 μg m^{-3} were estimated. It is interesting to note that the emulator results are estimating unrealistically high number of occurrences of hourly concentrations above 200 μg m^{-3} (Figure 7.9) for the baseline emissions, wind speed and wind direction of 2015, although the probabilities for the occurrence of a high number of occurrences of hourly concentrations above 200 μg m^{-3} are lower than 50%. This suggests that ADMS-Urban tends to overpredict the number of NO_2 hourly concentrations above 200 μg m^{-3}.

Chapter 8

Modelling and emulation of the NO_2 hourly concentrations in Glasgow

In Chapter 7, the numbers of NO_2 hourly concentrations over 200 μg m^{-3} were modelled as the Scottish regulation for hourly NO_2 concentrations allows for 18 occurrences over 200 μg m^{-3} a year as seen in Table 1.1. However, it is also of interest to be able to examine the time series for the NO_2 hourly concentrations from ADMS-Urban under varying emissions, wind speed and wind direction and assess how well the time series of NO_2 hourly concentrations can be emulated for new sets of inputs in terms of emissions, wind speed and wind direction. This will allow SEPA and other governmental agencies to be able to identify the specific conditions, which lead to periods of hourly concentrations of 200 μg m^{-3} or higher, and thus, enable them to focus their efforts on preventing the occurrence of such high concentrations. As Central Station in Glasgow is the only location where there are concentrations over 200 μg m^{-3}, the ADMS-Urban simulation scenarios for this station are modelled. The aim of this chapter is to create an emulator for the NO_2 hourly concentration time series which is more computationally efficient than producing the time series via ADMS-Urban. The chapter is organised as follows: Section 8.1 provides an exploratory analysis of the time series at Central Station across different simulation scenarios in order to establish a mean trend. In Section 8.2, a computationally efficient spatio-temporal emulator is introduced for producing the yearly time series for the ADMS-Urban hourly NO_2 concentrations under different conditions. The model introduced in Section 8.2 is then applied to a subset (of the number of time steps) of the time series simulation scenarios at Central Station in Section 8.3 due to the substantial size of the complete data set, and a more computationally efficient approach is implemented. Following that, emulation of the subsetted time series is performed and

compared to the observed hourly NO_2 concentrations in 2015. Section 8.4 provides a concluding discussion.

8.1 Exploratory analysis of ADMS-Urban time series

This section focuses on the exploratory modelling of one of the one hundred simulations from ADMS-Urban in order to establish the overall mean trend required for modelling the time series. Simulation scenario 16 was chosen as in terms of Euclidean distance (for the estimation of Euclidean distance, see [14]) it is the scenario closest to the $(0, 0, 0)$ coordinate point, which signifies the baseline conditions recorded in 2015 (i.e. no changes in emissions, wind speed and wind direction). Simulation 16 has emissions that are higher by 3.20% than the observed emissions for every hour, wind speed that is lower by 4.80% from the hourly wind speed in 2015, and wind direction that is changed by adding 1.49° to the hourly wind direction in 2015 (see Subsection 4.2.3). The simulation scenario has 15 hourly concentrations above 200 μg m^{-3}, hence although there are observations above the 200 μg m^{-3} hourly limit, the hourly regulation has not been breached. The exploratory modelling is based on the exploratory analysis of year long time series for emissions, temperature, wind speed and wind direction with the monitored NO_2 hourly concentrations presented in Appendix B. The log NO_2 concentrations will be used for modelling because the diurnal cycle for the emissions is based on a multiplication factor for each of the twenty-four hours in a day (see Appendix B), hence the log scale reduces this to an additive model. Other simulation scenarios have similar trends as each simulation follows the same overall trend but differs in the absolute value of the NO_2 concentrations, however the exploratory analysis of these additional simulations is not shown for brevity.

8.1.1 Data visualisation for simulation scenario 16

Firstly, the relationships between the simulated yearly time series for Scenario 16 for the log hourly NO_2 concentrations at Central Station and the corresponding meteorological conditions are explored. A time series plot comparing the time series of the hourly temperatures and the log hourly NO_2 concentrations in order to explore the seasonal trends is presented in Figure 8.1. From the time series plot, it can be observed that in periods of low temperatures, the log NO_2 concentration spikes. This trend is most visible when focusing on the temperature inversion in the second half of January, when the temperatures are at their lowest and the log NO_2 concentrations are highest with some of the concentrations going over the 200 μg m^{-3} regulation. This is logical given that higher temperatures act as a catalyst in chemical reactions, therefore resulting in

faster reactions between NO_2 and the other pollutants (for instance, the forming of ozone).

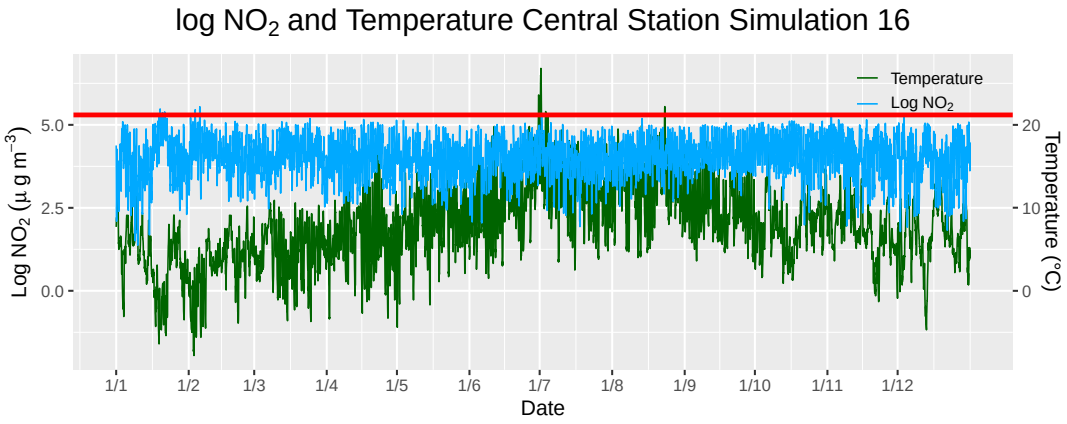

FIGURE 8.1: The time series for log NO_2 concentrations (μg m^{-3}) from Simulation 16 of the ADMS-Urban simulator (in light blue) imposed on the time series plot for the hourly temperatures (°C) (in dark green) for 2015 for Central Station. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation.

Additionally, a scatterplot for the hourly log NO_2 concentrations and hourly temperatures is presented in Figure 8.2. From the plot, there is a weak negative linear trend. The Pearson's correlation coefficient (see Subsection 2.1.2) is -0.10 with a 95% CI (-0.12, -0.07), indicating that although the relationship is weak, it is significant. This confirms the conclusions from Figure 8.1 that there is a negative relationship between the log NO_2 concentrations and the temperature.

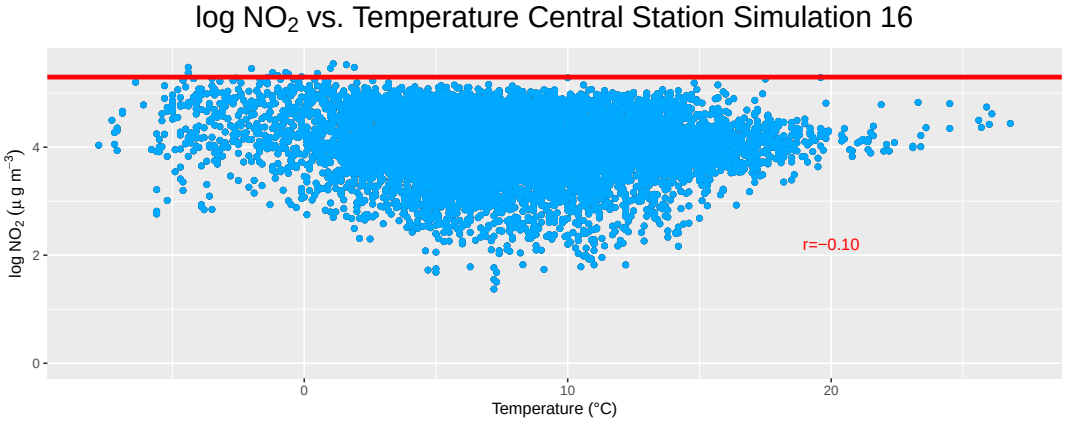

FIGURE 8.2: Scatterplot for the hourly temperatures (°C) against the log NO_2 concentrations (μg m^{-3}) from Simulation 16 of the ADMS-Urban simulator. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation. The correlation between temperature and log NO_2 concentrations is also provided.

Next, the seasonal relationship between simulated yearly time series for the log hourly NO_2 concentrations at Central Station from Simulation 16 with the hourly time series for wind speed in 2015 (adjusted for the LHC design by lowering the 2015 time series for wind speed by 4.80%) is investigated in Figure 8.3. From the plot, it is clear that

in moments when wind speed is 0 (for instance, the second half of January, beginning of February), there are spikes in the log NO_2 concentrations as the lower wind speed results in slower dispersion of NO_2.

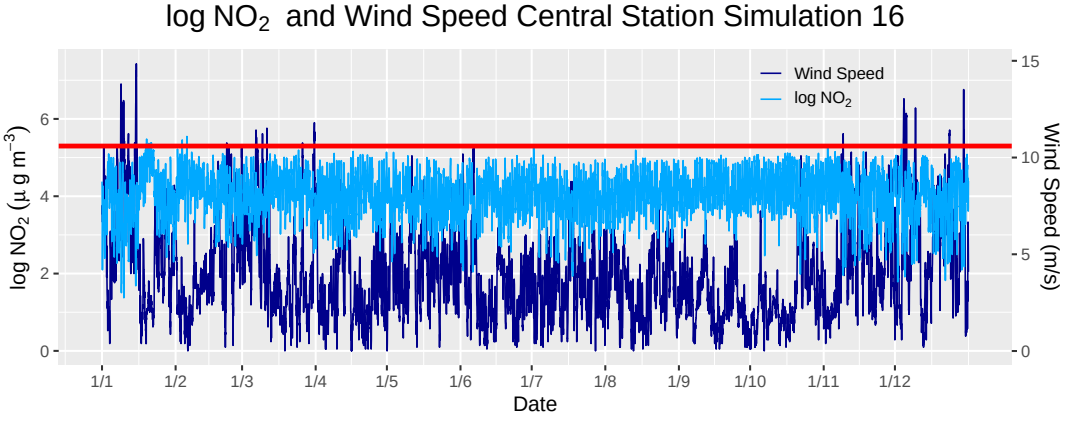

FIGURE 8.3: The time series for log NO_2 concentrations (μg m^{-3}) from Simulation 16 of the ADMS-Urban simulator (in light blue) imposed on the time series plot for the hourly wind speed (m/s) (in dark blue) for 2015 for Central Station. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation.

A scatterplot for the hourly log NO_2 concentrations and the hourly wind speed (lowered by 4.80% from the 2015 recordings) is presented in Figure 8.4. The Pearson's correlation coefficient is -0.44 with a 95% CI (-0.46, -0.42) suggesting a moderately strong negative correlation between the two variables. This confirms the conclusions from Figure 8.3 that higher wind speeds result in faster dispersion of the pollutants, and hence, lower log NO_2 hourly concentration.

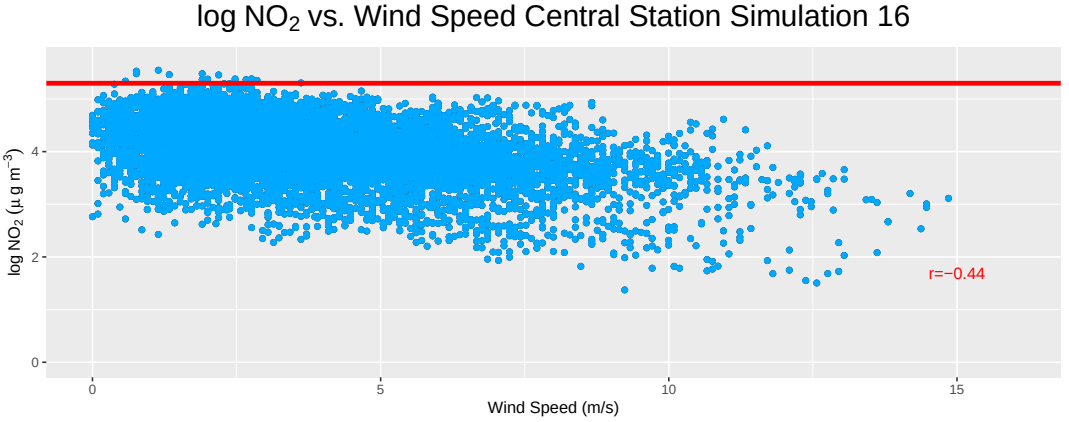

FIGURE 8.4: Scatterplot for the hourly wind speed (m/s) against the log NO_2 concentrations (μg m^{-3}) from Simulation 16 of the ADMS-Urban simulator. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation. The correlation between wind speed and log NO_2 concentrations is also provided.

Lastly, the relationship between simulated yearly time series for the log hourly NO_2 concentrations at Central Station from Simulation 16 with the hourly time series for wind direction in 2015 (adjusted for the LHC design by adding -1.47° to the 2015 time

series for wind direction) is investigated in Figure 8.5 using a pollution rose [34]. A wind rose is a type of plot, which shows how wind speed and wind direction conditions vary by year. The data are summarised by direction and by different wind speed categories represented by different width paddles. The plots show the proportion (in percentage) of time that the wind is from a certain angle and wind speed range. If the wind speed information is replaced by pollutant concentrations, a wind rose can be extended into a pollution rose. Hence, the pollutant concentrations by wind direction and the percentage time the concentration is in a particular range is visualised. Pollution roses can be created using the **openair** package [35] in R [157]. From the pollution rose in Figure 8.5, it is clear that the concentrations below 4.4 on the log scale (equivalent to 80 μg m^{-3}) are predominant for the western winds, whereas for all other directions the different segments appear relatively even. The orange segments are disproportionally larger for the eastern prevailing wind in comparison to the other directions. Hence, as the wind becomes more eastern prevailing, the hourly log NO_2 concentrations increase. Therefore, wind direction should be used as a circular variable when modelling the NO_2 hourly concentrations.

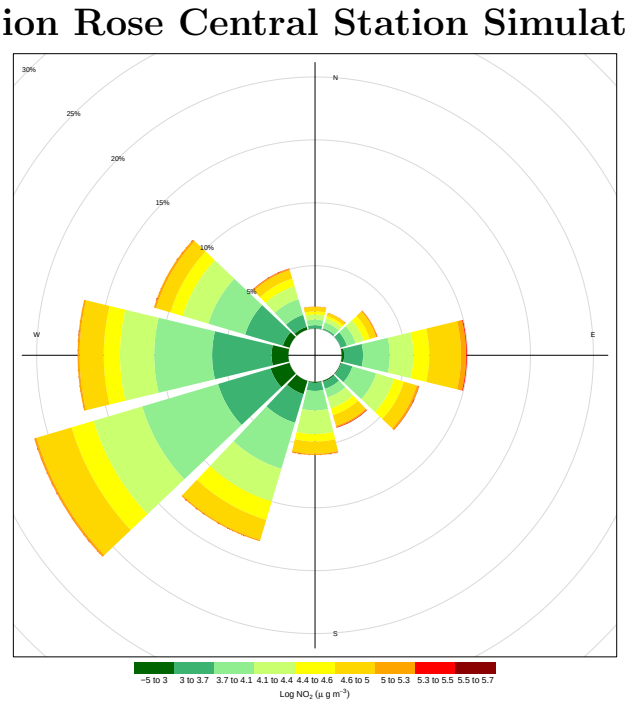

FIGURE 8.5: Pollution rose for the monitoring station at Central Station for Simulation 16 of the ADMS-Urban simulator. The corresponding log NO_2 concentration (μg m^{-3}) in 2015 are proportionally ordered to the modelled wind direction angle (°) at which the concentrations are recorded.

Besides temperature, wind speed and wind direction terms, an interaction between temperature and wind speed will be included in the modelling. This is done to account for the difference between summer and winter. In the summer, the hourly temperatures are

expected to be above 20°C and even though the wind speed will be close to 0 m/s, there will be lower NO_2 concentrations. On the other hand, in the winter, for hours with high wind speed (above 10 m/s), even though the temperatures are low (below 5°C), it is expected that the NO_2 concentrations will be lower.

In order to examine the log hourly NO_2 concentrations for each of the emissions (measured in g m^{-2} h) in the 24-hour pattern (emissions in ADMS-Urban follow a 24-hour pattern repeated for each of the days of the year as discussed in Subsection 4.1.3), boxplots are presented in Figures 8.6. There is only one boxplot for the emissions at 04:00 and 05:00 as the emissions are identical for the two hours. The emissions are ordered from smallest to largest going left to right to check for a factor effect in the covariate. There is a moderate linear trend suggesting that the higher the emissions, the higher log hourly NO_2 concentrations. The Pearson's correlation between the emissions and the log hourly NO_2 concentrations is 0.44 (with a 95% CI (0.42, 0.46)) confirming a moderate correlation between the two variables. Therefore, emissions will be included in the model.

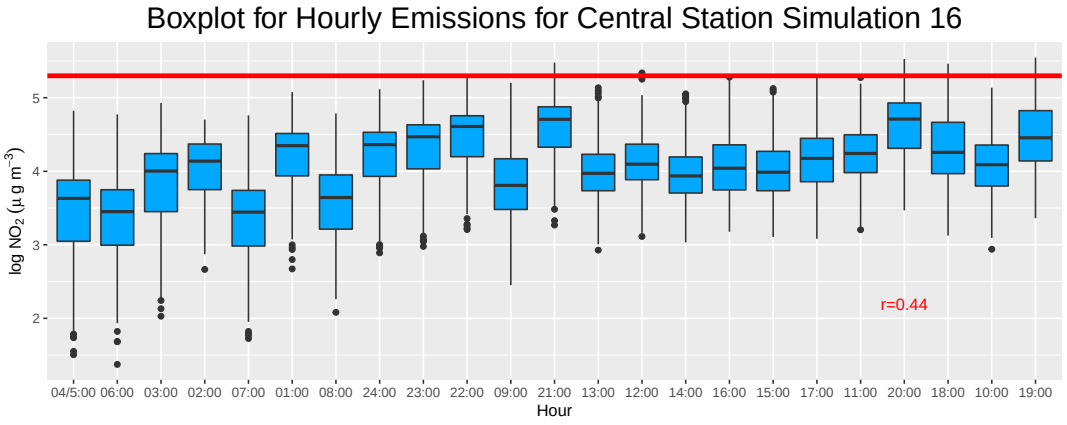

FIGURE 8.6: Boxplots for the ordered (by emissions' size ordered from smallest to largest) log NO_2 concentrations (μg m^{-3}) over a 24-hour cycle at Central Station for Simulation 16 of the ADMS-Urban simulator. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation. The correlation between emissions (g m^{-2} h) and log NO_2 concentrations is also provided.

Additionally, it is of interest to account for the temporal trends. The hourly variation in a day is examined in Figure 8.7, which presents the boxplots of the variation in the hourly log NO_2 concentrations in a day. The boxplots clearly indicate that there is a variation in concentrations between the hours during the day. The trend in the boxplots is similar to the magenta line, which signifies the daily emissions cycle for 24-hours, although the two trends are not identical. Hence, a factor term for the 24-hour cycle will be needed for modelling in addition to the emissions covariate. It also has to be noted that the boxes for 01:00 and 24:00 appear almost identical which suggests there

might be an identifiability issue between those two hours, regardless of the fact that there are different emissions for those two hours.

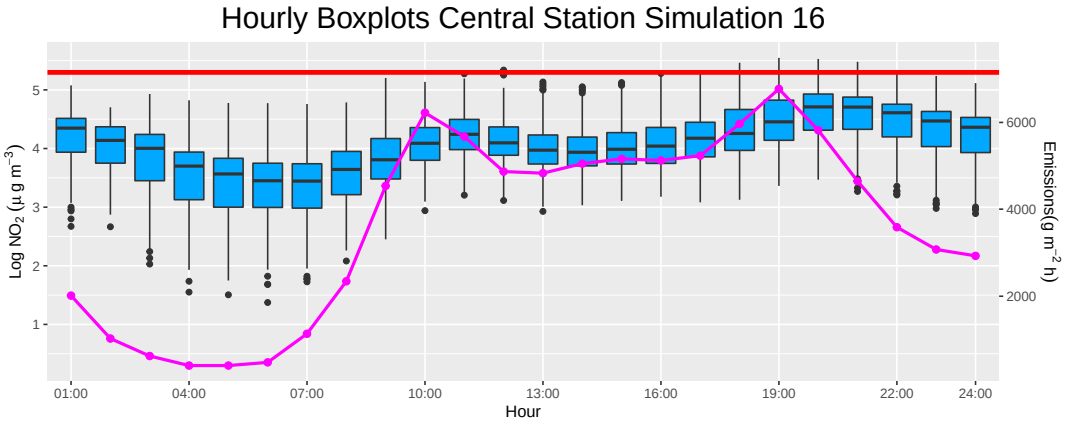

FIGURE 8.7: Boxplot for the log NO_2 concentrations (μg m^{-3}) over a 24-hour cycle at Central Station for Simulation 16 of the ADMS-Urban simulator. The emissions (g m^{-2} h) for each hour are superimposed in magenta. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation.

Next, the days in the week variation in the log NO_2 concentrations is examined in Figure 8.8. The boxes for the different days of the week are very similar to each other but not identical. This is a result of the fact that the simulations were produced without adjusting for the change of people's activities between workdays and weekends. Therefore, a day of the week variable is not included in the model due to the lack of day-to-day variation as illustrated in Figure 8.8. Furthermore, the lack of day-to-day variation is expected to result in over-estimation of the NO_2 hourly concentrations when comparing the simulated (and emulated) values to the actual observations.

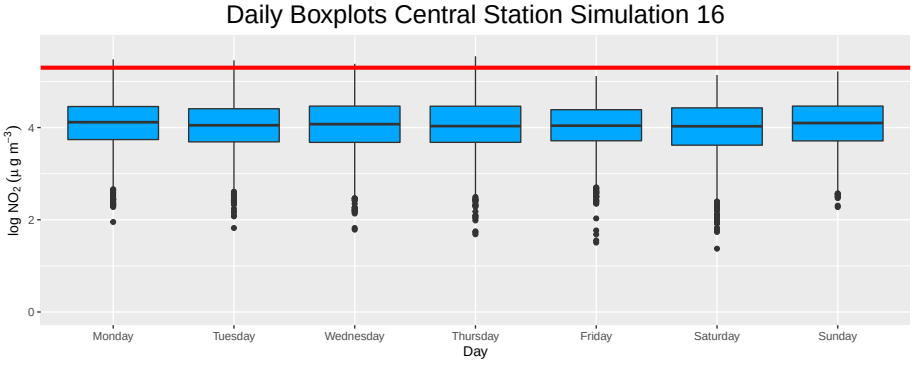

FIGURE 8.8: Boxplot for the log NO_2 concentrations (μg m^{-3}) over a week cycle for Simulation 16 of ADMS-Urban at Central Station. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation.

It is also of interest to also explore the daily and weekly variation in a year. This is visualised by plotting the daily and weekly averages (means) of the log hourly NO_2 concentrations in Figure 8.9. As expected there is a lot of variability in the daily means

which would suggest using a spline to smooth the curve. There is a peak around the beginning of the year (February), a trough in the summer (June) and another peak around the autumn (October). The weekly means further highlight the differences between the seasons. As the shape of the weekly variation has less fluctuations in comparison to the daily means suggesting that a weekly means b-spline (see Subsection 2.2.2) would be sufficient in modelling the time series movement.

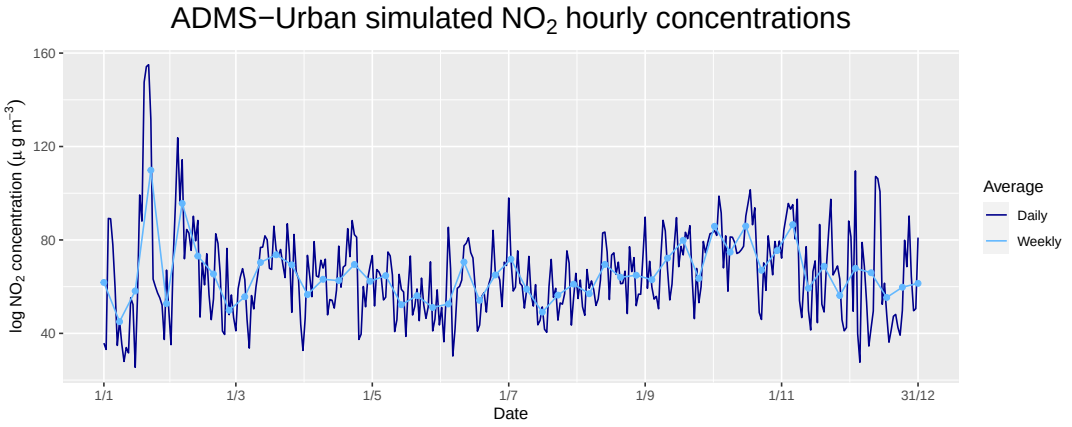

FIGURE 8.9: Daily and weekly means for log NO_2 concentrations (μg m^{-3}) for Simulation 16 of ADMS-Urban at Central Station.

8.1.2 Modelling of simulation scenario 16

Exploratory modelling is applied to Simulation 16 from the ADMS-Urban simulator for Central Station. This is done to help establish the main trend in the time series. The baseline model is fitted as a linear regression model (see Subsection 2.2.1). This baseline model includes only the meteorological conditions to explore how much of the variation in the data can be explained by them alone. Hence, for hour $t = 1, \ldots, 8760$:

$$\begin{aligned} \log\left(\mathrm{NO_2}\right)_t = \beta_0 + \beta_1\left(\mathrm{Temp}_t\right) + \beta_2\left(\mathrm{WS}_t\right) + \beta_3 \sin\left(\frac{2\pi \mathrm{WD}_t}{360}\right) \\ + \beta_4 \cos\left(\frac{2\pi \mathrm{WD}_t}{360}\right) + \beta_5\left(\mathrm{Temp}_t * \mathrm{WS}_t\right) + \epsilon_t\,. \end{aligned} \tag{8.1}$$

The model has $\mathrm{R}^2_{\mathrm{adj.}} = 20.89\%$ which indicates that the meteorological conditions explain about 21% of the variability in the log hourly NO_2 concentrations. The diagnostic plots in Figure 8.10 a) through d) indicate that the linear model assumptions hold and the model is a good fit to the data. However, the ACF and PACF plots of the residuals (plots e) and f)) indicate strong autocorrelation in the residuals with a clear seasonal pattern.

To address the seasonal pattern in the ACF and PACF plots in Figure 8.10, different variables will be added. To avoid repetition by presenting each of the models in full,

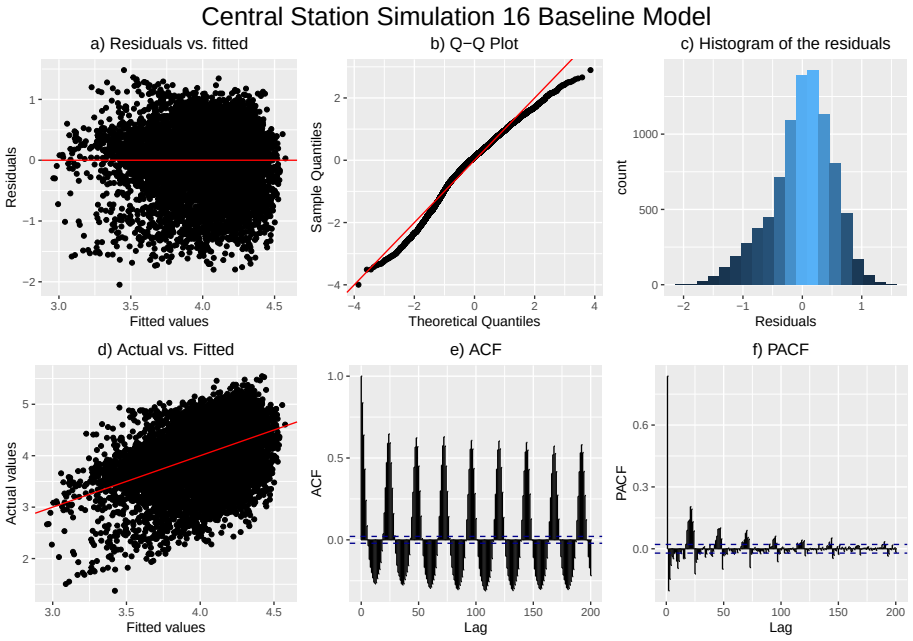

FIGURE 8.10: Diagnostic plots for the model for the log hourly NO_2 measurements (μg m^{-3}) for Simulation 16 of ADMS-Urban at Central Station with temperature (°C) , wind speed (m/s), an interaction between temperature and wind speed and wind direction (°) terms.

the models will be compared to each other using $R^2_{\text{adj.}}$ and AIC in Table 8.1. Each subsequent model given in Table 8.1 has an additional covariate (as specified in the name of the model) over the one preceding it unless otherwise specified. Additionally, the ACF and PACF plots for the residuals of the models are presented in Figure 8.11. Firstly, to the baseline model was added the emissions (referred to as Em) covariate. There is a clear improvement from the baseline model after the emissions variable is included in terms of the diagnostic criteria in Table 8.1. The ACF and PACF plots for the two models (in the first two rows) in Figure 8.11 also show an improvement but there is still temporal trend present which could be further reduced by adding more covariates to the model.

Therefore, a factor for the hour of the day is added to the model in order to remove the temporal trend in the ACF and PACF plots. However, the hours 01:00 and 24:00 are combined together (due to their similarity) as a condition to help estimate all factors. The diagnostics criteria in Table 8.1 indicate an improvement. The ACF and PACF plots (plots e) and f), respectively) in Figure 8.11 also show an improvement. The temporal trend on the ACF plot has been reduced and the PACF plot shows only one significant correlation at lag 1. Instead of using factors for the hours of the day, a circular variable for the hours of the day was added to the model but the temporal trend reduction was smaller than when using a factor variable. For brevity, this comparison is omitted.

Lastly, the week numbers (as continuous numbers) are also included as a b-spline in the model to reduce the variability in the ACF plots. Weekly variation is important not only in terms of weather but also because it accounts for variation in people's behaviour such as school holidays. The b-spline basis has degree 3. Initially, 10 knots were fitted and some coefficients were removed until all coefficients for the spline are significant. Therefore, for the final model the spline has 3 coefficients. From Table 8.1, the model including the week numbers as a b-spline has the largest $R^2_{\text{adj.}}$ value in comparison with the other models as well as the smallest AIC value. The ACF and PACF plots (plots g) and h), respectively) in Figure 8.11 show further reduced variability in the ACF plot and the PACF plot has only lag 1 as significant (same as plot f)). Although the ACF plots show some variability, it has to be noted that the plots are produced for a large dataset and hence, the error bars are very small. Furthermore, the reduction in the variability as more covariates were added to the model is clear. Therefore, it appears that only a temporal random effects structure is left and this model is chosen to be the final model.

Model	**df**	**AIC**	$\mathbf{R^2_{adj.}}$
Baseline	7	13 162.85	20.89%
Em	8	9 001.32	50.81%
24-Hour factor	30	5 230.91	68.10%
Week Number Spline	34	5 078.15	68.66%

TABLE 8.1: Comparing models with different temporal covariates to account for the autocorrelation in the residuals when modelling the log hourly NO_2 measurements (μg m^{-3}) for Simulation 16 of ADMS-Urban at Central Station.

The AIC and $R^2_{\text{adj.}}$ values from Table 8.1 and the ACF and PACF plots from Figure 8.11 indicate that the most appropriate model for the log NO_2 hourly concentrations from Simulation 16 of ADMS-Urban at Central Station contains the meteorological data (hourly temperatures, hourly wind speed and hourly wind direction for the year as well as an interaction between temperatures and wind speed) with a continuous covariate for emissions, a factor covariate for hour of the day, and a b-spline for the week of the year. Hence, for hour $t = 1, \ldots, 8760$:

$$\begin{aligned}\log(\mathrm{NO}_2) = \beta_0 &+ \beta_1(\mathrm{Temp}_t) + \beta_2(\mathrm{WS}_t) + \beta_3 \sin\left(\frac{2\pi \mathrm{WD}_t}{360}\right) \\ &+ \beta_4 \cos\left(\frac{2\pi \mathrm{WD}_t}{360}\right) + \beta_5(\mathrm{Temp}_t * \mathrm{WS}_t) \\ &+ \beta_6(\mathrm{EM}_t) + \sum_{j=2}^{23} \beta_j^H \mathbb{I}[t \text{ is the } j^{\text{th}} \text{ hour of the day}] \\ &+ f(\text{Week Numbers}_t) + \epsilon_t\,,\end{aligned} \tag{8.2}$$

where:

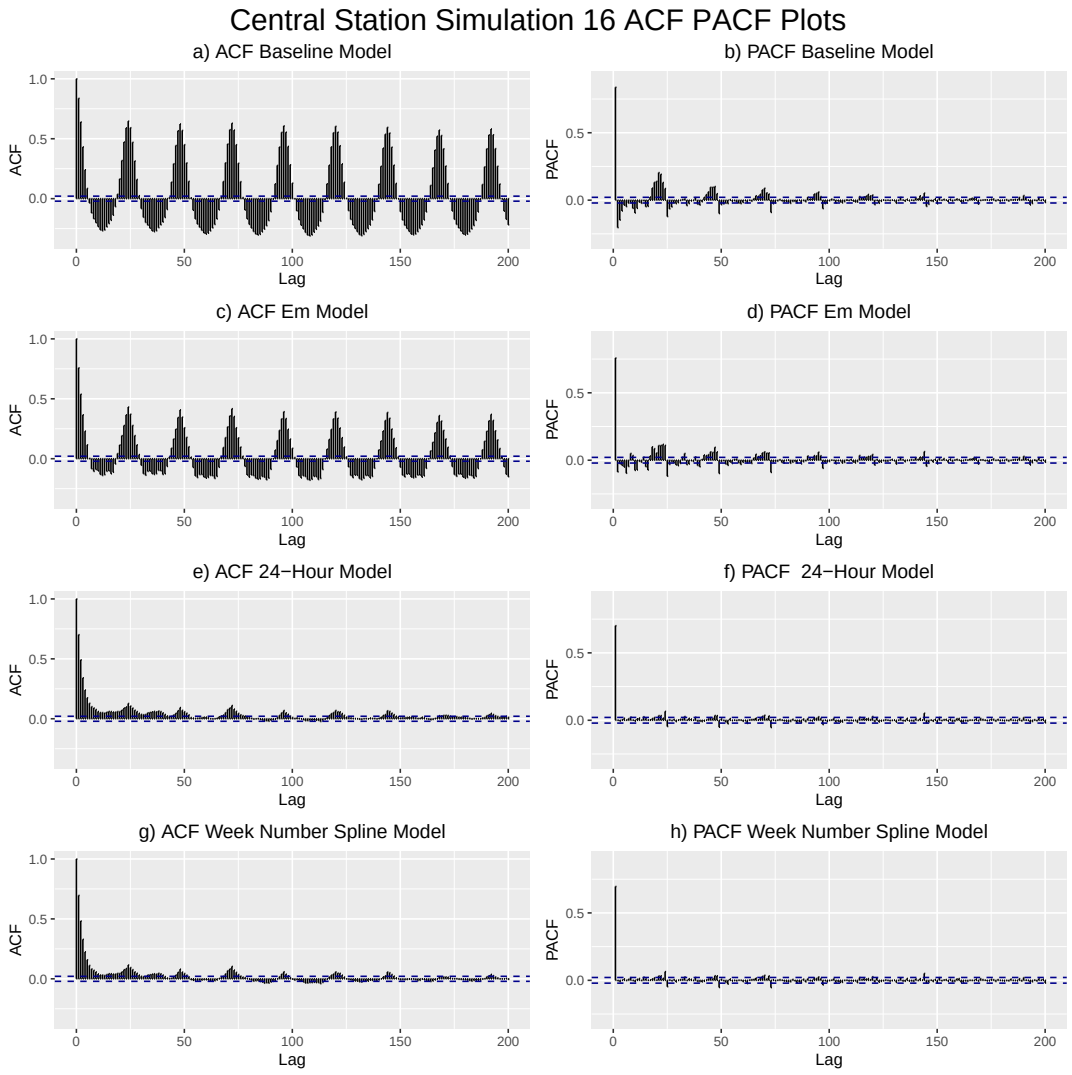

FIGURE 8.11: ACF and PACF plots for the residuals for the different models for the log hourly NO_2 measurements (μg m^{-3}) for Simulation 16 of ADMS-Urban at Central Station.

- $\sum_{j=2}^{23} \beta_j^H \mathbb{I}[t \text{ is the } j^{\text{th}} \text{ hour of the day}]$ is an indicator variable for the factor for the hour j of the day for the t^{th} hour in the year. The baseline for the factor is set for 01:00 and 24:00 (due to an identifiability issue) and hence, the parameter for those two hours is 0. The parameter β_j^H is the difference between the other observed hours ($j =$ 02:00, ..., 23:00) and the baseline hour; and

- $f(\text{Week Numbers}_t) = \beta_8 B_{1,d}(\text{Week Numbers}_t) + \beta_9 B_{2,d}(\text{Week Numbers}_t) + \beta_{10} B_{3,d}(\text{Week Numbers}_t) + \beta_{11} B_{4,d}(\text{Week Numbers}_t)$ is the b-spline with $d = 3$ degrees of freedom and 3 basis functions for the week of the year variable.

The diagnostic plots for this model are presented in Figure 8.12. The residuals vs. fitted values points on plot a) are randomly scattered around zero and there is no fanning out. Hence, the residuals appear normally distributed with mean zero and constant variance.

In plot b), the residuals on the qq-plot are lying on the normality line indicating the residuals are normally distributed. This is further confirmed by the histogram of the residuals in plot c) which is symmetric and centred around zero. The actual vs. fitted points in plot d) are lying around the equivalence line indicating the model performs well in predicting the observed data. However, the ACF in plot e) and the PACF in plot f) indicate that there is strong autocorrelation in the residuals. In order to account for the autocorrelation, the model was refitted as an AR(1) model (using the *gls* function in the **nlme** [150] package in R) as the PACF plot has only lag 1 significant, whereas the significant lags on the ACF plot are slowly decaying.

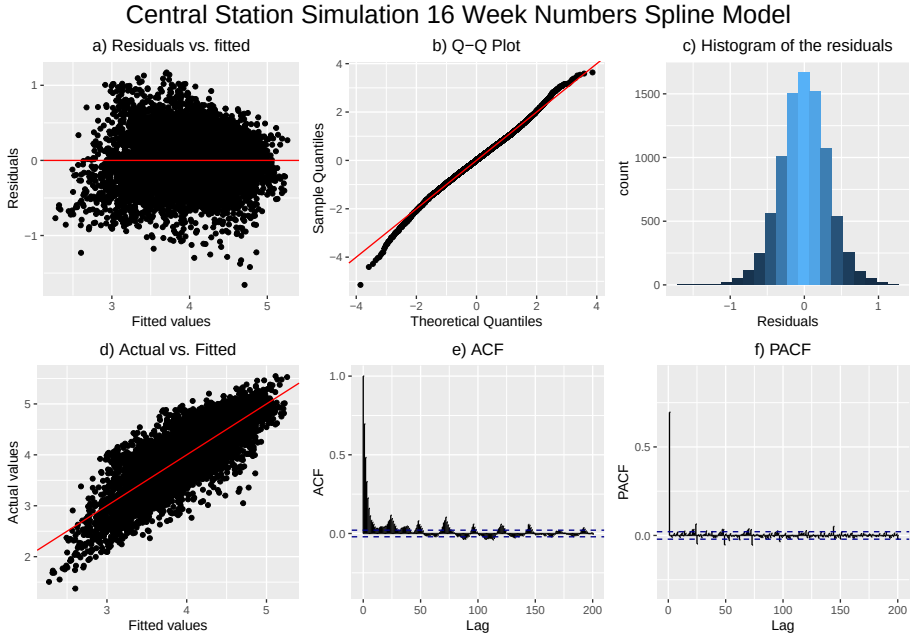

FIGURE 8.12: Diagnostic plots for the model for the log hourly NO_2 measurements (μg m^{-3}) for Simulation 16 of ADMS-Urban at Central Station with temperature (°C) , wind speed (m/s), an interaction between temperature and wind speed, wind direction (°), emissions (g m^{-2} h), factor for the hour of the day and b-spline for the week number as covariates.

The ACF and PACF plots for the AR(1) model are presented in Figure 8.13. Lag 0 is not displayed on the ACF plots to allow the examination of subsequent lags in more detail. The plots show that the autocorrelation in the residuals has been accounted for with the AR(1) structure as almost all the bars in the plots are within the error bands. However, it has to be noted that for every 24 lags, there is still a significant lag which is caused due to the condition of setting the factor for the 24^{th} hour of the day to be equal to the 1^{st} hour of the day. The AR(1) parameter is estimated to be 0.75 (with standard error of 0.01) indicating a strong correlation between two consecutive hours.

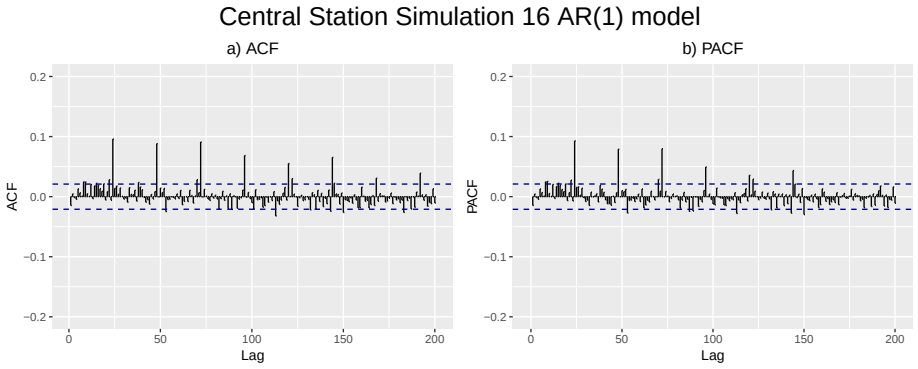

FIGURE 8.13: ACF and PACF plots for the residuals of the log hourly NO_2 measurements (μg m^{-3}) for Simulation 16 of ADMS-Urban at Central Station with temperature (°C), wind speed (m/s), an interaction between temperature and wind speed, wind direction (°) , emissions (g m^{-2} h), factor for the hour of the day and b-spline for the week number as covariates and AR(1) model to account for the autocorrelation in the residuals.

8.1.3 Findings

In this section, simulation scenario 16 was chosen (due to its proximity to the $(0, 0, 0)$ baseline based on Euclidean distance) in order to examine the effects of different variables on the hourly NO_2 concentrations. However, as the emissions used for the ADMS-Urban simulation come from a multiplicative factor, log NO_2 concentrations were used for modelling. Based on the exploratory modelling of Simulation 16, several key observations were made. Firstly, the meteorological covariates (temperature, wind speed, an interaction between temperature and wind speed, and wind direction) explain about 20% of the variation in the data. Secondly, the factor for hour of the day helps smooth out the seasonality trend in the ACF and PACF plots in the residuals. However, due to an identifiability issue, the factor for 01:00 and 24:00 is the same. Thirdly, b-spline is used for the Week Number covariate which is also used to smooth the seasonality trend in the ACF and PACF plots of the residuals. Furthermore, an AR(1) model appears sufficient to explain the residual autocorrelation. Lastly, due to conditioning the factor for 01:00 and 24:00 to be equal, in the final ACF and PACF plots, there are still significant lags every 24$^{\text{th}}$ hour. Other simulation scenarios were also examined but omitted in order to avoid repetition. It was found that the models for other simulation scenarios are broadly similar to the one of scenario 16 and that the larger the NO_2 simulated concentrations, the more variability in the data is explained. Overall, the modelling suggests that using temperature, wind speed, an interaction between temperature and wind speed, wind direction, emissions, a factor variable for the hours of the day and a b-spline with degree 3 is sufficient modelling which should be adjusted with an AR(1) correlation structure for the autocorrelation in the residuals. Based on these findings, a

hyperspatial-temporal model is developed in the next section to model all one hundred simulations from ADMS-Urban for Central Station in Glasgow together.

8.2 Hyperspatial-temporal model

This section introduces a hyperspatial-temporal model for emulation of the hourly NO_2 time series across the LHC space as simulated by ADMS-Urban. This emulator is created to create framework for stochastic simulation models as recommended by [169]. As in previous chapters, the term hyperspatial is used as a reference to the locations within the LHC space rather than physical locations in Glasgow. The proposed model is for a single station only. The section is organised as follows: Subsection 8.2.1 introduces the theoretical background for the hyperspatial-temporal model. Subsection 8.2.2 discusses the prediction of time series at new locations in the LHC space using the hyperspatial-temporal model.

8.2.1 Theoretical background

In Section 6.1, the methodology for fitting a spatial multivariate Bayesian GP model was presented. In this subsection, the work is extended to a hyperspatial-temporal model to allow the modelling of the time series (with length T) for all the scenarios (n in total) in the LHC design at a single monitoring station. Part of this change is that the design matrix has to be adjusted to allow for the different time series of emissions, wind speed, and wind direction at each of the LHC design points. The proposed model is expressed in a vector-matrix form as:

$$\mathbf{y} = \mathbf{S}\boldsymbol{\beta} + \mathbf{z}\,, \tag{8.3}$$

where:

- $\mathbf{y}$ is a response vector ($Tn \times 1$), where the hourly NO_2 time series (in this specific case, $T = 1, \ldots, 8760$) for each scenario n in the LHC design are stacked on each other. Hence, for the one hundred ADMS-Urban simulations in this example, $\mathbf{y}$ is given as:

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,8760} \\ y_{2,1} \\ \vdots \\ y_{2,8760} \\ \vdots \\ y_{100,1} \\ \vdots \\ y_{100,8760} \end{bmatrix}; \tag{8.4}$$

- $\mathbf{S}$ is a block-diagonal design matrix ($Tn \times pn$), where each block $\mathbf{B}_i$ ($T \times p$) contains the time series of the p covariates for scenario i ($i = 1, \ldots, n$). Therefore, different scenarios can have different time series for some or all of their covariates. Therefore, for the one hundred ADMS-Urban simulation scenarios in the LHC space:

$$\mathbf{S} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{B}_{100} \end{bmatrix}, \tag{8.5}$$

  where each $\mathbf{B}_i$ has the form:

$$\mathbf{B}_i = \begin{bmatrix} 1 & x_{1i,1} & x_{2i,1} & \ldots & x_{(p-1)i,1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{1i,8760} & x_{2i,8760} & \ldots & x_{(p-1)i,8760} \end{bmatrix}; \tag{8.6}$$

- $\boldsymbol{\beta}$ is a vector ($pn \times 1$) of the fixed parameters to be estimated. In terms of the one hundred ADMS-Urban scenarios, $\boldsymbol{\beta}$ has the form:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{0,1} \\ \beta_{x_1,1} \\ \vdots \\ \beta_{x_{p-1},1} \\ \beta_{0,2} \\ \beta_{x_1,2} \\ \vdots \\ \beta_{x_{p-1},2} \\ \vdots \\ \beta_{0,100} \\ \beta_{x_1,100} \\ \vdots \\ \beta_{x_{p-1},100} \end{bmatrix}, \tag{8.7}$$

where each $\beta_{0,i}$ is the intercept term for the i^{th} simulation scenario and each $\beta_{x_q,i}$ is the fixed effect term for the q^{th} $(q = 1, \ldots, (p-1))$ covariate under the i^{th} simulation scenario; and

- $\mathbf{z}$ is an error vector $(Tn \times 1)$, which has a normal distribution $\mathbf{z} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Lambda})$. For the one hundred ADMS-Urban simulation scenarios, $\mathbf{z}$ has the form:

$$\mathbf{z} = \begin{bmatrix} z_{1,1} \\ \vdots \\ z_{1,8760} \\ z_{2,1} \\ \vdots \\ z_{2,8760} \\ \vdots \\ z_{100,1} \\ \vdots \\ z_{100,8760} \end{bmatrix}. \tag{8.8}$$

Hence, the response follows a normal distribution:

$$\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Lambda} \sim \mathbf{N}(\mathbf{S}\boldsymbol{\beta}, \boldsymbol{\Lambda}). \tag{8.9}$$

The variance-covariance matrix $\boldsymbol{\Lambda}$ $(Tn \times Tn)$ is assumed to be separable in terms of hyperspatial correlation (between simulation scenario correlation within the LHC space in a similar fashion as the model in Section 6.1) and time series correlation. It is defined as:

$$\boldsymbol{\Lambda} = \sigma^2 \, \mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho) \,, \tag{8.10}$$

where:

- σ^2 is the overall variance parameter to be estimated;
- $\mathbf{R}(\boldsymbol{\theta})$ is the hyperspatial (i.e. between simulation scenarios) correlation matrix $(n \times n)$ based on the LHC design as described in Section 6.1. Here, an exponential correlation function between two rows $\mathbf{u}_i = [u_{i\mathrm{E}}, u_{i\mathrm{WS}}, u_{i\mathrm{WD}}]^\top$ and $\mathbf{u}_j = [u_{j\mathrm{E}}, u_{j\mathrm{WS}}, u_{j\mathrm{WD}}]^\top$ of the input space (designed by the LHC) is specified as:

  $$\mathbf{R}_{ij}(\boldsymbol{\theta}) = \exp\left(-\left\{\left(\frac{|u_{i\mathrm{E}} - u_{j\mathrm{E}}|}{\theta_{\mathrm{E}}}\right) + \left(\frac{|u_{i\mathrm{WS}} - u_{j\mathrm{WS}}|}{\theta_{\mathrm{WS}}}\right) + \left(\frac{|u_{i\mathrm{WD}} - u_{j\mathrm{WD}}|}{\theta_{\mathrm{WD}}}\right)\right\}\right), \tag{8.11}$$

  with the set of hyperspatial range parameters $\boldsymbol{\theta} = [\theta_{\mathrm{E}}, \theta_{\mathrm{WS}}, \theta_{\mathrm{WD}}]^\top$ to be estimated; and
- $\boldsymbol{\Sigma}(\rho)$ is the temporal correlation matrix $(T \times T)$ with an AR(1) structure (based on the preliminary analysis in Section 8.1) of the form:

  $$\boldsymbol{\Sigma}(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \vdots & \vdots & \ddots & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{bmatrix}, \tag{8.12}$$

  where ρ is the temporal correlation parameter to be estimated. It is assumed that $\boldsymbol{\Sigma}(\rho)$ is the same for all scenarios.

The set of parameters ($\boldsymbol{\beta}$, σ^2, $\boldsymbol{\theta}$, ρ) are assigned a joint prior distribution, where the parameters are assumed to be independent:

$$f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \rho) = f(\boldsymbol{\beta}) f(\sigma^2) f(\boldsymbol{\theta}) f(\rho) \,. \tag{8.13}$$

The prior for the set of fixed effect parameters $\boldsymbol{\beta}$ is decomposed as:

$$f(\boldsymbol{\beta}) = f(\beta_{0,1}) f(\beta_{x_1,1}) \cdots f(\beta_{x_{p-1},100}) \,, \tag{8.14}$$

where each $f(\beta_l) \propto 1$ for $l = 1, \dots, pn$ is a non-informative improper flat prior on the positive real line so that the "data speak for themselves".

A weakly informative univariate flat prior (as suggested in [88]) is chosen for the overall variance parameter σ^2:

$$f(\sigma^2) \propto \text{Un}(0.01, 10000)\,. \tag{8.15}$$

A uniform prior is chosen for the autocorrelation coefficient ρ as it is expected that the temporal correlation between successive hours is positive:

$$f(\rho) \sim \text{Un}(0, 1)\,. \tag{8.16}$$

The prior for the set of hyperspatial range parameters $\boldsymbol{\theta}$ is decomposed as:

$$f(\boldsymbol{\theta}) = f(\theta_{\text{E}}) f(\theta_{\text{WS}}) f(\theta_{\text{WD}})\,, \tag{8.17}$$

where $f(\theta_{\text{E}}) \propto 1$, $f(\theta_{\text{WS}}) \propto 1$ and $f(\theta_{\text{WD}}) \propto 1$ are non-informative improper flat priors on the positive real line. Hence, the parameters are estimated by using the posterior distribution:

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \rho|\mathbf{y}) &\propto f(\boldsymbol{\beta}) f(\sigma^2) f(\boldsymbol{\theta}) f(\rho) \\ &\quad \times \exp\left[-\frac{1}{2}\log\left(|\boldsymbol{\Lambda}\left(\sigma^2, \boldsymbol{\theta}, \rho\right)|\right) - \frac{1}{2}\left(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Lambda}\left(\sigma^2, \boldsymbol{\theta}, \rho\right)^{-1} \left(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}\right)\right]. \end{aligned} \tag{8.18}$$

The log of this posterior distribution can be simplified and is given by:

$$\begin{aligned} \log\left(f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \rho|\mathbf{y})\right) \propto &-\frac{Tn}{2}\log\left(\sigma^2\right) - \frac{1}{2}\log\left(|\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)|\right) \\ &- \frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}\right)^{\top} \left[\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)\right]^{-1} \left(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}\right). \end{aligned} \tag{8.19}$$

It is possible to obtain closed-form solutions for $(\boldsymbol{\beta},\, \sigma^2)$ using differentiation. Differentiating the log-likelihood with respect to $\boldsymbol{\beta}$ gives:

$$\frac{\partial l(\mathbf{y})}{\partial \boldsymbol{\beta}} = \frac{\mathbf{S}^{\top}\left[\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)\right]^{-1}\mathbf{y}}{\sigma^2} - \frac{\mathbf{S}^{\top}\left[\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)\right]^{-1}\mathbf{S}\boldsymbol{\beta}}{\sigma^2}\,. \tag{8.20}$$

Setting this partial derivative to zero and solving for $\boldsymbol{\beta}$ produces:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{S}^{\top}\left(\mathbf{R}\left(\boldsymbol{\theta}\right) \otimes \boldsymbol{\Sigma}\left(\rho\right)\right)^{-1}\mathbf{S}\right)^{-1} \mathbf{S}^{\top}\left(\mathbf{R}\left(\boldsymbol{\theta}\right) \otimes \boldsymbol{\Sigma}\left(\rho\right)\right)^{-1}\mathbf{y}\,, \tag{8.21}$$

which is equivalent to the GLS closed form solution (see Subsection 2.2.1). Hence, the variance-covariance matrix of the parameters is estimated as:

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{S}^{\top}\left(\mathbf{R}\left(\boldsymbol{\theta}\right) \otimes \boldsymbol{\Sigma}\left(\rho\right)\right)^{-1}\mathbf{S}\right)^{-1}. \tag{8.22}$$

In a similar way, the log-likelihood is differentiated with respect to σ^2:

$$\frac{\partial l(\mathbf{y})}{\partial \sigma^2} = -\frac{Tn}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{S}\boldsymbol{\beta})^\top [\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)]^{-1} (\mathbf{y} - \mathbf{S}\boldsymbol{\beta})}{2\sigma^4} . \tag{8.23}$$

Setting this partial derivative to zero and solving for σ^2 yields:

$$\widehat{\sigma}^2 = \frac{1}{(Tn)} \left(\mathbf{y} - \mathbf{S}\widehat{\boldsymbol{\beta}}\right)^\top (\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho))^{-1} \left(\mathbf{y} - \mathbf{S}\widehat{\boldsymbol{\beta}}\right) . \tag{8.24}$$

Nonetheless, this estimate is biased [98]. Therefore, an alternative is used:

$$\widehat{\sigma}^2 = \frac{1}{(Tn - pn)} \left(\mathbf{y} - \mathbf{S}\widehat{\boldsymbol{\beta}}\right)^\top (\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho))^{-1} \left(\mathbf{y} - \mathbf{S}\widehat{\boldsymbol{\beta}}\right) . \tag{8.25}$$

However, the log-likelihood cannot be differentiated with respect to $\boldsymbol{\theta}$ or ρ and closed form solutions cannot be obtained. A natural approach would be to perform a Markov Chain Monte Carlo (MCMC) method to sample from the posterior distributions for $\boldsymbol{\theta}$ and ρ. However, the approach is "computationally cumbersome" [143] as it would require inverting and taking derivatives of $\mathbf{R}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\rho)$ at each step of the MCMC process. Hence, a plug-in numerical approximation is recommended by both [49] and [143]. Therefore, the BFGS algorithm as described in Subsection 2.2.4 and applied in Chapters 5 and 6 is instead applied via the *optim* function in **R** [157]. The BFGS algorithm is used to maximise the log-posterior expression over the set of parameters $(\boldsymbol{\theta}, \rho)$. The estimates of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are plugged into the log-posterior and, thus a profile log-posterior expression is produced. The profile log-posterior is then optimised over ($\boldsymbol{\theta}$, ρ):

$$f(\boldsymbol{\theta}, \rho|\mathbf{y}) \propto -\frac{Tn}{2}\log\left(\widehat{\sigma}^2\right) - \frac{1}{2}\log\left(|\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)|\right) . \tag{8.26}$$

Hence, point estimates are obtained for $\boldsymbol{\theta}$ and ρ and closed form solutions for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$. Since the variance-covariance matrix $\boldsymbol{\Lambda}$ (σ^2, $\boldsymbol{\theta}$, ρ) has dimensions $Tn \times Tn$, there is a need for computational efficiency. Some properties of Kronecker products and AR(1) correlation matrices are used as described in Appendix C.

8.2.2 Prediction

The main reason to develop the hyperspatial-temporal emulator is to be able to predict the yearly time series of the (log) hourly NO_2 concentrations for a new location in the hyperspace. This will be done using the principles of kriging, which were introduced in Subsection 2.3.1. Let $\mathbf{y_0}$ be a vector $(Tn_\mathbf{0} \times 1)$ of stacked time series for untried sets of inputs of ADMS-Urban, with $\mathbf{S_0}$ being a block-diagonal design matrix $(Tn_\mathbf{0} \times n_\mathbf{0}p)$

containing the observed covariates at these $n_{\mathbf{0}}$ new locations in the LHC space. Then, it is assumed that $\mathbf{y_0}$ and $\mathbf{y}$ follow a joint multivariate normal distribution:

$$\begin{pmatrix} \mathbf{y_0} \\ \mathbf{y} \end{pmatrix} \sim \mathbf{N}\left[\begin{pmatrix} \boldsymbol{\mu_0} \\ \boldsymbol{\mu} \end{pmatrix}, \quad \begin{pmatrix} \boldsymbol{\Lambda_0} & \boldsymbol{\Lambda_{12}} \\ \boldsymbol{\Lambda_{21}} & \boldsymbol{\Lambda} \end{pmatrix} \right], \tag{8.27}$$

where:

- $\boldsymbol{\mu_0} = \mathbf{S_0}\boldsymbol{\beta_0}$ is the mean for $\mathbf{y_0}$, where $\boldsymbol{\beta_0}$ are the fixed effect parameters for the new set of covariates observed at the new locations $n_{\mathbf{0}}$;
- $\boldsymbol{\mu} = \mathbf{S}\boldsymbol{\beta}$ is the mean for $\mathbf{y}$;
- $\boldsymbol{\Lambda_0} = \sigma^2\,\mathbf{R_0}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)$ is the hyperspatial-temporal variance-covariance matrix $(Tn_{\mathbf{0}} \times Tn_{\mathbf{0}})$ for $\mathbf{y_0}$. $\mathbf{R_0}(\boldsymbol{\theta})$ is the hyperspatial correlation matrix $(n_{\mathbf{0}} \times n_{\mathbf{0}})$ for the new hyperspatial scenarios based on the estimated $\boldsymbol{\theta}$. $\boldsymbol{\Sigma}(\rho)$ is the temporal correlation matrix for $\mathbf{y}$ as it is assumed that the time series at the new hyperspatial scenarios also have length T;
- $\boldsymbol{\Lambda} = \sigma^2\,\mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)$ is the hyperspatial-temporal variance-covariance matrix for $\mathbf{y}$; and
- $\boldsymbol{\Lambda_{21}} = \boldsymbol{\Lambda_{12}}^{\top} = \sigma^2\,\mathbf{T}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)$ is the cross hyperspatial-covariance matrix $(Tn \times Tn_{\mathbf{0}})$ between $\mathbf{y_0}$ and $\mathbf{y}$. Therefore, $\mathbf{T}(\boldsymbol{\theta})_{ij}$ is the hyperspatial correlation between the i^{th} scenario from $\mathbf{y_0}$ with the j^{th} scenario from $\mathbf{y}$.

Using a property of the multivariate normal distribution, the conditional distribution of $\mathbf{y_0}$ is:

$$\mathbf{y_0}|\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu_0} + \boldsymbol{\Lambda_{12}}\boldsymbol{\Lambda}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \boldsymbol{\Lambda_0} - \boldsymbol{\Lambda_{12}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda_{21}})\,. \tag{8.28}$$

In this case, the variance of $\mathbf{y_0}$ can also be written as:

$$\text{Var}(\mathbf{y_0}) = \boldsymbol{\Lambda_0} - \boldsymbol{\Lambda_{21}}^{\top}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda_{21}}\,. \tag{8.29}$$

In order to estimate the predicted values for $\mathbf{y_0}$ and their respective variances, the hyperspatial correlation matrix $\mathbf{R_0}(\boldsymbol{\theta})$ needs to be estimated based on the already calculated hyperspatial range parameters $\widehat{\boldsymbol{\theta}}$ from the model for $\mathbf{y}$:

$$\widehat{\mathbf{R_0}(\boldsymbol{\theta})} = \mathbf{R_0}(\widehat{\boldsymbol{\theta}})\,. \tag{8.30}$$

Similarly, the temporal correlation matrix $\boldsymbol{\Sigma}(\rho)$ is calculated using the AR(1) parameter estimate $\widehat{\rho}$ from the model for $\mathbf{y}$:

$$\widehat{\boldsymbol{\Sigma}(\rho)} = \boldsymbol{\Sigma}(\widehat{\rho})\,. \tag{8.31}$$

By definition, in the hyperspatial-temporal model for $\mathbf{y}$, each time series i ($i = 1, \ldots, 100$) has a different set of fixed effect parameters $\boldsymbol{\beta}_i = [\beta_{0,i}, \beta_{x_1,i}, \ldots, \beta_{x_{p-1},1}]^\top$, which reflect the change in the covariates within the LHC space from varying the three inputs (emissions, wind speed and wind direction). Hence, the full set of fixed effect parameters $\widehat{\boldsymbol{\beta}}$ are calculated using the estimated hyperspatial range parameters $\widehat{\boldsymbol{\theta}}$ and the estimated AR(1) temporal parameter $\widehat{\rho}$. Therefore, the fixed effect parameters for the new set of time series $\boldsymbol{\beta_0}$ need to be estimated to reflect the changed emissions, wind speed and wind direction based on the new scenario locations in the LHC space. To do this, univariate hyperspatial models (such as the ones used in Chapter 5) are used. Firstly, $\widehat{\boldsymbol{\beta}}$ is split into subsets based on the type of entry (intercept, temperature, etc.). In this specific case, $\widehat{\boldsymbol{\beta}}_{\text{intercept}} = [\beta_{0,1}, \ldots, \beta_{0,100}]^\top$ is a vector of size 100×1, which contains the one hundred intercepts estimated for each of the one hundred ADMS-Urban simulation scenarios. Then, $\widehat{\boldsymbol{\beta}}_{\text{intercept}}$ is taken to be the response and the three inputs (emissions and wind speed as % change, and wind direction as $^\circ$ change) are used as covariates. Based on that univariate hyperspatial model, $\boldsymbol{\beta}_{0,\text{intercept}}$ can be estimated for untried sets of inputs for ADMS-Urban. The process is then repeated to estimate each of the sets $\boldsymbol{\beta}_{0,x_1}, \ldots, \boldsymbol{\beta}_{0,x_{p-1}}$. Details of this model are given in Section 5.1.

Using $\widehat{\sigma}^2$, the $\boldsymbol{\Lambda_0}$ and $\boldsymbol{\Lambda_{21}}$ (and respectively, $\boldsymbol{\Lambda_{12}}$) elements of the variance-covariance matrix between $\mathbf{y}$ and $\mathbf{y_0}$ can be calculated as follows:

$$\widehat{\boldsymbol{\Lambda}}_{\mathbf{0}} = \widehat{\sigma}^2\, \mathbf{R_0}(\widehat{\boldsymbol{\theta}}) \otimes \boldsymbol{\Sigma}(\widehat{\rho})\,, \tag{8.32}$$

and

$$\widehat{\boldsymbol{\Lambda}}_{\mathbf{21}} = \widehat{\sigma}^2\, \mathbf{T}(\widehat{\boldsymbol{\theta}}) \otimes \boldsymbol{\Sigma}(\widehat{\rho})\,. \tag{8.33}$$

8.3 Application of the hyperspatial-temporal model

In Appendix D, it is shown that the hyperspatial-temporal model provides parameter estimates for the hyperspatial range and temporal parameters that are very close to the true values for much smaller data sets than the one hundred simulation scenarios each with 8760 hourly concentrations time series. However, using the hyperspatial-temporal model on the full data set is not possible. The main problem is the size of the matrices, which are needed for the estimates. For instance, the kronecker product of $\mathbf{R}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\rho)$, required to estimate the fixed effect parameters $\boldsymbol{\beta}$ and the overall variance σ^2, would have a size of around 6TB (this estimate is based on the fact that as the time

steps are doubled, the size of the kronecker product matrix is quadrupled). Furthermore, it is estimated that one iteration of the BFGS algorithm would take 16 days (based on the fact that as time steps are doubled, the time to estimate one iteration of the BFGS algorithm triples). Hence, at a rate of 15 iterations (the lowest number of iterations in the model testing in Appendix D), running the full model would take 240 days, i.e. almost 9 months. These estimates are based on a 2016 MacBook Pro with 16GB memory and 2.9 GHz Quad-Core Intel Core i7 processor. Therefore, applying the hyperspatial-temporal model to the full data set available is not practical as the emulator would not be computationally faster than ADMS-Urban. Hence, an alternative approach is adopted in this section. Instead of applying the hyperspatial-temporal model to the full data set, the model would be applied to overlapping blocks of time and the blocks will then be put back together to get the larger time frame. In this specific application, the hyperspatial-temporal model will be applied to blocks of time periods, where breaches of the 200 μg m^{-3} regulation are observed, and it will be assessed how well the blocks of time periods match onto each other. Additionally, when splitting the data set into blocks, multiple models can be run simultaneously on different cores of the computer resulting in further time reduction.

8.3.1 Subsetting the data

In order to decide on a period of interest, the time series for simulation 16 is re-examined in Figure 8.14. As previously discussed, each simulation scenario is created based on varying emissions (in % change), wind speed (in % change) and wind direction (in $^\circ$ change) to an observed baseline. Therefore, the one hundred ADMS-Urban simulations have different absolute NO_2 hourly concentrations but the time series have similar trend. Looking at Figure 8.14, the highest peak is in the beginning of February (5/02). However, that peak has been preceded by another peak around mid-January (18/01). Although there are other periods (7/07 and 7/11) when the hourly NO_2 concentration is close to the regulation of 200 μg m^{-3}, there are no further clear breaches. Therefore, it is concluded that in terms of exploring the conditions which result in breaches, it is of interest to focus on the period between 17/01 to 7/02. In order to be able to assess how well the proposed segmenting into overlapping time blocks of interest performs, the period is split into two parts: 17/01 - 29/01, which will be referred to as the **January data**, and 28/01 - 7/02, which will be referred to as the **February data**.

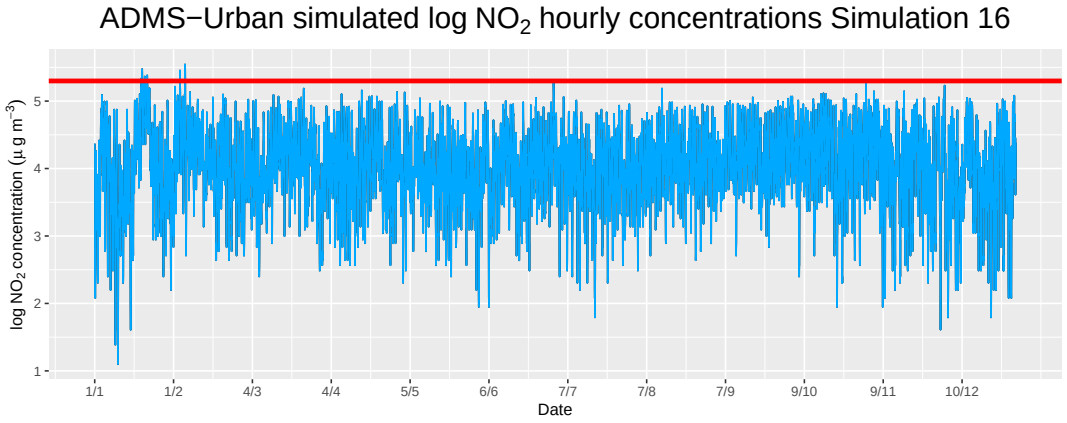

FIGURE 8.14: Time series for the ADMS-Urban simulation scenario 16 for the log NO_2 hourly concentrations. A red line signifies the log 200 μg m^{-3} regulation.

8.3.2 Hyperspatial-temporal modelling

Firstly, the hyperspatial-temporal model was fitted for the January data set (17/01 - 29/01). The model has the same covariates as in Equation 8.2 but as the subset has a relatively small time frame, the factor variable for hour of the day and the b-spline for Week Number covariates are not used. The estimated random effect parameters are $\widehat{\rho} = 0.67$ and $\widehat{\boldsymbol{\theta}} = [\theta_{EM} = 243.07, \theta_{WS} = 931.85, \theta_{WD} = 1000.00]^{\top}$. [1] The hyperspatial-temporal parameter estimates indicate high temporal correlation within each of the time series and high hyperspatial correlation between the time series in the LHC space. As in previous chapters, the out-of-sample predictive performance of the model was then assessed by using a 10-fold CV. Diagnostic plots were produced based on the out-of-sample 10-fold CV predictions and are presented in Figure 8.15. From the diagnostic plots in Figure 8.15, there is a clear issue with the tails in the qq-plot (plot c)). However, the histogram of the residuals (plot d)) shows that almost all residuals are very close to zero. After further investigation, it was found that the points, which cause the heavy tails on the qq-plot, come from a simulation scenario at the edge of the LHC space and removing it as part of the CV results in predictions for that scenario being extrapolated. However, the scenario has very low log NO_2 hourly concentrations, which also result in outliers on all other diagnostic plots and therefore, the heavy tails are not an indication of an overall issue with the fit. Furthermore, there are no indicators that there is need for more covariates. The points for simulation 16 in blue indicate the model performs very well especially for scenarios where breaches were observed.

[1] As the hyperspatial range parameter for wind direction is at the upper limit boundary, an alternative was adapted where *sin* (wind direction) and *cos* (wind direction) were also tested in order to define the hyperspace but both the *sin* and *cos* (wind direction) hyperspatial range parameters were also estimated as reaching the upper limit boundary and therefore, the *sin* and *cos* (wind direction) were discarded as hyperspatial range parameters and wind direction was used as it requires estimating a smaller number of hyperspatial range parameters which is consistent with the findings in [212].

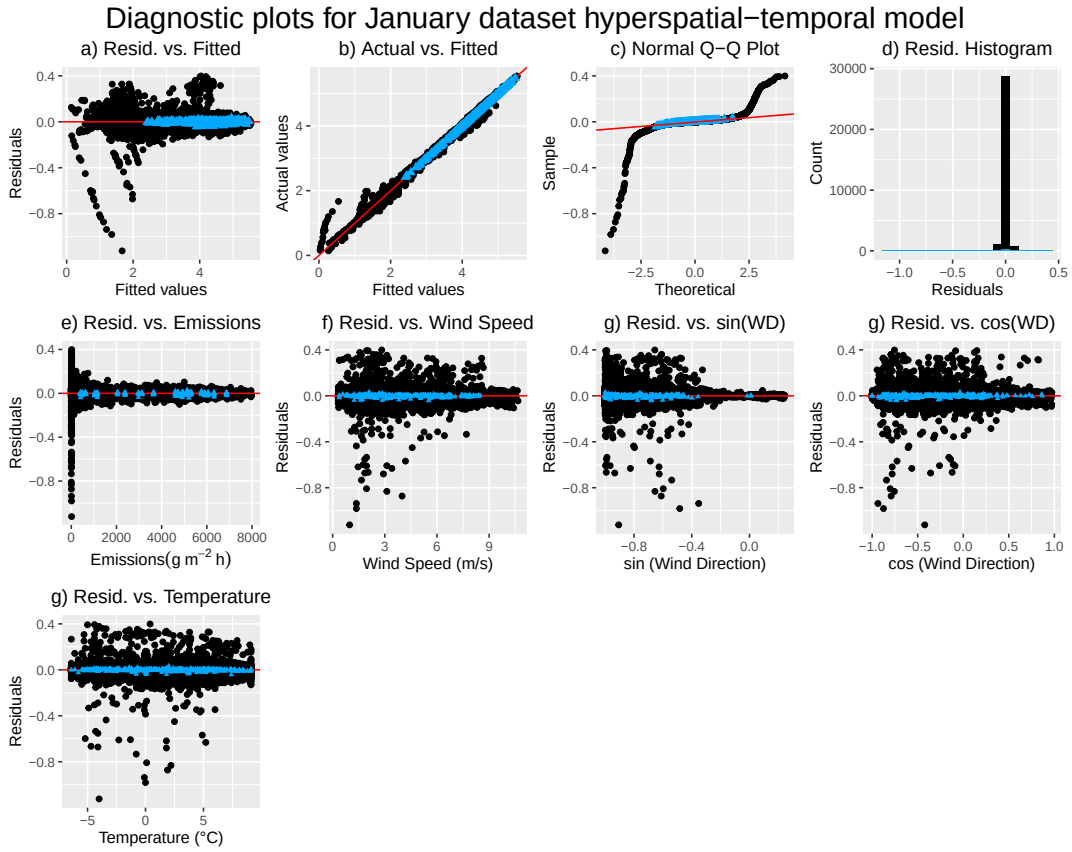

FIGURE 8.15: Diagnostic plots for the model for the January subset log hourly NO_2 measurements (μg m^{-3}) of ADMS-Urban at Central Station with temperature (°C), wind speed (m/s), an interaction between temperature and wind speed, sin and cos wind direction (°), and emissions (g m^{-2} h). The points for simulation scenario 16 are highlighted in blue.

Next, the hyperspatial-temporal model was fitted on the February data set (28/01 - 7/02). Similarly to the January modelling, the February model has the same covariates as in Equation 8.2 but as the subset has a relatively small time frame, the factor variable for hour of the day and the b-spline for Week Number covariates are not used. The random effect parameter estimates are $\widehat{\rho} = 0.51$ and $\widehat{\boldsymbol{\theta}} = [\theta_{EM} = 224.50, \theta_{WS} = 726.58, \theta_{WD} = 1000.00]^\top$. It is interesting to note that the temporal parameter and the emissions and wind speed hyperspatial range parameters for the February data set are lower than the ones for the January data set but for both data sets the hyperspatial range parameter for wind direction remains at the upper boundary. A 10-fold CV was performed to assess the out-of-sample predictive performance of the model. Based on the 10-fold CV results, diagnostic plots are produced and are presented in Figure 8.16. Similarly to the January model, the tails of the qq-plot (plot c)) are quite heavy but the histogram of the residuals (plot d)) show that in fact almost all residuals are very close to zero. Furthermore, the points, which cause the heavy tails on the qq-plot, are from the same scenario as in the January modelling and these points are a result of extrapolation. An interesting difference to the January modelling is that the outlier

points for the February modelling are over-predicted as opposed to under-predicted as in January. It has to be noted that the over-predicted values are for different simulation scenarios than in the January modelling. There are no indications that more covariates are needed for the modelling. Moreover, the points from simulation 16, highlighted in blue, show that the model performs well especially for scenarios with breaches.

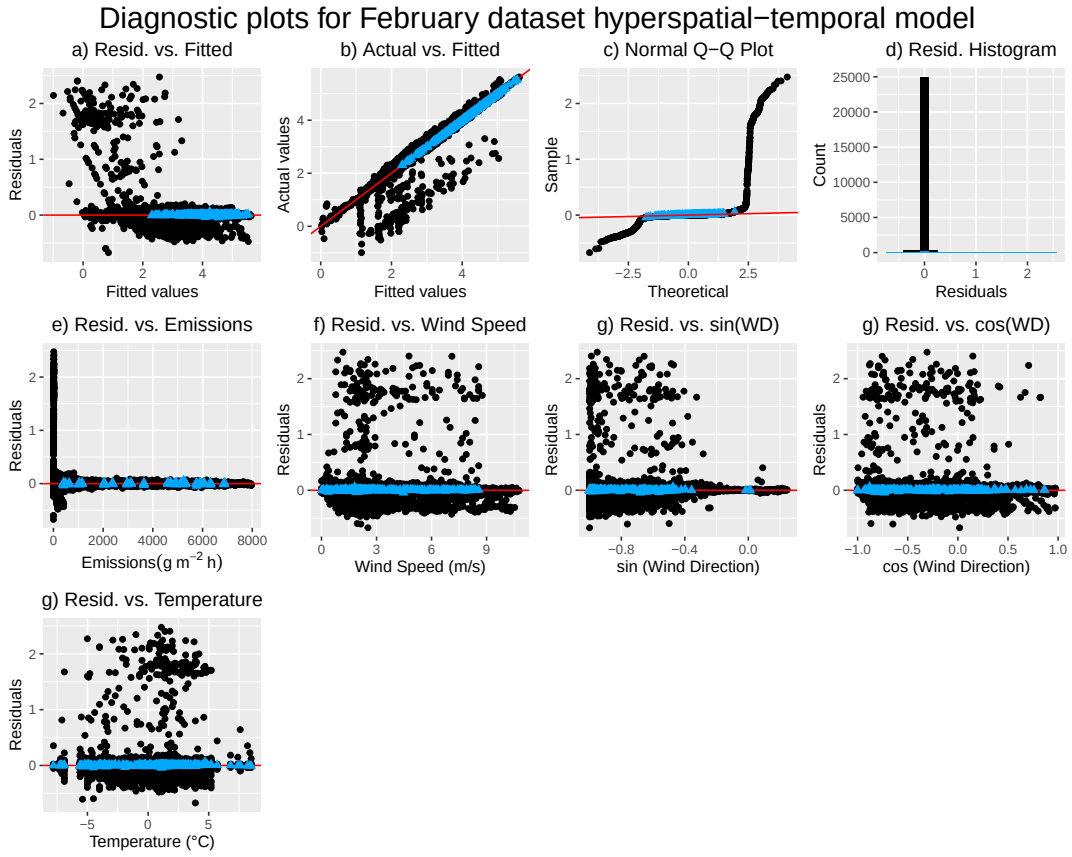

FIGURE 8.16: Diagnostic plots for the model for the February subset log hourly NO_2 measurements (μg m^{-3}) of ADMS-Urban at Central Station with temperature (°C), wind speed (m/s), an interaction between temperature and wind speed, sin and cos wind direction (°), and emissions (g m^{-2} h). The points for simulation scenario 16 are highlighted in blue.

Lastly, the predicted estimates for the log hourly NO_2 measurements from 28/01 to 29/01 (i.e. 48 hours) from both data sets are compared in order to establish whether there is a smooth transition between the two models. This is done in the scatterplot in Figure 8.17. The majority of the points lie on the equivalence line. Simulation scenario 16 (in blue triangles) lies perfectly on the equivalence line suggesting that there is a good overlap between the predictions from the two models. The plot shows that applying the model to overlapping blocks of time provides estimates of almost perfect absolute values for the log NO_2 concentrations. There a few points under the equivalence line but they are the result of the aforementioned extrapolation of a single scenario (with the January model predicting higher values than the February model) and hence, are not considered

an issue. Therefore, the two models for the January and February data will be used to predict time series for the NO_2 hourly concentrations within the time period of interest.

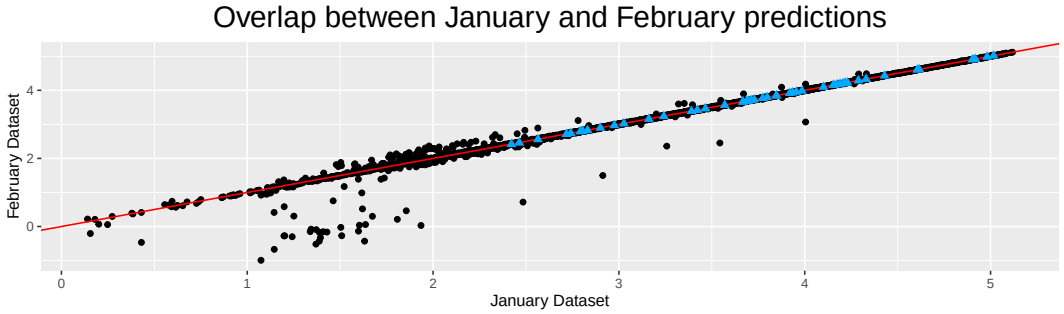

FIGURE 8.17: Scatterplot comparing the predictions for log hourly NO_2 measurements (μg m^{-3}) of ADMS-Urban at Central Station from 28/01 to 29/01 for both the January and February data sets. The points for simulation scenario 16 are highlighted in blue.

8.3.3 Hyperspatial-temporal emulation

To assess the quality of the January and February models, they were used to emulate three simulation scenarios. Simulation scenario 16 was emulated as it is the scenario closest to the $(0, 0, 0)$ coordinate point and hence, was used in the exploratory analysis. Simulation 16 has emissions that are higher by 3.20% than the observed emissions, wind speed that is lower by 4.80% from the wind speed in 2015, and wind direction that is changed by adding 1.49° to the hourly wind direction in 2015. Additionally, simulation 84 and 24 were chosen. Scenario 84 was chosen as it is closest to the centre of the LHC in terms of Euclidian distance with coordinates $(-42.42, 9.51, 2.07)$, which means that the emissions were 42.42% lower than the observed emissions in 2015, the wind speed was 9.51% higher than the observed in 2015 and the wind direction is changed by adding 2.07° to each hourly observation in 2015. Simulation 24 was chosen as it is the furthest away from the centre of the LHC in terms of Euclidian distance with coordinates $(-99.75, -2.61, -5.89)$. Therefore, simulation 24 has emissions which are 99.75% lower than the observed ones in 2015, wind speed lower by 2.61% than the 2015 recordings and 5.89° has been taken away from the observed hourly wind direction 2015. The 10-fold CV results from the modelling diagnostics from Subsection 8.3.2 are used in this subsection again. As the regulation is for the actual hourly NO_2 concentrations, these are presented in the plots rather than the log scale predictions.

Firstly, for simulation 16, the true simulation values were plotted against the emulated values for the January and February subsets in Figure 8.18. From the figure, it is clear that the emulated values are almost perfectly superimposed on top of the NO_2 hourly concentrations for scenario 16. Furthermore, the emulated values from the two models appear to overlap each other almost perfectly as well. The standard deviation

estimate for the January model is 0.03, whereas for the February model the standard error estimate is 0.02. For both models this results in very small prediction intervals for the NO_2 hourly measurements and hence, not included in the plot.

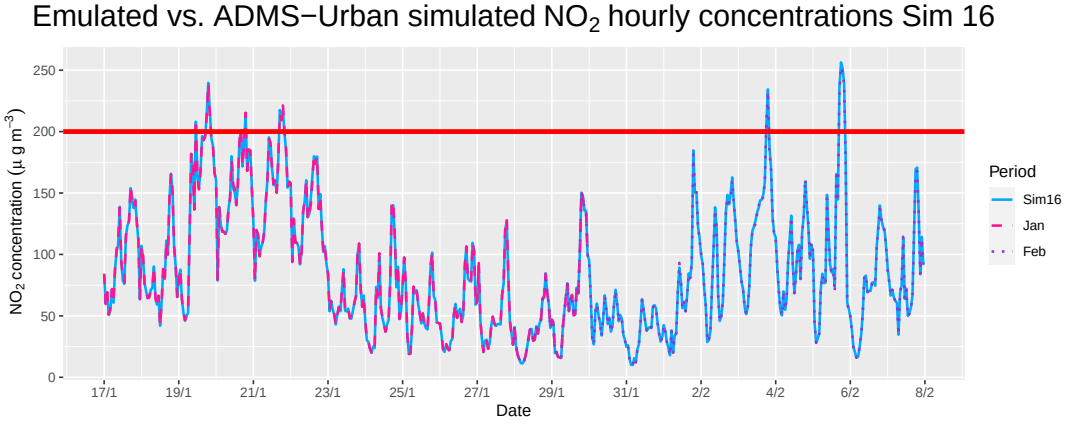

FIGURE 8.18: Time series comparing the emulated against the true simulation scenario 16 NO_2 hourly concentrations (μg m^{-3}) for the period 17/01-7/02. Simulation 16 is a blue solid line, the January emulated data is a pink dashed line and the February emulation is a purple dotted line. A red line signifies the 200 μg m^{-3} regulation.

Secondly, for simulation 84, the true simulation values were plotted against the emulated values for the January and February subsets in Figure 8.19. As with the plot for Simulation scenario 16, the lines from the January and February plots are almost perfectly superimposed on the true NO_2 concentrations for scenario 84. Again, the overlap between the emulated values from the two models is complete. Furthermore, the standard deviation estimate for the January model is 0.03, whereas for the February model the standard error estimate is 0.02. Again, this results in very small prediction intervals for the NO_2 hourly measurements and hence, not included in the plot.

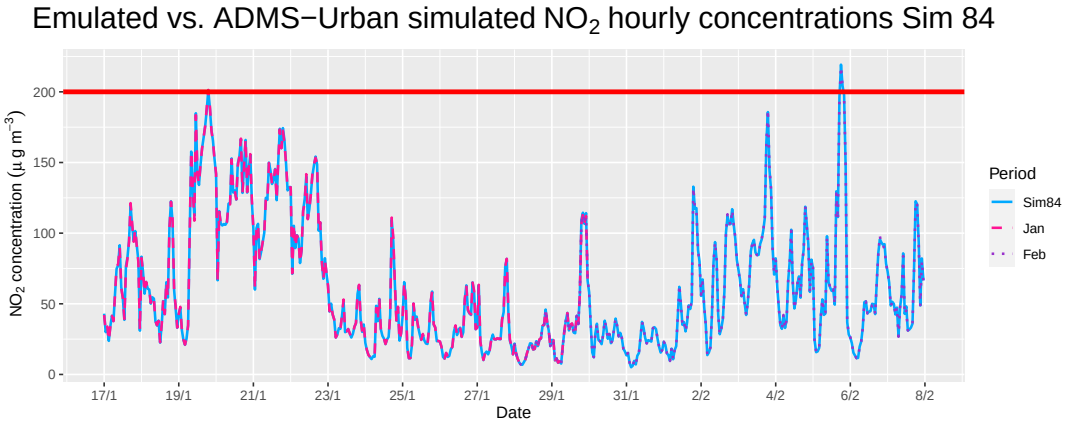

FIGURE 8.19: Time series comparing the emulated against the true simulation scenario 84 NO_2 hourly concentrations (μg m^{-3}) for the period 17/01-7/02. Simulation 84 is a blue solid line, the January emulated data is a pink dashed line and the February emulation is a purple dotted line. A red line signifies the 200 μg m^{-3} regulation.

Thirdly, the true simulation values for scenario 24 were plotted against the emulated values for the January and February subsets in Figure 8.20. As previously discussed, the emulation of scenario 24 is in fact an extrapolation and therefore, it is expected that there will be more clear differences between the lines on the plot in comparison to the previous two scenarios. From Figure 8.20, it is clear that both the January and February models have under-emulated the values of the scenario 24 NO_2 hourly concentrations in comparison with the actual simulated values. It appears that the February model produces lower values than the January model as for the 48 hour overlap period between the two models, the February model values are lower than the January ones. This is consistent with the observations in Figure 8.17. However, although in terms of absolute values, both models have under-predicted the magnitude of the NO_2 hourly concentrations, both models have identified the underlying general trend and correctly identified periods with spikes in the NO_2 hourly concentrations. Additionally, the standard deviations for both the January and February models are 0.03 (due to rounding). The resulting prediction intervals are not included in the plot due to their small size.

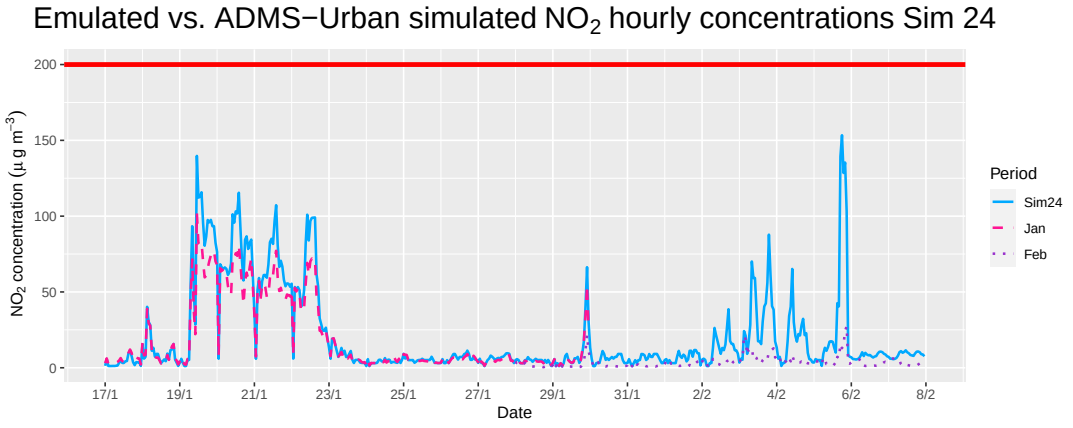

FIGURE 8.20: Time series comparing the emulated against the true simulation scenario 24 NO_2 hourly concentrations (μg m^{-3}) for the period 17/01-7/02. Simulation 24 is a blue solid line, the January emulated data is a pink dashed line and the February emulation is a purple dotted line. A red line signifies the 200 μg m^{-3} regulation.

Based on the good performance of the out-of-sample emulation on the overlapping January and February data sets for simulation scenarios 16, 84 and 24, the two models from Subsection 8.3.2 are used to emulate the ADMS-Urban time series for the baseline scenario, i.e. 0% change in emissions and wind speed, and 0° degrees change in wind direction. This will allow the comparison of the emulated simulated scenario to the real data, which is important for the reliability of using the hourly ADMS-Urban simulations for creating and justification of governmental policies. From Figure 8.21, it is clear that the January model has predicted that there will be a period of four days (between 19/1 and 23/1), when the hourly regulation of 200 μg m^{-3} will be breached for almost every

hour (74 occurrences above 200 μg m^{-3}). In reality, there were no breaches above 200 μg m^{-3}. The fact that the model has over-predicted the number of exceedances is not surprising given that in Chapter 7, it was emulated that there are expected 17 occurrences over 200 μg m^{-3} when only 4 are observed in reality. However, the number of exceedances is more than 4 times larger than expected, which indicates a period, where breaches are very likely to occur. Furthermore, the January model predicts another exceedance on 30/01. However, this is part of the overlap between the two models and interestingly, the February model has almost perfectly predicted the true hourly NO_2 concentrations for that time. As opposed to the January model, the February model follows very closely the true concentrations. The only exception is the exceedance on 5/2, where the model under-predicts the value of the breach over 200 μg m^{-3}. The February model appears to follow more closely the true concentrations. The January model has estimated the standard deviation at 0.03 and the February model estimate is 0.02, which are similar to those observed for scenarios 16 and 84. Once again, the prediction interval values were not included in the plot as they are too similar to the mean values.

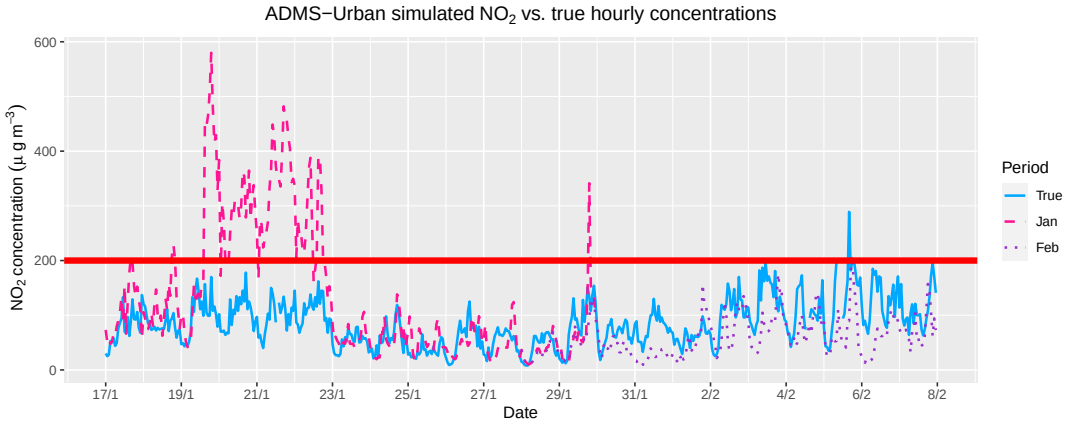

FIGURE 8.21: Time series comparing the emulated ADMS-Urban scenario against the true NO_2 hourly concentrations (μg m^{-3}) for the period 17/01-7/02. A red line signifies the 200 μg m^{-3} regulation.

8.3.4 Findings

The hyperspatial-temporal model was applied to two overlapping blocks of the ADMS-Urban data set. The models had fixed effect terms as in Equation 8.2 but as the subsets have a relatively small time frame, the factor variable for hour of the day and the b-spline for Week Number covariates are not used. The estimates for the temporal parameters indicated moderate autocorrelation within scenarios and the estimates for the hyperspatial range parameters indicated high between scenario correlation within the LHC space. Both models performed well in mimicking the ADMS-Urban simulations

based on their out-of-sample predictions. However, both models had different issues with extrapolating a single scenario during the 10-fold CV - the January model struggled by slightly under-predicting low log NO_2 hourly concentrations, whereas the February model struggled by over-predicting log NO_2 hourly concentrations. Nevertheless, the overlapping log hourly NO_2 measurements from both models matched each other well and the overlapping blocks method provides a good alternative to not being able to run the hyperspatial-temporal model over the full time steps simulated by ADMS-Urban.

These issues were further highlighted when the plots comparing the ADMS-Urban NO_2 hourly concentrations time series for simulation scenarios 16, 84 and 24 were compared to the emulated time series from the January and February models. For scenarios 16 and 84, the overlap between the two models is perfect and the emulated values are almost perfectly imposed on top of the ADMS-Urban values. For Scenario 24, the absolute NO_2 hourly concentrations are lower than the true ADMS-Urban ones. Furthermore, for the overlapping period between the two models, the February model has emulated lower values than the January model. However, both models have very small standard deviation estimates and have managed to capture the overall underlying trend in the data. Overall, there appears to be good agreement in the overlapping times in two of the scenarios and the emulated values follow the underlying trend in the data.

Therefore, the two models were used to emulate what ADMS-Urban would simulate for the observed conditions in 2015. It was found that the January model over-predicts breaches over 200 μg m^{-3}, whereas the February model identifies correctly a breach, but under-predicts its magnitude by about 40 μg m^{-3}. These results are in line with the out-of-sample 10-fold CV prediction results. According to the emulated simulation, the hourly regulation for NO_2 concentrations would have been broken in Glasgow by 23/1 but in reality that is not the case. Therefore, the period between 19/1 and 23/1 could be classified of high interest to governmental agencies to investigate the discrepancy between the emulated simulation and the true concentrations in order to identify how breaches of the regulation were avoided. This discrepancy is likely to be coming from the fact that there is a lack of day-to-day variation in these ADMS-Urban simulations.

8.4 Conclusion

In this chapter, the hourly NO_2 concentrations for a year as simulated by ADMS-Urban for varying emissions, wind speed and wind direction were examined in order to create a modelling technique with which to examine how these varying conditions affect the hourly NO_2 concentrations and hence, enable governmental agencies to identify the conditions which lead to high pollutant concentrations. Firstly, an exploratory analysis

of the simulated by ADMS-Urban time series under one hundred different sets of emissions, wind speed and wind direction was performed in Section 8.1. The exploratory analysis of simulation scenario 16 (chosen due to its proximity to $(0, 0, 0)$) was presented in full detail. Other scenarios were also examined and it was found that they required broadly similar modelling. Therefore, it was concluded that temperature, wind speed, an interaction between temperature and wind speed, wind direction, emissions, a factor variable for the hours of the day and a b-spline with degree 3 are sufficient covariates with an AR(1) correlation structure adjusting for the autocorrelation in the residuals.

Based on the findings from the exploratory analysis, a hyperspatial-temporal model was proposed in Section 8.2. The model proposes using a block-diagonal design matrix, which allows for different covariate values to be used for the modelling of different simulation scenarios. Properties of Kronecker products and AR(1) correlation matrices are used for computational efficiency. As in Chapters 5 and 6, the proposed models struggled with correctly estimating the hyperspatial parameters, a short model testing was performed (see Appendix D) and it was found that even for a much smaller subset of the full ADMS-Urban scenarios (reduced number of time steps, but the same number of scenarios), the proposed hyperspatial-temporal model estimates the hyperspatial range parameters with almost no bias (less than 1%) and the temporal parameter with a slight negative bias of 10%.

As the hyperspatial-temporal model provides reasonable estimates for the random effects, the model was then applied to the ADMS-Urban simulations in Section 8.3. Applying the hyperspatial-temporal model to the full data set was deemed impractical. An alternative approach of identifying overlapping blocks of data for periods of interest was adopted. The January period from 17/01 to 29/01 and the February period from 28/01 to 7/02 were chosen based on the fact that simulation scenario 16 contains breaches over 200 μg m^{-3}. The periods were chosen to overlap in order to assess how well the predictions from the two models would match onto each other. The January model seemed to perform really well but slightly under-predicting the simulated hourly NO_2 concentrations for one of the scenarios as a result of extrapolation. On the other hand, the February model over-predicted the simulated hourly NO_2 concentrations for several scenarios. The results from the 10-fold CV were used to emulate three scenarios and compare and assess the performance of the overlapping technique across the LHC space. For two of the three scenarios (the one closest to the $(0, 0, 0)$ baseline point and the one closest to the centre of the LHC), the emulated NO_2 hourly concentrations from the January and February model overlapped each other perfectly. Furthermore, the emulated NO_2 hourly concentrations are almost perfectly superimposed on top the true ADMS-Urban simulated values. However, for the third scenario (furthest away from the centre of the LHC space), the emulated NO_2 hourly concentrations did not overlap each

other well with the February values being lower than the January ones, and both the January and February model under-predicted the true ADMS-Urban simulated values. Although the emulated NO_2 hourly concentrations from both the January and February models were lower than the true ADMS-Urban simulated values, both models had identified the underlying general trend in the time series. The issues with the quality of the emulated NO_2 hourly concentrations in the third scenario were caused due to extrapolation. Nevertheless, the standard deviation estimates for all models were very low (0.02 to 0.03). Overall, the two models mimicked very closely out-of-sample the ADMS-Urban simulated hourly NO_2 concentrations.

Therefore, the two models were then used to emulate the simulated hourly NO_2 concentrations based on the conditions for 2015. The January model over-predicted the hourly NO_2 concentrations for a period of almost 4 days by forecasting hourly concentrations of above 200 μg m^{-3} but in fact there were no breaches in reality. This overestimation is likely the result of the lack of variability in terms of day of the week. On the other hand, the February model followed very closely the true NO_2 concentrations and correctly identified a breach, although under-predicted its magnitude. The overlapped values between the two models and true NO_2 concentration were almost identical with the exception of the January model predicting a single breach above 200 μg m^{-3} on 30/01 which the February model did not predict and in fact such a breach did not occur in reality. The overlapping method gives good results and identifies a period (19/01 to 23/01) which is expected to have high hourly NO_2 concentrations which do not occur. Hence, it would be of interest to governmental agencies to further investigate the conditions at that time to identify the discrepancy. Overall, the hyperspatial-temporal model does well in mimicking ADMS-Urban but struggles with identifying when the true hourly NO_2 concentrations would be above 200 μg m^{-3} and the magnitude of the breaches.

The emulation by using overlapping blocks of time provides a computational efficiency without comprising the quality of the emulated NO_2 hourly concentrations. However, the method would still require fitting 35 models to be able to emulate a full year ADMS-Urban run, which would take over 3 months on a 2016 MacBook Pro with 16GB memory and 2.9 GHz Quad-Core Intel Core i7 processor. The running time can be further reduced if multiple models are fitted simultaneously. Therefore, the overlapping blocks of time approach provides certain computational efficiency as opposed to running the ADMS-Urban model.

Chapter 9

Discussion and conclusions

Air pollution is one of most serious environmental problems faced by modern society. Multiple international organisations (WHO, EU, SEPA) are investigating air pollution and developing strategies to reduce air pollution due to its effect on people's health. In this thesis, the main focus is on the statistical modelling of both measured (monitored) and simulated from an air quality model (ADMS-Urban) data in Scotland. The air pollution regulations in Scotland are based on EU Directive 2008/50/EC [76], although the Scottish government aims to become the first country in the world to introduce much stricter regulations outlined by WHO in [208]. Air pollution regulations define the monitoring of multiple pollutants as different pollutants have different effects on people's health. In Scotland, monitoring NO_2 is specifically crucial as NO_2 breaches the Scottish regulation at multiple locations. The regulation is split into two parts - annual average mean regulation, which cannot exceed 40 μg m^{-3}; and hourly mean regulation, which cannot exceed 200 μg m^{-3} more than 18 times a year.

However, air pollution data are expensive to acquire as the monitoring systems used (for instance, AURN monitors) are very expensive to operate. This results in a very sparse monitoring network. There are two possible solutions to this problem discussed in the thesis. The first option is using miniature automated sensors, which are lower cost than the AURN monitors. The second option is to use data from process models such ADMS-Urban. The advantage of data from ADMS-Urban is that it provides estimates for the air pollution for meteorological conditions (such as wind speed and wind direction), which have not been observed, and for locations, which have not been monitored. Nevertheless, model runs can be computationally time consuming. Emulation can be used to model the simulation results and hence, attempt to reduce the time required for simulating data. Details on these air pollution modelling investigations are presented below.

9.1 Assessing the quality of miniature automated sensors

In Chapter 3, a study is performed to assess the quality of NO_2 and O_3 hourly concentrations from the miniature automated sensor ALPHASENSE B2 in a realistic setting for citizen science application. The sensors produce two types of measurements from its two electrodes - the auxiliary and the working electrodes. Three sensors were deployed next to an AURN sensor at St. Leonards monitoring station in Edinburgh. Bland-Altman analysis was used in order to establish the consistency between the hourly measurements taken by the three sensors. It was found that the auxiliary electrode measurements taken by the sensors are not consistent with each other but the working electrode measurements were mostly consistent with each other. Following that, linear regression was applied to examine the relationship between the measurements from the sensors and the AURN monitor in order to assess how well air pollution is measured by the miniature automated sensors. Although for all models the main pollutant from the reference monitor was significant and captured the changes in the hourly pollutants' concentrations, it was found that the miniature automated sensors' hourly measurements are also heavily influenced by changes in hourly temperature and relative humidity. From the analysis it can be concluded that while these lower cost sensors are useful, this experiment indicated that the technology is not yet fully foolproof when used by non-specialists. Looking forward there will be interesting statistical questions concerning how a network of monitoring stations could be designed, combining both the high and low quality air pollution sensors.

9.2 Models and emulation of the ADMS-Urban simulation scenarios

The rest of the thesis focused on the idea of emulating simulated data generated from ADMS-Urban model, which has been used in many Scottish cities, to further explore the conditions, which cause increased NO_2 pollutant concentrations, than just using the observed monitoring data. The main aim of the work is to emulate a deterministic computer simulation model ADMS-Urban and establish a framework to work with similar stochastic simulation models as suggested in [169]. Doing so allows to better understand under what conditions both the annual and hourly exceedance regulations are breached. In Chapter 4, two data sets were introduced. Firstly, the ADMS-Urban simulated data set for the city of Aberdeen, which SEPA has previously used to investigate the NO_2 concentrations for six monitoring stations in the city, was presented. ADMS-Urban was also used to create a similar data set for the NO_2 concentrations across the city

of Glasgow with specific attention being paid to the eight monitoring stations' locations. Both simulated data sets were created using a Latin Hypercube (LHC) design and the locations in the hyperspace were chosen by varying the percentage change in emissions and wind speed, and the degree change in wind direction. The main difference between the two data sets is that for Glasgow there are more data available in terms of hourly measurements for emissions, temperature, wind speed and wind direction. Based on various numerical and graphical summaries, three monitoring stations in Aberdeen (Market Street 2, Union Street and Wellington Road) and one monitoring station in Glasgow (Central Station) were identified as "at risk" as both their actual NO_2 annual average concentration and the median for the simulation data is above 40 μg m^{-3}.

Exploratory analysis of the LHC space for both Aberdeen and Glasgow was done using variograms in order to establish whether there is hyperspatial correlations between scenarios present. From the variograms it was noted that some of the hyperspatial range parameters would not converge within the LHC space explored. This suggested that some of the hyperspatial parameters would be difficult to estimate due to the design of the LHC space, but confirms that there is correlation between the points within the LHC. Therefore, an emulation approach is reasonable to undertake.

9.2.1 Univariate hyperspatial modelling

Chapter 5 presented models for the NO_2 annual averages from the ADMS-Urban scenarios for each of the monitoring stations in both Aberdeen and Glasgow. Two types of modelling approaches were compared - linear regression and Gaussian Process (GP) models. The difference between the two approaches comes from the fact that the GP models were used to also account for the hyperspatial correlation between scenarios in the LHC using the **DiceKriging** package in R. Different kernels were tested to assess the level of smoothness required for the hyperspatial correlation. All models were compared in terms of out-of-sample prediction power based on a 10-fold cross validation (CV) and for both Aberdeen and Glasgow the GP models with exponential kernel were chosen as the best. The exponential kernel was the roughest hyperspatial correlation structure suggesting that the changes between different sets of inputs are very abrupt. It is interesting to note that for Glasgow, both the linear regression and GP models required more covariates than the models for Aberdeen, which is consistent with the larger size of Glasgow and hence, larger variety of pollution contributors. As expected from the exploratory analysis in Chapter 4, the GP models struggled with estimating the three hyperspatial range parameters in the models by estimating very large values for some of the monitoring stations. Therefore, an upper boundary limit of 1000 was set for all hyperspatial range parameters as it is a value much larger than the span of

any of the hyperspatial variables (emissions, wind speed and wind direction). In this chapter, a relatively simple model for each individual station was developed allowing the exploration of the risks of exceeding the annual average regulation value.

9.2.2 Multivariate hyperspatial modelling and emulation

The simulated data from ADMS-Urban are estimated not by using the actual spatial distances between locations but by a factor (combining information about the city outline, traffic, road and background emissions, meteorological data and chemistry) to vary the predictions at different locations at the city. Therefore, Chapter 6 extends the work from Chapter 5 by proposing a multivariate Bayesian GP model with an exponential kernel for the between scenario correlation and free form covariance matrix between stations. Hence, it is assessed whether modelling all monitoring stations in a city together would improve the prediction quality and reduce the number of models fitted. In order to assess the quality of the proposed multivariate Bayesian GP model to estimate the hyperspatial range parameters, a validation study was performed and it was found that the multivariate Bayesian model underestimates the hyperspatial range parameters as their values increase. However, a second validation study showed that even with "sensible" [50] estimates for the hyperspatial range parameters, the RMSPE remains relatively unaffected. Hence, the multivariate Bayesian GP model for the NO_2 annual average across multiple stations was applied to both Aberdeen and Glasgow but it was found that the NO_2 annual average individual station models from Chapter 5 perform very slightly better in terms of out-of-sample prediction error (based on a 10-fold CV). The results from the comparison of the multivariate Bayesian GP model to the single station frequentists GP models suggest that estimating a different set of hyperspatial parameters for each monitoring station slightly improves the prediction quality. However, a more detailed assessment with non-deterministic data is required to confirm these conclusions.

As using the multivariate model requires less models to fit, the multivariate GP models from Chapter 6 were used to emulate the ADMS-Urban NO_2 annual averages for untested scenarios (new sets of inputs for emissions, wind speed and wind direction) in order to identify emissions values and meteorological conditions for which compliance with the annual average regulation is achieved. In Aberdeen, it was found that for two of the monitoring stations (Anderson Drive and Errol Place), the NO_2 regulation will be breached under no conditions. Anderson Drive is away from the city centre, whereas Errol Place is an urban background station so it was expected that no breaches of the regulation will occur there. For one of the stations (King Street), some combinations of higher than the observed emissions in 2012 will result in breaches of the NO_2 regulation.

This result is logical given that the King Street monitoring station is away from the city centre but there is some heavy goods vehicles traffic nearby it. For three of the monitoring stations (Market Street 2, Union Street and Wellington Road), a reduction of at least 40% of the emissions from 2012 is required to ensure compliance. As the three monitoring stations are close to the harbour, they are directly affected by the traffic associated with the oil industry. Overall, the emulated NO_2 annual averages results were in agreement with the actual observed value in 2012 in Aberdeen.

In Glasgow, for three monitoring stations (Burgher Street, Townhead and Waulkmillglen Reservoir), the NO_2 regulation will never be breached, which is logical by the distance from the city centre and the background character of Townhead and Waulkmillglen Reservoir. For three other monitoring stations (Byres Road, Great Western Road and High Street) emissions must not be larger than the observed emissions in 2015 to ensure compliance. The grouping of these three stations is logical as they are close to the city centre and the west end of the city. For one of the monitoring stations (Dumbarton Road) in the west end but close to heavy goods vehicles traffic, a 5% reduction in the emissions from 2015 will result in no breaches of the NO_2 regulation. Lastly, for one of the monitoring stations (Central Station) a 60% reduction in the baseline emission will result in compliance for all meteorological conditions. The result is expected given that the monitoring station is located at a transport hub in the city centre. As with Aberdeen, the emulated simulated NO_2 annual averages results are in agreement with the actual observed values in 2015 in Glasgow.

9.2.3 Modelling and emulation of the NO_2 hourly exceedances over 200 μg m^{-3}

There is an hourly regulation for the NO_2 concentrations - no more than 18 breaches over 200 μg m^{-3}. Therefore, there is a need to create a model and use it to emulate the number of hourly NO_2 exceedances over 200 μg m^{-3} as simulated by ADMS-Urban. Given that count data are the response in this case, a Poisson generalised linear model (GLM) was used in Chapter 7. As hourly data are only available for Glasgow, where Central Station is the only location where any breaches above 200 μg m^{-3} are observed, that is the only monitoring station examined in that chapter. The exploratory plots showed that exceedances only occur for emissions larger than -60% from the baseline. Therefore, a Poisson GLM and a segmented Poisson GLM were compared based on their out-of-sample predictive performance (based on a 10-fold CV). However, the diagnostic plots for both models indicated under-dispersion issues and had to be refitted as quasi-Poisson models. The quasi-Poisson GLM had better out-of-sample performance and its diagnostic plots indicated a better fit.

Hence, the quasi-Poisson GLM was used to emulate the number of NO_2 exceedances for untested scenarios (new sets of inputs for emissions, wind speed and wind direction). It was found that for all conditions when the emissions are 60% less than the baseline recorded in 2015, there will be no occurrences over 200 μg m^{-3}, which is consistent with the findings from the NO_2 annual average emulation in Chapter 6. Additionally, it was found that the more western prevailing the wind, the more exceedances over 200 μg m^{-3} will occur. When the baseline conditions were emulated, it was estimated that there were 17 exceedances expected to occur. However, in reality only 4 have occurred. Based on these results, it can be concluded that the emulator predicts quite well the Central Station Glasgow ADMS-Urban scenarios results but the scenarios over-predict in comparison with what has actually occurred in reality.

9.2.4 Hyperspatial-temporal modelling and emulation

Chapter 8 presented the biggest statistical challenge in terms of modelling and prediction of high resolution air pollution data - the modelling and prediction of the yearly time series of NO_2 hourly concentrations as simulated by ADMS-Urban. The chapter builds on the models from Chapters 5 and 6 by including a temporal random effect in addition to the hyperspatial random effects. Continuing from Chapter 7, only the Central Station monitoring station in Glasgow was modelled. In the chapter, a new hyperspatial-temporal model with a block-diagonal design matrix is proposed to allow for different covariates to be used to model the different scenarios at a single monitoring location. The model uses a separable (in terms of hyperspace for the between scenarios correlation and within scenario time correlation) variance-covariance matrix. The separability allowed for some computational efficiency techniques to be used. However, due to the size of the variance-covariance matrix (6TB), the model was not applied to the full data set of yearly time series of NO_2 hourly concentrations as it is expected that it would take almost nine months to run the model. This is contrary to the idea of using an emulator to provide a faster alternative to running the actual simulated model. Instead, an alternative approach of modelling overlapping blocks of data for periods of interest is adopted. After examining the full time series over a year from the simulation scenarios, it was identified that the period from 17/01 to 7/02 is of most interest as all NO_2 hourly exceedances above 200 μg m^{-3} occurred in the period in 2015.

In order to test the idea of overlapping blocks of data, two models were fit - a January model (17/01 to 29/01) and a February model (28/01 to 7/02). In that way, in both periods of the simulation scenarios breaches over 200 μg m^{-3} are observed. Both models had hourly temperature, hourly wind speed, hourly wind direction (as a circular variable), an interaction between temperature and wind speed, and emissions as covariates.

Based on out-of-sample predictive power using 10-fold CV, the two models were found to perform well when emulating the ADMS-Urban scenarios out-of-sample.

It has to be noted, that when performing the 10-fold CV, one of the scenarios is in effect outside of the LHC space defined the 90 scenarios used to fit the model. This results in an extrapolation and affects some of the diagnostic plots. The January model slightly over-predicted the close-to-zero NO_2 hourly concentrations for one of the scenarios, whereas the February model under-predicted some NO_2 hourly concentrations for a few of the scenarios. Three scenarios were chosen based on their locations in the LHC and emulated. It was found that for two of the scenarios, the overlapping periods match perfectly onto each other and almost all emulated NO_2 hourly concentrations are superimposed on the true ADMS-Urban simulated NO_2 hourly concentrations. The third scenario is the extrapolated one and although both models under-predicted the absolute values of the NO_2 hourly concentrations, both models have captured the overall underlying trend. Hence, the January and February models mimicked very closely the ADMS-Urban simulations for the NO_2 hourly concentrations out-of-sample.

Hence, the January and February models were used to emulate the NO_2 hourly concentrations at the baseline. The emulated values were then compared to the true NO_2 concentrations from 2015. The January model estimated that there will be a period of almost four days, where the hourly occurrences will be over 200 μg m^{-3} and up to almost 600 μg m^{-3}. This would have resulted in a breach of the hourly regulation but in fact no breaches actually occurred. The overestimation of reality by ADMS-Urban and its emulated values are heavily impacted by the fact that all days of the week are treated as equal instead of adjusting for weekday vs. weekend activities. The February model performed better than the January model and correctly estimated one breach at the same time as it had actually occurred, but under-estimated the magnitude of the breach by 40 μg m^{-3}. A 48-hour overlap between the two models was estimated and the hourly concentrations between the two models were almost identical with the exception of the final hourly concentration from January being a breach, which does not occur in either real-life or the emulation based on the February model. Overall, it was found that the hyperspatial-temporal model does well in emulating the Central Station Glasgow ADMS-Urban concentrations. The comparison to the actually observed NO_2 hourly data found that for Central Station Glasgow, the ADMS-Urban is expected to fail to identify the moments when most breaches over 200 μg m^{-3} are actually observed, and their magnitudes, but overall the underlying trend is captured.

9.3 Discussion and future work

In this thesis, the issues in the sparsity of air pollution data were studied by examining the quality of MA sensors and emulating ADMS-Urban. Different modelling techniques were developed (such as the hyperspatial-temporal model in Chapter 8) and applied. However, there are some limitations to the work presented in this thesis. This last section of the conclusion aims to discuss these limitations and ways they can be overcome with future work.

The main limitation in Chapter 3 was the quality and amount of the data. Firstly, only one location was used to place the MAS, which does not provide a sufficient spatial representativity for locations with more urban (e.g. kerbside) or rural background characteristics. As the objective of this study was to evaluate an exemplary application within a citizen science scenario, this lack of spatial representation is acceptable. However, to derive more general findings, a wider application in a more diverse pollution context is advised. Furthermore, the experiment only lasted 20 days. Running the experiment for a longer period would allow exposure to more diverse meteorological conditions and to draw more robust conclusions with regard to the covariates. The MAS' data contained some unusual observations which require further analysis as well as investigating whether there are specific conditions in which the sensors struggle to operate reliably. Last but not least, temperature and relative humidity were not ratified against reference instruments for this deployment but previous field tests with the same packages indicate robust performance with regard to the trends. The values used were averaged from the MAS' measurements for temperature and relative humidity. Once more robust results from the MAS are available as well as ratified data for relative humidity and temperature, it would be beneficial to use an inverse regression model and produce estimates for the true pollutant concentration based on the measurements from the sensors.

A challenge in the modelling and emulation in Chapters 5, 6 and 8 was that hyperspatial range parameters are hard to estimate. The issue was first noted in Chapter 4, when variograms were used for initial impression for the correlation in the LHC space for both Aberdeen and Glasgow. The issue continued through Chapters 5 and 6. The issue can be overcome with larger sets of data as seen in the model testing in Appendix D. Additionally, in Chapters 6 and 8, the hyperspatial range parameters are estimated using the exponential function based on the analysis from Chapter 5, where it was found that the surface of the area of interest is very rough. However, it would be of value in future works to test the predictive performance of other hyperspatial correlation functions when applying the proposed models from Chapters 6 and 8.

Another important point is the choice of the LHC design for choosing the ADMS-Urban simulation scenario runs. One possible alternative is to use grid ensemble design for multivariate emulation as proposed in [133]. Alternatively, a design combining LHC and Sobol indices can be applied as suggested in [53]. It would also be of interest to assess how much effect does the choice of simulation runs have on quality of the emulated results.

Furthermore, the variance-covariance matrices used for both models in Chapters 6 and 8 are defined as separable as there was no reason to believe there is a trade-off between two parts of the variance-covariance matrices. The separability also allowed for computational efficiency in the estimation of the hyperspatial-temporal model from Chapter 8. Nevertheless, it would be of interest to explore whether there is in fact a trade-off between two parts of the variance-covariance matrices, especially between the hyperspatial and temporal correlations in Chapter 8. This could lead to an improvement in the predictions.

Additionally, the multivariate Bayesian GP emulator in Chapter 6 uses a free form correlation matrix to model the between stations correlation. However, it would be beneficial to explore more complex correlation structures which would take into account the geographical locations although they are not explicitly specified in ADMS-Urban. This would allow the emulation to be applied to a more regularised grid across a city which in turn would result in better understanding of the air pollution and its movement across a city.

The modelling and emulation in Chapter 7 could be further explored by using a spatial Poisson model, which would allow for more accurate prediction of the number of NO_2 hourly concentrations which breach the 200 μg m^{-3} in a year. For instance, [130] proposes a spatial Poisson regression for modelling spatial autocorrelation of geographical observations which can be extended to work in hyperspace, while [101] introduces a multivariate spatial Poisson model which would allow modelling the number of exceedances for all stations within a city.

A future work of interest would be to improve the emulation of hourly concentrations by using the results from the Chapter 7 quasi-Poisson GLM for the number of breaches over 200 μg m^{-3} in a year to calibrate the number of breaches when using the hyperspatial-temporal model from Chapter 8. Alternatively, exceedances over thresholds analysis can be applied to the ADMS-Urban simulations for more robust modelling of the high NO_2 hourly concentrations above 200 μg m^{-3}.

The emulator presented in Chapter 8 requires further improvements to make it computationally faster. It would be of interest to explore block-design computation using

cloud sharing as discussed in [178]. One way to try and overcome the issue would be to use an alternative language such as C++ through the R package Rccp [67] or Python [196] as suggested in [103] and [9] respectively. Furthermore, sparse GPs can be used by inducing variables and thus, reducing the time complexity as suggested in [190]. Alternatively, local GPs can be applied as suggested in [96] and compared in performance to the sparse GPs.

Overall, the thesis aimed to address the sparsity of air pollution data by assessing lower cost alternatives to the AURN sensors and emulating simulated data to provide further information about air pollution to the observed data. Although the results from the linear regression on the MAS' data showed that the sensors record fluctuations in temperature and relative humidity, the main driving force was the pollutant concentrations indicating that the measurements from lower cost sensors have their merits. The emulation of the ADMS-Urban data showed that the NO_2 annual averages are almost perfectly emulated. The emulated results from using ADMS-Urban simulation data showed that when NO_2 annual averages are being emulated, they are not only close to mimicking the simulation values but also the true observed values in both the single station hyperspatial and multi-station hyperspatial modelling. Moreover, the modelling of count data in terms of the number of hourly concentrations above a regulatory limit using a quasi-Poisson GLM had an very accurate emulation for the ADMS-Urban scenarios. When the hyperspatial-temporal emulator was applied to the NO_2 hourly data, the emulated values were almost perfectly superimposed with those produced by the ADMS-Urban simulations. Generally, the emulators created in this thesis have performance very close to the simulated data, with very small standard deviation estimates and allow for more air pollution data for unobserved emissions levels and meteorological conditions to be produced.

Appendix A

Matrix distributions

In Chapter 6, a Bayesian multivariate emulator is applied to the ADMS-Urban simulation runs for Aberdeen and Glasgow to model the NO_2 annual average across all monitoring stations in both cities. The emulator is built using the matrix Normal distribution and a matrix t-distribution, which are described below.

A.1 Matrix Normal Distribution

The matrix Normal Distribution was introduced in [55]. Let the matrix $\mathbf{A}$ $(m \times n)$ follow the matrix Normal distribution, then:

$$\mathbf{A} \sim \mathbf{MN}(\mathbf{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}), \tag{A.1}$$

where:

- $\mathbf{M}$ is the mean matrix $(m \times n)$;
- $\boldsymbol{\Sigma}$ is a row-scaling positive definite matrix $(m \times m)$; and
- $\boldsymbol{\Omega}$ is a column-scaling positive definite matrix $(n \times n)$.

The probability density function for $\mathbf{A}$ is:

$$p(\mathbf{A}|\mathbf{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Omega}^{-1}(\mathbf{A}-\mathbf{M})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{A}-\mathbf{M})\right]\right)}{(2\pi)^{\frac{nm}{2}}|\boldsymbol{\Omega}|^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}}, \tag{A.2}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

It is important to note that a Matrix Normal distribution can be re-written as a n-variate-Normal distribution by stacking the columns of $\mathbf{A}$ in a vector as:

$$\text{vec}(\mathbf{A}) \sim \mathbf{N}(\text{vec}(\mathbf{M}), \mathbf{\Omega} \otimes \mathbf{\Sigma})\,. \tag{A.3}$$

A.2 Matrix t-Distribution

The matrix t-Distribution was introduced in [106]. Let the matrix $\mathbf{A}$ $(m \times n)$ follow the matrix t-distribution, then:

$$\mathbf{A} \sim \mathbf{MT}(\mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}, \nu)\,, \tag{A.4}$$

where:

- $\mathbf{M}$ is the mean matrix $(m \times n)$;
- $\mathbf{\Sigma}$ is a row-scaling positive definite matrix $(m \times m)$;
- $\mathbf{\Omega}$ is a column-scaling positive definite matrix $(n \times n)$; and
- ν are the degrees of freedom.

Then, the probability density function for $\mathbf{A}$ is:

$$p(\mathbf{A}|\mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}, \nu) = \frac{\Gamma_n\left(\frac{\nu+m+n-1}{2}\right)}{\left(\pi^{\frac{mn}{2}}\right)\Gamma_n\left(\frac{\nu+n-1}{2}\right)} |\mathbf{\Omega}|^{-\frac{m}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} |\mathbb{I}_m - \mathbf{\Sigma}^{-1}(\mathbf{A}-\mathbf{M})\mathbf{\Omega}^{-1}(\mathbf{A}-\mathbf{M})^\top|^{\frac{\nu+m+n-1}{2}}\,, \tag{A.5}$$

where $\Gamma_n(\cdot)$ is a multivariate gamma function defined in [104] as:

$$\Gamma_n(\alpha) = \pi^{\frac{1}{4}n(n-1)} \sum_{i=1}^{n} \Gamma\left(\alpha - \frac{1}{2}(i-1)\right)\,. \tag{A.6}$$

Similarly to the matrix Normal distribution, the matrix t-distribution can be re-written as a multivariate t-distribution by stacking the columns of $\mathbf{A}$ into a vector:

$$\text{vec}(\mathbf{A}) \sim \mathbf{T}(\text{vec}(\mathbf{M}), \mathbf{\Sigma} \otimes \mathbf{\Omega}, \nu)\,. \tag{A.7}$$

Appendix B

Exploring the emissions and meteorological effect on the NO_2 hourly concentrations in Glasgow in 2015

Modelled hourly emissions and meteorological data for each of the eight monitoring stations in Glasgow for 2015 is available. The emissions data is a 24-hour cycle of emissions (in g m^{-2} h). The meteorological data set used consists of the hourly measurements for temperature (in degrees °C), wind speed (in m/s) and wind direction (in degrees °). The data were downloaded from the Air Quality in Scotland website (http://www.scottishairquality.co.uk/) on 16/11/2016. Exploratory analysis of each of these meteorological conditions and emissions will be presented in this subsection as well as checks for any trends in the hourly observations of the meteorological data themselves as well as with the NO_2 hourly concentrations. The log NO_2 hourly concentrations will be used as they appear normally distributed as seen in Figure 4.15.

B.1 Temperature

Since there is modelled hourly data for the temperatures (will be referred to simply as temperature onwards) at each of the stations. The time series for the temperatures at each of the stations are presented in Figure B.1. Overall, the temperatures for all stations appear identical in shape - with low concentrations on both ends of the time series (winter) and a peak in the middle (summer). It is interesting to note that all

stations except for High Street are missing 192 hours, which is equal to 8 days. The missing days are 12/02, 11/03, 27-30/03, 8/4, 1/12. When the hourly observations from a single day were missing, the data were imputed by averaging the neighbouring days temperatures for the same hour, and assigning it to the missing observation. When the hourly observations from multiple days were missing, the data were imputed by averaging the neighbouring weeks temperatures for the same day and hour, and assigning it to the missing observation.

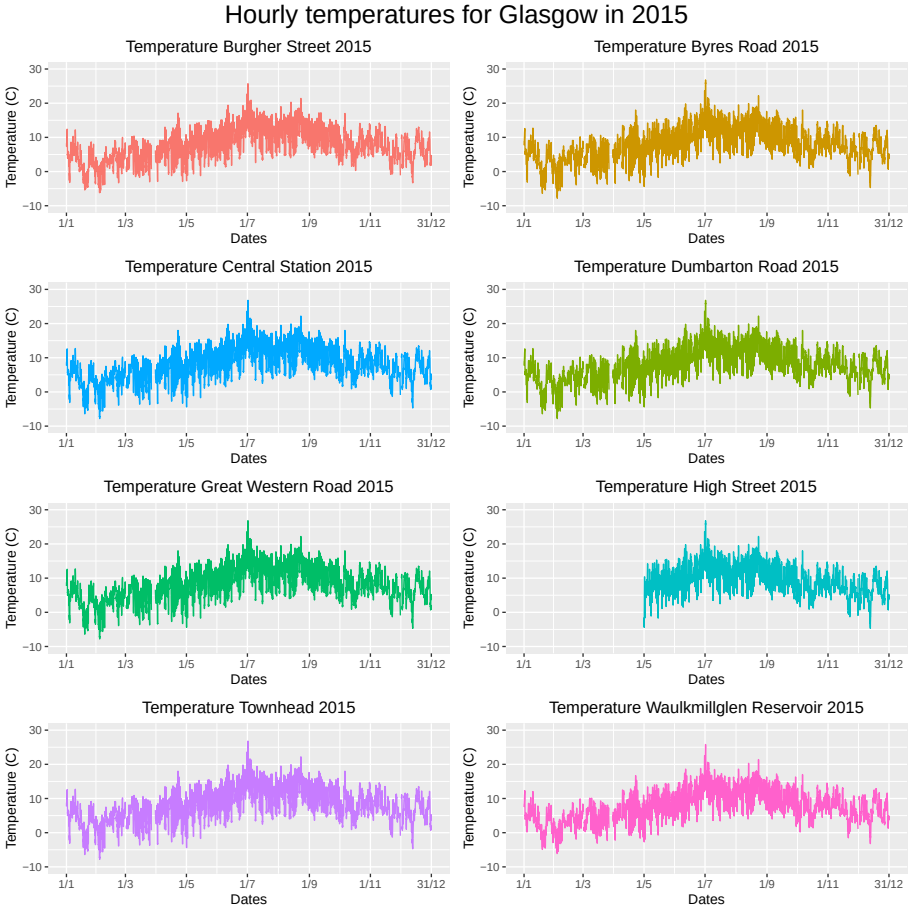

FIGURE B.1: Time series plot for the hourly temperatures (°C) for each of the eight monitoring stations in Glasgow in 2015.

However, when comparing the magnitudes, the stations can be split into three groups. Byres Road, Central Station, Dumbarton Road, Great Western Road, High Street and Townhead have identical temperature values, whereas Burgher Street has slightly lower temperatures than this group of six stations and Waulkmillglen Reservoir has even lower temperatures than Burgher Street. This is not surprising given the fact that the group of six monitoring stations are located within the city centre, whereas Burgher Street is in the east end of the city and Waulkmillglen Reservoir is outside the city as previously seen in Figure 4.12. As High Street has identical values as five other stations (Byres Road, Central Station, Dumbarton Road, Great Western Road and Townhead), the

missing values for the first four months of the year were imputed as the same values recorded at the other five stations. It is interesting to note that the missing values for temperatures are not the same as the ones missing in Figure 4.13 but the station with the most missing values in both cases is High Street.

In Figure B.2, time plots comparing the log NO_2 hourly concentrations and the hourly temperatures for all the monitoring stations in Glasgow in 2015 are presented. It appears that when there are low temperatures, the log NO_2 hourly concentrations increase. This trend is most visible when focusing on the temperature inversion in late January, when the temperatures are at their lowest values, whereas the log NO_2 concentrations are the highest. Similarly, in the the beginning of July, the temperatures are the highest and this results in the lowest log NO_2 concentrations. These conclusions are expected as high temperatures is a catalyst in chemical reactions and causes faster reactions between the NO_2 and other pollutants.

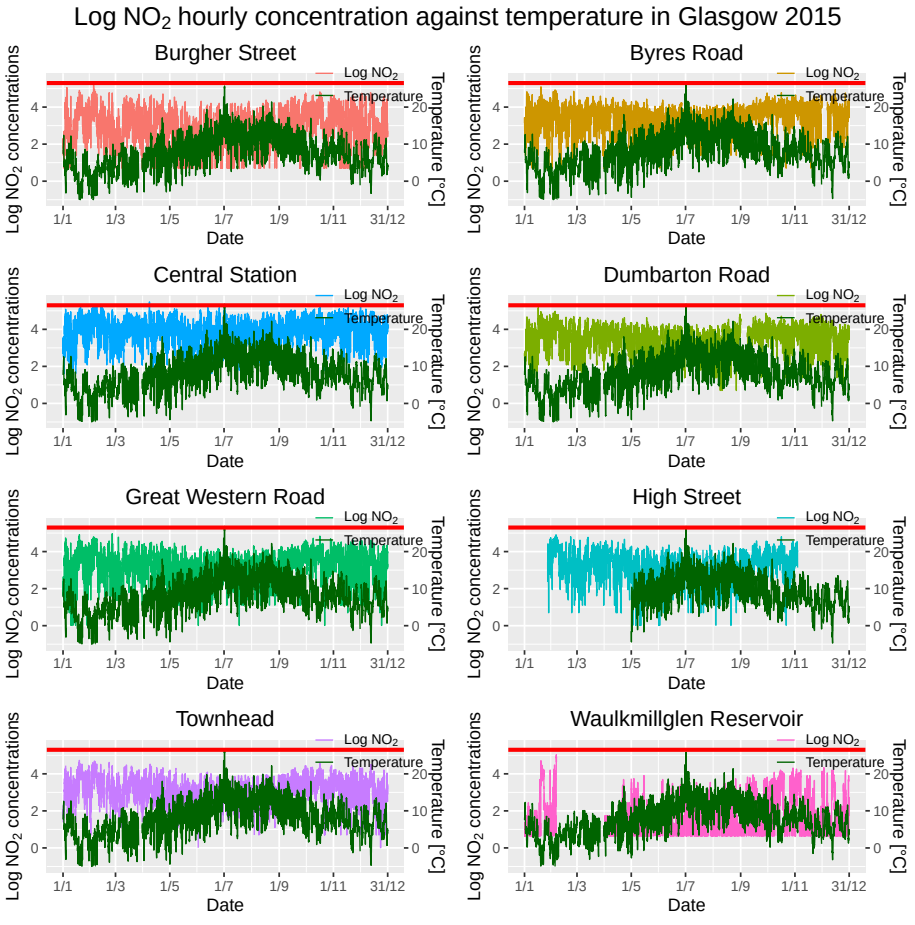

FIGURE B.2: Joint time series plot of the hourly NO_2 concentrations (μg m^{-3}) and the hourly temperatures (°C) for each of the eight monitoring stations in Glasgow in 2015. The hourly limit of log 200 μg m^{-3} is represented by the red line.

Lastly, scatterplots for the log NO_2 hourly concentrations and the hourly temperatures for each of the eight monitoring stations in Glasgow in 2015 are examined in Figure B.3.

Station	**Pearson's correlation** (95% CI)
Burgher Street	-0.27 (-0.29, -0.25)
Byres Road	-0.06 (-0.09, -0.04)
Central Station	-0.05 (-0.07, -0.03)
Dumbarton Road	-0.15 (-0.17, -0.13)
Great Western Road	-0.18 (-0.20, -0.16)
High Street	-0.09 (-0.12, -0.06)
Townhead	-0.38 (-0.40, -0.36)
Waulkmillglen Reservoir	-0.19 (-0.21, -0.16)

TABLE B.1: The Pearson's correlation coefficients and their corresponding 95% intervals for the log hourly NO_2 concentrations (μg m^{-3}) and temperatures (°C) across the eight monitoring stations in Glasgow in 2015.

For all stations, there is a weak linear negative correlation. This is further investigated in Table B.1 where the 95% CIs for the Pearson's correlation are presented. The table confirms that there is a significant negative correlation between the log NO_2 concentrations and the temperatures at all eight monitoring stations in Glasgow in 2015 as all the 95% CIs are entirely negative and do not contain zero. It has to be noted that the scatterplots contain a cloud containing the majority of the points and below the cloud for almost all stations, there appear to be lines of points. These lines are due to the aforementioned rounding to whole numbers of the NO_2 hourly observations.

B.2 Wind speed

There is modelled hourly data for the wind speed (will be referred to as wind speed onwards) for each of the stations. Firstly the time series for the wind speeds at each of the stations are presented in Figure B.4. Overall, the wind speeds for all stations appear very similar in shape. There is no clear trend when low or high wind speeds appear as opposed to temperatures. Similarly to the temperatures, for all stations except for High Street there are 192 hours missing, which is equal to 8 days. The missing days are the same as for temperature. Therefore, the imputations were performed in the same way as for temperature.

As opposed to the temperatures, the plots in Figure B.4 clearly show the three groups by which the stations can be grouped when taking into account the actual magnitudes for the wind speeds. The stations can be split into the same three groups as by temperatures.

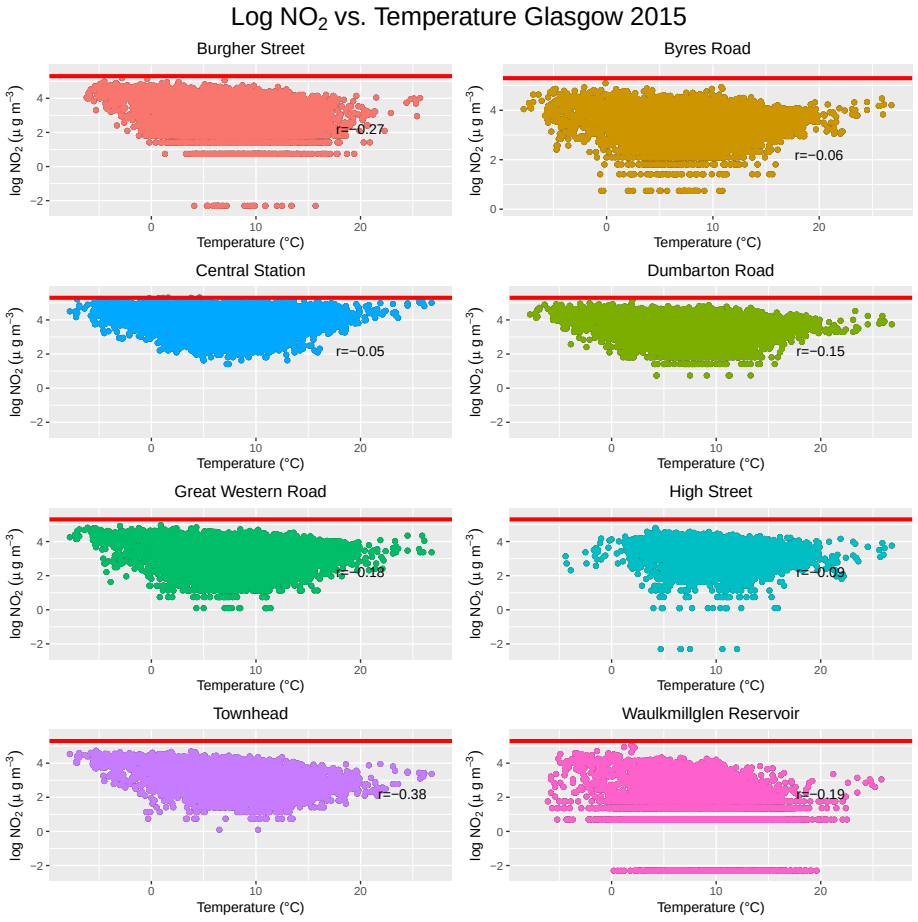

FIGURE B.3: Scatterplot of the hourly NO_2 concentrations (μg m^{-3}) and the hourly temperatures (°C) for each of the eight monitoring stations in Glasgow in 2015. The hourly limit of log 200 μg m^{-3} is represented by the red line. The correlations for each pairing are also provided.

Byres Road, Central Station, Dumbarton Road, Great Western Road, High Street and Townhead have identical wind speed values, whereas Burgher Street has slightly higher wind speed than this group of six stations and Waulkmillglen Reservoir has higher wind speed than Burgher Street. This is not surprising given the geographical location of the monitoring stations as seen in Figure 4.12. As High Street has identical values as five other stations (Byres Road, Central Station, Dumbarton Road, Great Western Road and Townhead), the missing values for the first four months of the year were imputed as the same values recorded at the other five stations.

Next, time plots comparing the time series of the log NO_2 hourly concentrations and the hourly wind speeds for all the monitoring stations in Glasgow in 2015 are compared in Figure B.5. Initially, it appears that when there are low wind speeds, the log NO_2 hourly concentrations increase. This trend is most visible when focusing on the temperature inversion in late January, when the temperatures and wind speed are at their lowest values, whereas the log NO_2 concentrations are the highest. However, in the the beginning

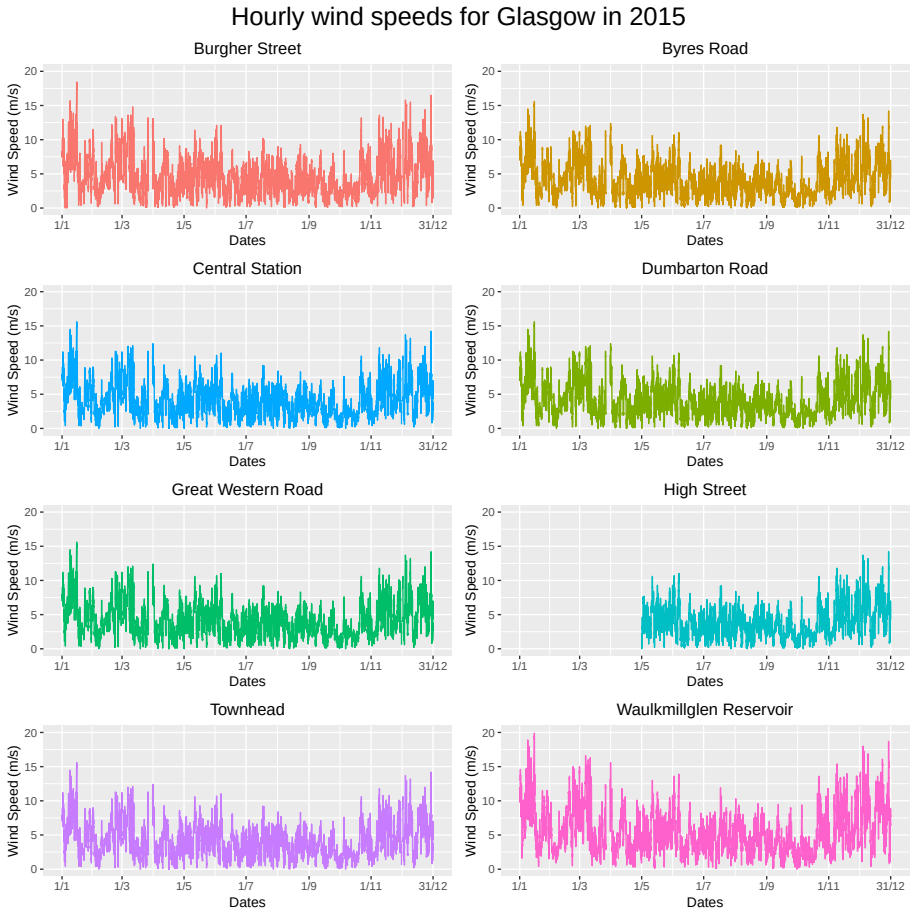

FIGURE B.4: Time series plot for the time series of the hourly wind speed (m/s) for each of the eight monitoring stations in Glasgow in 2015.

of July, the wind speeds are also low but this does not cause the log NO_2 concentrations to increase. These implies that wind speed's effect on the NO_2 concentrations has a seasonal effect.

Finally, the scatterplots for the log NO_2 hourly concentrations and the hourly wind speeds for each of the eight monitoring stations in Glasgow in 2015 are presented in Figure B.6. For all stations, there is a moderate linear negative relationship. This is further examined in Table B.2 where the 95% CIs for the Pearson's correlation are shown. As all the 95% CIs are entirely negative and do not contain zero, it is concluded that there is a significant negative correlation between the log NO_2 concentrations and the wind speeds at all eight monitoring stations in Glasgow in 2015.

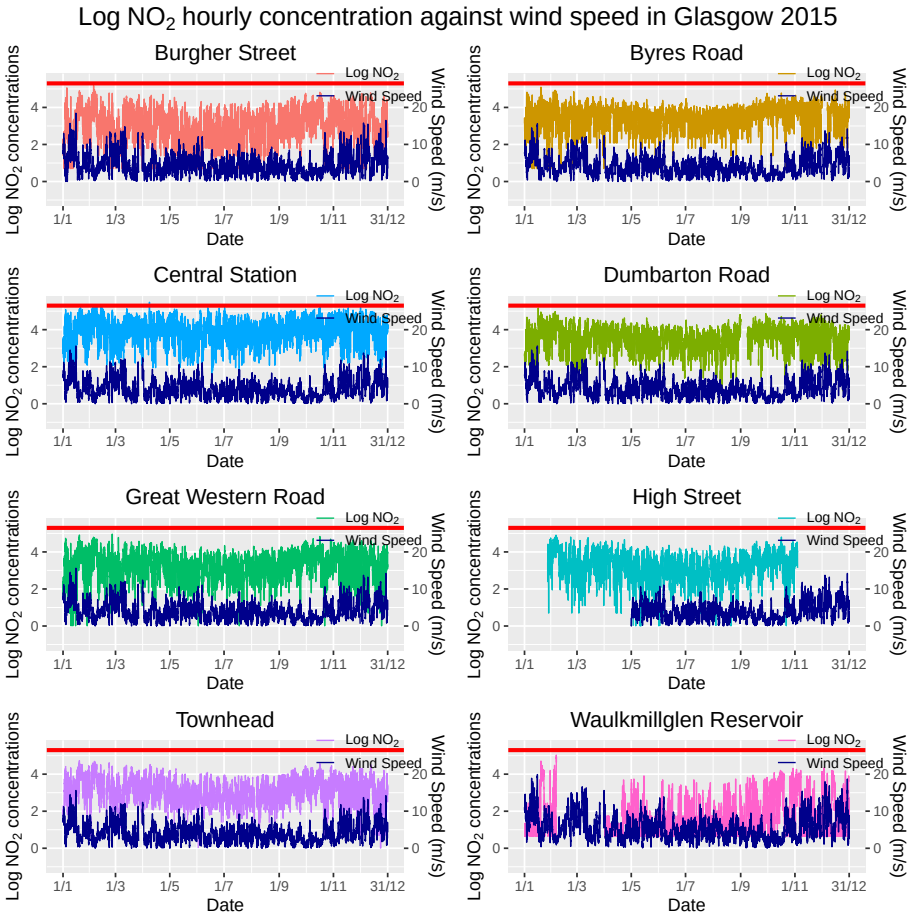

FIGURE B.5: Joint time series plot of the hourly NO_2 concentrations (μg m^{-3}) and the hourly wind speed (m/s) for each of the eight monitoring stations in Glasgow in 2015. The hourly limit of log 200 μg m^{-3} is represented by the red line.

Station	**Pearson's correlation** (95% CI)
Burgher Street	-0.43 (-0.44, -0.41)
Byres Road	-0.30 (-0.32, -0.28)
Central Station	-0.32 (-0.34, -0.30)
Dumbarton Road	-0.39 (-0.41, -0.37)
Great Western Road	-0.45 (-0.47, -0.43)
High Street	-0.39 (-0.41, -0.36)
Townhead	-0.46 (-0.47, -0.44)
Waulkmillglen Reservoir	-0.22 (-0.24, -0.20)

TABLE B.2: The Pearson's correlation coefficients and their corresponding 95% intervals for the log hourly NO_2 concentrations (μg m^{-3}) and wind speeds (m/s) across the eight monitoring stations in Glasgow in 2015.

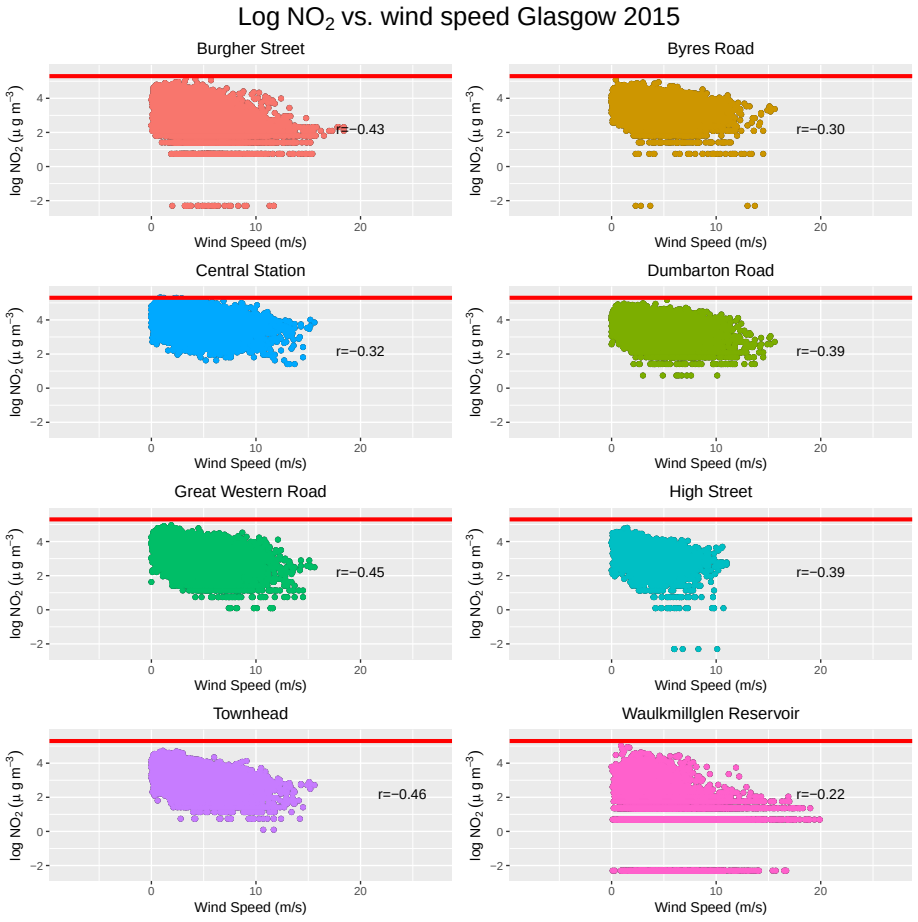

FIGURE B.6: Scatterplot of the hourly NO_2 concentrations (μg m^{-3}) and the hourly wind speeds (m/s) for each of the eight monitoring stations in Glasgow in 2015. The hourly limit of log 200 μg m^{-3} is represented by the red line. The correlations for each pairing are also provided.

B.3 Wind direction

Lastly, the modelled hourly data for wind direction (will be referred to as wind direction onwards) at each of the stations in Glasgow in 2015 are examined using histograms in Figure B.7. The histograms for all stations appear identical in shape with the exception of Burgher Street and Waulkmillglen Reservoir. However, for all stations, the majority of recorded wind directions have angles of about 250° making the predominant wind south-western. There is a second smaller peak around 100° indicating an east-southeastern wind. As with temperature and wind speed, there are 8 days of missing data for all stations except for High Street. Imputations were performed in a similar way to the ones for temperature and wind speed.

The differences in shape and occurrence frequencies for Burgher Street and Waulkmillglen Reservoir are due to the different geographical location of the two stations in comparison to the others and is expected. For Burgher Street, the difference comes

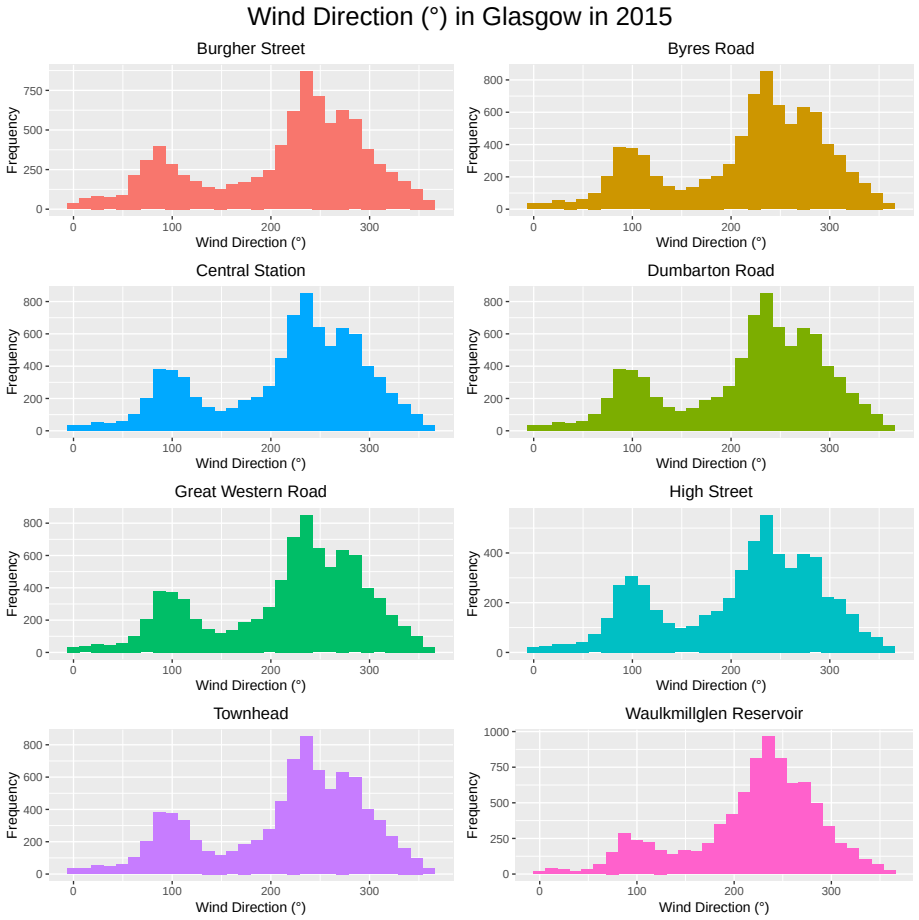

FIGURE B.7: Histograms for the modelled wind direction (°) for 2015 for eight of the monitoring stations in Glasgow in 2015.

from the wind directions for less than 200°, where there is more variation in the observed degrees of freedom. The wind direction at Burgher Street is very similar to the six stations with identical observations with a predominant southwestern wind direction and a small peak at east-southeastern direction. At Waulkmillglen Reservoir, the whole histogram has a different shape with the larger peak at about 250° having higher occurrence frequency (of almost 1000), whereas the bars for the second peak at about 100° are shorter in comparison to the second peak for the other stations. This suggests that at Waulkmillglen Reservoir, the wind has mostly been in the southwestern direction. It is not surprising that the wind direction at Waulkmillglen Reservoir is most different given that the monitoring station is outside the city.

To visualise the relationship between the hourly NO_2 concentrations and the wind direction, pollution roses (as described in Subsection 8.1.1) are used. The pollution roses for the eight monitoring stations in Glasgow are presented in Figures B.8 and B.9.

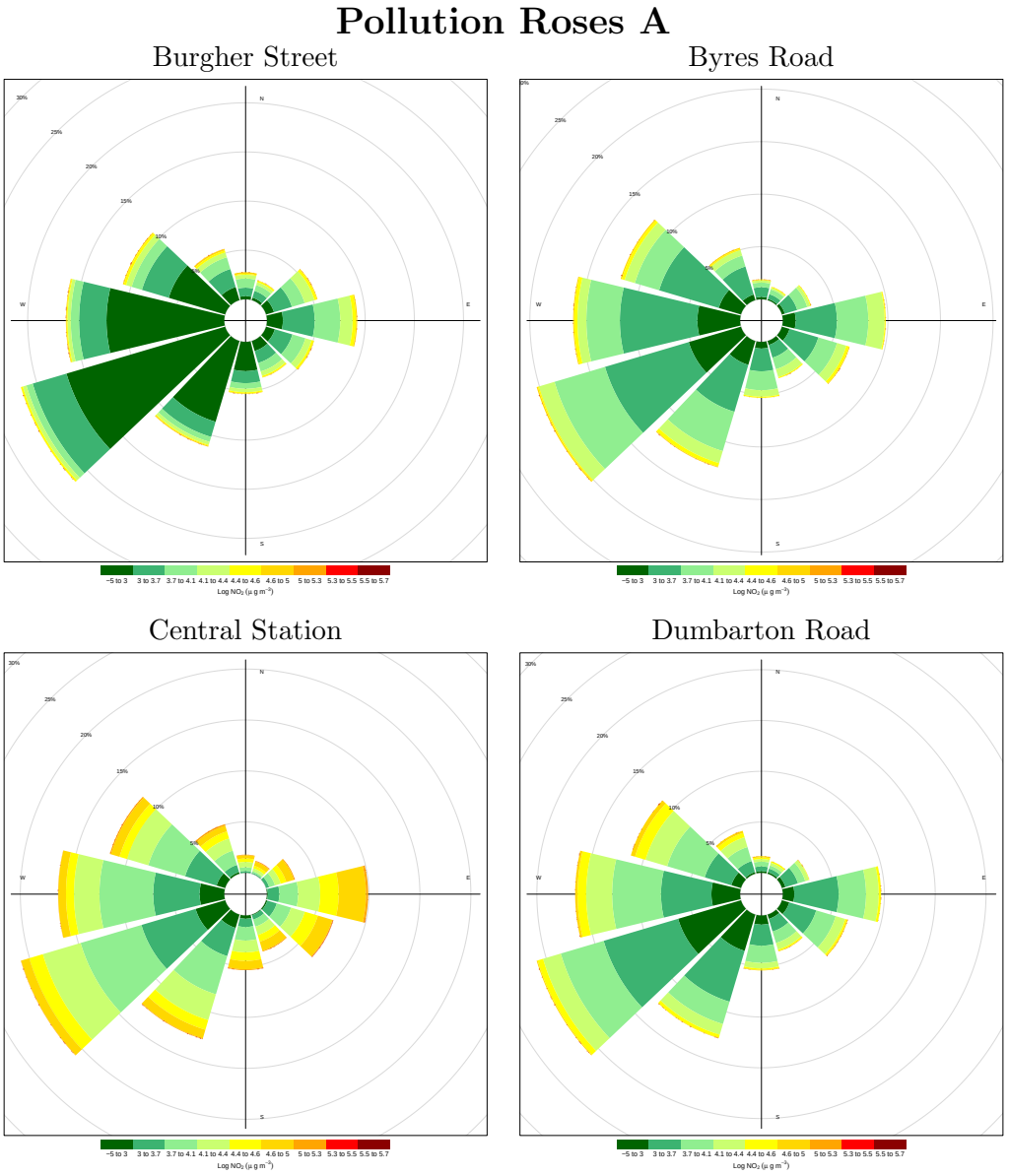

FIGURE B.8: Pollution roses for the monitoring stations at Burgher Street, Byres Road, Central Station and Dumbarton Road. The corresponding log NO_2 concentration (μg m^{-3}) in 2015 are proportionally ordered to the wind direction angle (°) at which the concentrations are recorded.

For all stations, there is a clear trend for western to southern prevailing winds, low NO_2 concentrations (below 4.1, which is equivalent to 60 μg m^{-3}) are predominant, whereas for all other directions, the values up to 4.6 (100 μg m^{-3}) appear almost equally. The only exception is Waulkmillglen Reservoir, which as the background monitoring station has almost all recordings below 3 (20 μg m^{-3}). Overall, there is an impression that during the predominant southeastern winds, the log NO_2 concentrations are quite low but as the wind becomes more eastern, the log NO_2 concentrations are increased and yellow and orange colours are present.

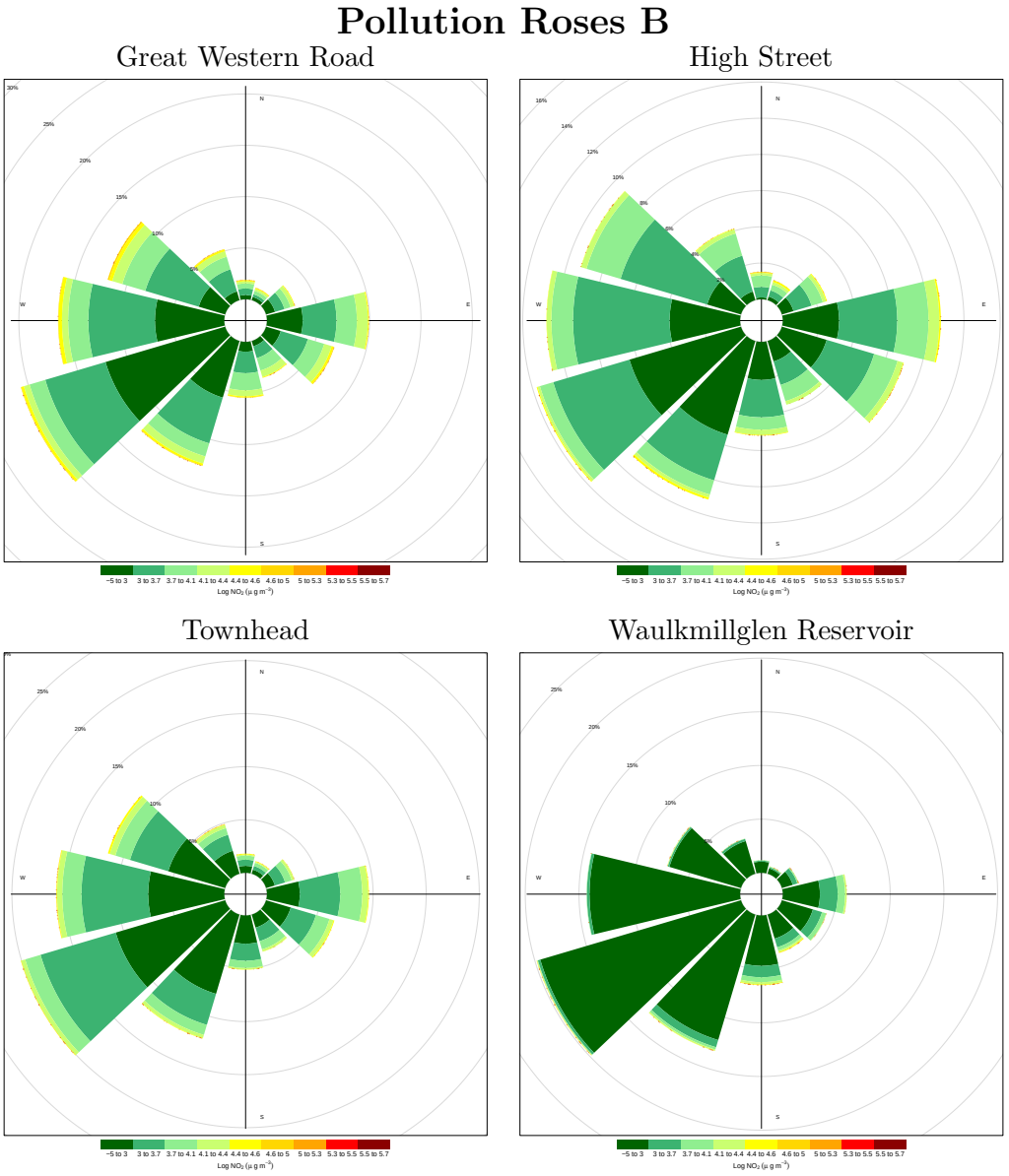

FIGURE B.9: Pollution roses for the monitoring stations at Great Western Road, High Street, Townhead and Waulkmillglen Reservoir. The corresponding log NO_2 concentration (μg m^{-3}) are ordered in proportion to the wind direction angle (°) at which the concentrations are recorded.

B.4 Emissions

As previously mentioned, there is an 24-hour emissions cycle based on the data collected of one day of the year based on background emissions and an average flow of the vehicles. The emissions are visualised in Figure B.10. In order to explore the relationship between 24:00 and 01:00, 01:00 is plotted twice. On the plot, there is a clear distinction between the night and day hours with the day hours having emissions that are more than 3 times higher. The highest emissions are at 19:00, although there is a smaller peak at 10:00. These two peaks are at the end of the rush hour when people get to and from work.

There is almost a plateau of high emissions between 10:00 and 19:00 indicating that throughout the day the emissions are high. The lowest emissions are at 04:00 and 05:00 in the morning.

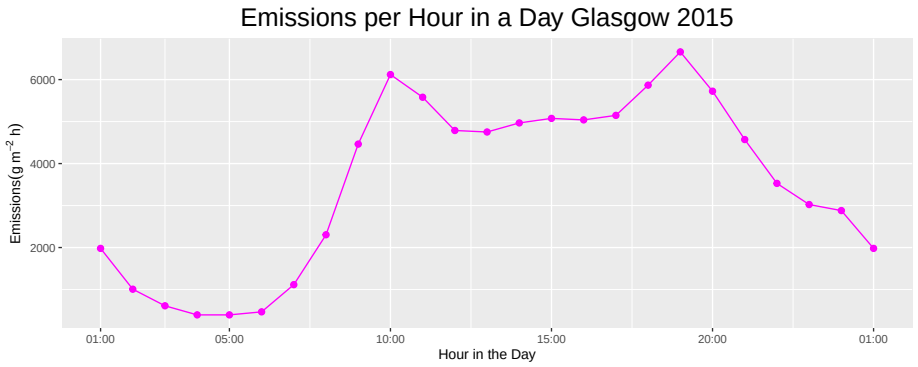

FIGURE B.10: Scatterplot for the emissions (g m^{-2} h) in Glasgow in 2015.

Next, boxplots for the log NO_2 concentrations for each hour are shown in Figure B.11. The boxplots are ordered by emissions going from smallest to largest from left to right. Since for 04:00 and 05:00 the same emissions are estiated, the two hours have been combined in one box. There is a weak to moderate positive linear correlation for all eight monitoring stations which indicates that the larger emissions, the higher the log NO_2 concentrations. This is further confirmed by Table B.3 where the Pearson' correlation coefficients (see Subsection 2.1.2) and their respective 95% CIs are entirely positive. The boxplots in Figure B.11 highlight that the monitoring station at Waulkmillglen Reservoir does not record the log NO_2 concentrations at 02:00. This is due to the nature of the station as a background monitoring location and requires daily calibration.

Station	**Pearson's correlation** (95% CI)
Burgher Street	0.22 (0.20, 0.24)
Byres Road	0.48 (0.46, 0.49)
Central Station	0.46 (0.44, 0.48)
Dumbarton Road	0.48 (0.46, 0.50)
Great Western Road	0.35 (0.33, 0.37)
High Street	0.33 (0.30, 0.35)
Townhead	0.17 (0.15, 0.19)
Waulkmillglen Reservoir	0.08 (0.06, 0.10)

TABLE B.3: The Pearson's correlation coefficients and their corresponding 95% confidence intervals for the log hourly NO_2 concentrations (μg m^{-3}) and emissions (g m^{-2} h) across the eight monitoring stations in Glasgow in 2015.

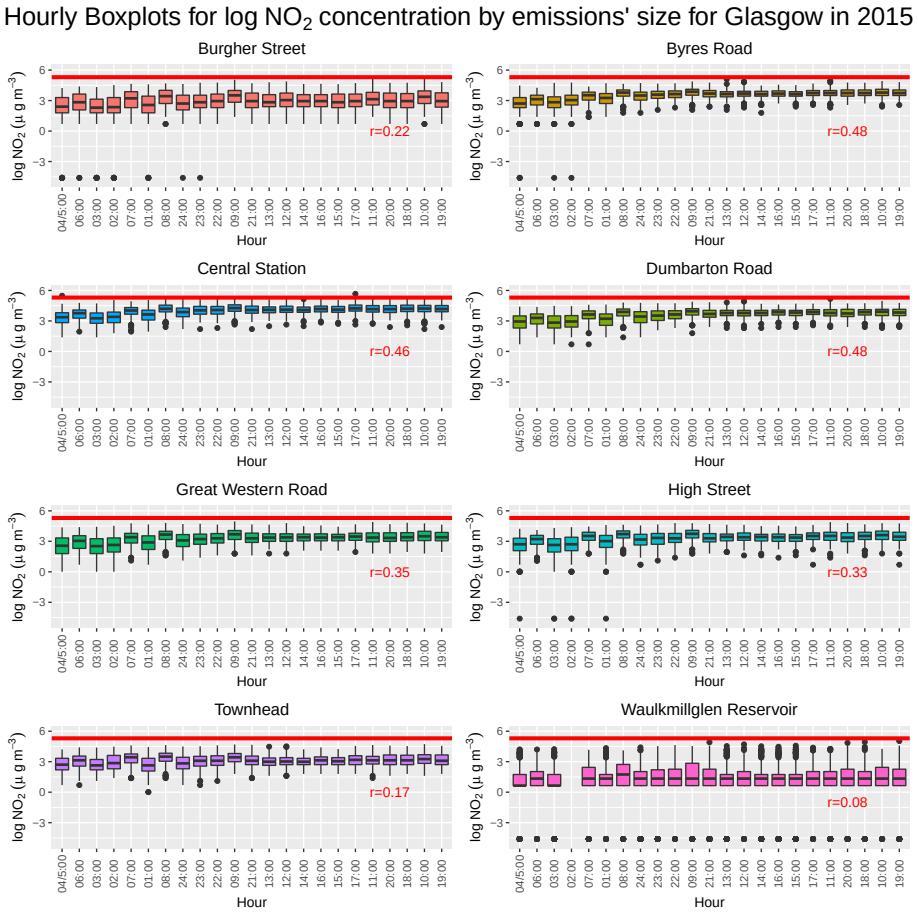

FIGURE B.11: Boxplots for the ordered (by emissions' size ordered from smallest to largest) log NO_2 concentrations (μg m^{-3}) over a 24-hour cycle at the eight monitoring stations in Glasgow in 2015. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation. The correlations for each pairing are also provided

Lastly, the emissions cycle is superimposed (in magenta) over the boxplots for the log NO_2 concentrations for the eight monitoring stations in Figure B.12. The boxplots show that there is a hour-to-hour variation in concentrations during the day for all stations. The trend for the boxplots is similar to the emissions line, though not identical. Central Station is the only monitoring station where the hourly limit of 200 μg m^{-3} has been breached and it is interesting to note that the breaches have occurred at 05:00 and 17:00, which is before the peak emissions. This could be explained by the fact that traffic around Central Station is always heavy and the high buildings do not allow for pollutants to escape, while for the other locations the higher pollutant concentrations are quicker to disperse.

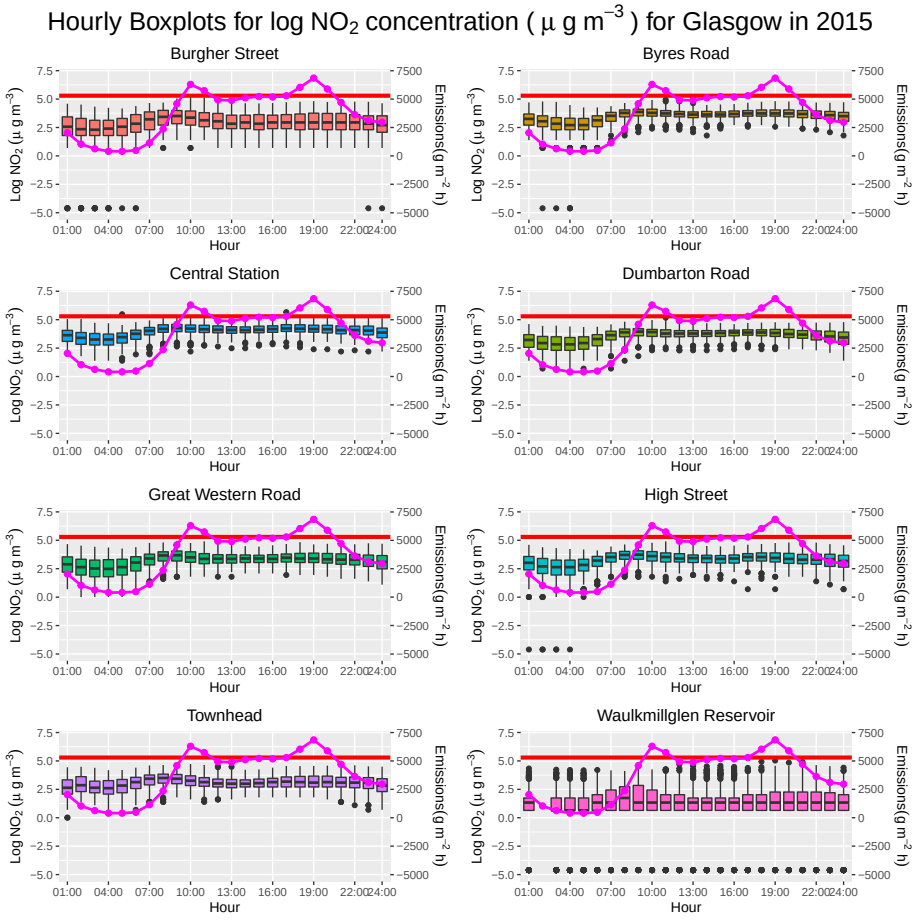

FIGURE B.12: Boxplots for the log NO_2 concentrations (μg m^{-3}) over a 24-hour emissions (g m^{-2} h) cycle at the eight monitoring stations in Glasgow in 2015. The emissions (g m^{-2} h) for each hour are superimposed in magenta. A red line at 5.30 is added for the log of the 200 μg m^{-3} regulation.

B.5 Findings

Overall, it was found that the true NO_2 hourly concentrations taken in Glasgow in 2015 are negatively correlated with temperature and wind speed, whereas the concentrations are positively correlated with emissions. For wind direction, a circular variable has to be used as winds get more eastern prevailing, the NO_2 concentrations appear to increase.

Appendix C

Matrix properties

In Section 8.2, a hyperspatial-temporal model is proposed for the modelling and emulation of the hourly NO_2 time series across the LHC space as simulated by ADMS-Urban. Here, some properties of matrices used to provide computational efficiency are discussed. Firstly, properties of Kronecker products are applied to simplify the computation. Let $\mathbf{A}$ be an $m \times m$ matrix and $\mathbf{B}$ be an $n \times n$ matrix. The Kronecker product of the two matrices is $\mathbf{A} \otimes \mathbf{B} = \mathbf{C}$, where $\mathbf{C}$ has dimensions $mn \times mn$. Then:

- the inverse of a Kronecker product [142] is:

$$\mathbf{C}^{-1} = (\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \,; \text{and} \tag{C.1}$$

- the determinant of a Kronecker product [142] is:

$$|\mathbf{C}| = |\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^n |\mathbf{B}|^m \,. \tag{C.2}$$

Secondly, simplifications can be applied to the estimation of the inverse and determinant of the temporal correlation matrix, given that it has an AR(1) correlation structure. Let $\mathbf{D}$ $(T \times T)$ be an AR(1) correlation matrix with correlation coefficient ρ, then:

- the determinant of $\mathbf{D}$ [84] is:

$$|\mathbf{D}| = (1 - \rho^2)^{T-1} \,. \tag{C.3}$$

  Hence, the logarithm of the determinant is:

$$\log|\mathbf{D}| = (T-1)\log(1 - \rho^2) \,; \text{and} \tag{C.4}$$

- the inverse of $\mathbf{D}$ [110] is a tri-diagonal matrix of the form:

$$\mathbf{D}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & \dots & -\rho & 1 \end{bmatrix}. \tag{C.5}$$

Appendix D

Model testing for the hyperspatial-temporal model

D.1 Background

In both Chapters 5 and 6, it was found that the hyperspatial range parameters are difficult to estimate correctly. In Section 6.2, a simulation study on estimating the hyperspatial range parameters was performed and it was found that all of the tested models struggled with estimating the hyperspatial range parameters. This is in agreement with Cressie who has stated hyperspatial range parameters are hard to estimate correctly [50]. However, it was shown that even with "sensible" [50] estimates for the hyperspatial range parameters the RMSPE is relatively unaffected. In a similar way to the first simulation study in Section 6.2, a short study is performed to assess the ability of the hyperspatial-temporal model to correctly identify the hyperspatial range and temporal parameters in a setting very close to the real-life data.

The study will be based on the simulated NO_2 concentrations for the first one hundred hours (January 1^{st} to 5^{th}) for all 100 locations in the LHC space. Such a small size dataset was chosen as it takes between 9-12 hours to run the model on a 2016 MacBook Pro with 16GB memory and 2.9 GHz Quad-Core Intel Core i7 processor. In Figure D.1, a boxplot for the log hourly NO_2 concentrations as simulated by ADMS-Urban is presented with the corresponding actual concentrations for 2015 superimposed in blue points. The simulations capture the pattern of the true concentrations quite well and 85% of the blue points lie within the ADMS-Urban simulation corresponding interquartile range on Figure D.1. The subset is chosen as a good representation of the real data by the simulated concentrations. The subset of the simulated data provides values both close

to zero and to the breach limit of 200 μg m^{-3} (or the equivalent log limit of 5.30 μg m^{-3}).

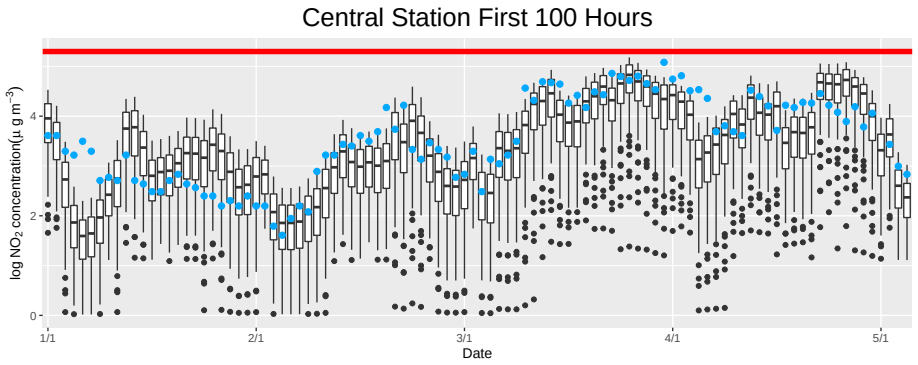

FIGURE D.1: Boxplots for the log NO_2 hourly concentrations (μg m^{-3}) for the first one hundred hours from the ADMS-Urban simulations. The true log NO_2 hourly concentrations for the corresponding hour are the coloured points in blue on top of the boxplots. The log NO_2 hourly limit of 5.30 μg m^{-3} is represented by a red line.

Since the first one hundred hours subset only contains the hourly times series for just over 5 days, the model from Equation 8.2 will be used with the factor variable for hour of the day and the b-spline for Week Number covariates removed. To check that the reduced model remains appropriate, an AR(1) model was applied to the univariate time series models for simulation scenario 16 to check whether the model does account for the temporal correlation. The corresponding ACF and PACF plots are shown in Figure D.2 and show that there is no residual temporal correlation after an AR(1) structure is applied.

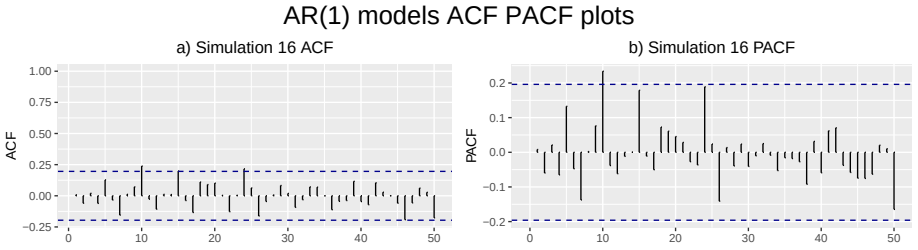

FIGURE D.2: ACF and PACF plots for ARIMA models with AR(1) correlation structure for the log NO_2 concentrations (μg m^{-3}) with temperature (°C), wind speed (m/s), an interaction between temperature and wind speed, wind direction (°) and emissions (g m^{-2} h) as covariates from the first 100 hours for simulation scenario 16 at Central Station.

In order to get an initial assessment of the possible values of the hyperspatial range parameters across the different hours, the variograms for the residuals for an intercept model with log NO_2 concentrations as response for each hour across the 100 locations in the LHC space were examined in a similar way as in Subsection 4.2.4. An intercept only model is used to estimate the largest possible values for the hyperspatial parameters even when no fixed effects are accounted for. The variograms for log NO_2 concentrations

residuals for the 1st hour (1/01 at 01:00) are presented in Figure D.3. The variograms for the other hours are similar and hence, omitted to avoid repetition. The variogram shows that the individual variograms for emissions, wind speed and wind direction do not plateau and the ranges of the hyperspatial range parameters will be hard to estimate. The 3D variogram mimics the emissions variogram but given that the emissions have the widest span of values (from -100% to +20%), the 3D variogram could be reflecting that rather than emissions being the most dominant hyperspatial variable. Overall, the annual average variograms in Figure 4.24 and the log NO_2 hourly concentration residuals variograms in Figure D.3 are very similar in shape with only the values of the semivariance being different. Therefore, the issues with estimating the parameters at extreme values are likely to be repeat themselves.

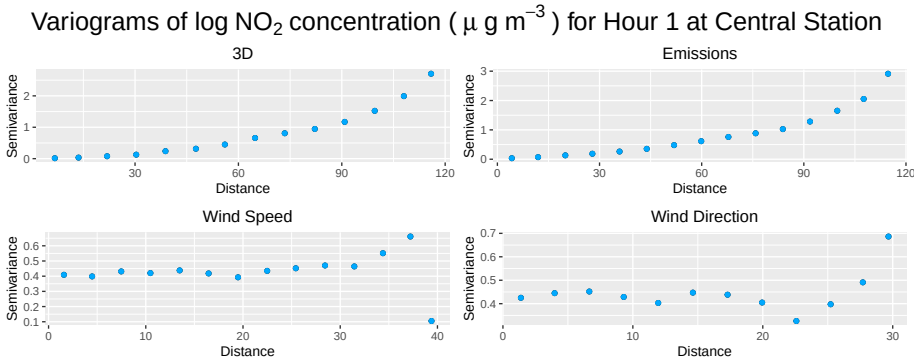

FIGURE D.3: Variograms for the log NO_2 concentrations (μg m^{-3}) for the 1st hour at Central Station in Glasgow.

D.2 Initialising the study

In order to asses the quality of the hyperspatial range and temporal parameters estimates, fifty datasets are simulated using the following parameters:

- the response vector $\mathbf{y} = [y_{1,1}, \ldots, y_{1,100}, \ldots, y_{100,1}, \ldots, y_{100,100}]^\top$ (10000×1) is the simulated responses for one hundred time steps at one hundred locations in the LHC for Central Station;
- the block-diagonal matrix $\mathbf{S}$ (10000×700) will have block $\mathbf{B}_i$ with seven covariates (intercept, temperature, wind speed, sin (wind direction), cos (wind direction), an interaction between temperature and wind speed, and emissions). The covariates used are the temperature, wind speed, wind direction and emissions time series as presented in Section 8.1;

- the fixed effect parameter $\boldsymbol{\beta}$ is estimated by applying the gls model from the nlme package in **R** to each of the first one hundred hours of ADMS-Urban with an AR(1) correlation structure;
- the overall variance parameter σ^2 is set to 0.4 in a similar way to the simulation studies in Section 6.2;
- the hyperspatial correlation matrix $\mathbf{R}(\boldsymbol{\theta})$ (100×100) is estimated based on $\boldsymbol{\theta}$ being set to be a vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^\top$. The values for $\boldsymbol{\theta}$ were chosen based on the variograms for each of the inputs in Figure D.3. Since the variograms do not plateau and never suggest range values, the set $\boldsymbol{\theta} = [75, 20, 15]^\top$ was chosen;
- the temporal correlation matrix $\boldsymbol{\Sigma}(\rho)$ (100×100) is an AR(1) structure with $\rho = 0.82$ (based on the average estimate for ρ from the univariate time series models with which the values of $\boldsymbol{\beta}$ were chosen); and
- the error vector $\mathbf{z}$ (10000×1) is randomly drawn from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Lambda} = \sigma^2\, \mathbf{R}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}(\rho)$.

D.3 Results

The bias for the estimates for the hyperspatial range and temporal parameters are presented in Table D.1. For the temporal parameter, the bias is negative, whereas for the hyperspatial range parameters the bias is positive. The largest bias is for the temporal parameter and it equates to 10% of the true value, whereas for the hyperspatial range parameters the bias is less than 1%. The hyperspatial-range and temporal parameter estimates are relatively well estimated by the proposed hyperspatial-temporal model (in general in Equation 8.3 and with the same covariates as in Equation 8.2).

Parameter	**True**	**Mean**	**Bias**
$\widehat{\rho}$	0.82	0.75	-0.07
$\widehat{\theta_1}$	75	75.42	0.42
$\widehat{\theta_2}$	20	20.21	0.12
$\widehat{\theta_3}$	15	15.13	0.13

TABLE D.1: Comparing the hyperspatial range and temporal parameter estimates and their bias from the hyperspatial-temporal model testing for simulated log NO_2 concentrations (μg m^{-3}) with temperature (°C), wind speed (m/s), an interaction between temperature and wind speed, sin/cos (wind direction) (°) and emissions (g m^{-2} h) as covariates for the first 100 hours for the hundred simulation scenarios based on the ADMS-Urban NO_2 concentrations at Central Station.

D.4 Findings

A simulation study (over one hundred hours for all 100 locations in the LHC space) was performed to assess how well the hyperspatial-temporal model estimates the hyperspatial range and temporal parameters. It was found that the model estimates almost perfectly (less than 1% bias) the hyperspatial-range parameters and there is a slight negative bias (10%) in estimates the temporal parameter. Therefore, the model appears to provide more accurate parameter estimates in comparison to the multivariate GP model in Section 6.2. Even for a small dataset with high variability in the observations, the hyperspatial range and temporal parameters are well estimated and there is no need to assess the effect of mis-estimating these parameters on the RMSPE (as it was shown in Section 6.2 that even for larger mis-estimates, the RMSPE remains relatively unaffected). Hence, the hyperspatial-temporal model is appropriate and will be applied to the ADMS-Urban simulated dataset for Central Station.

Bibliography

[1] (2021). Clean Air Zones - what are they and where are they? *RAC*. https://www.rac.co.uk/drive/advice/emissions/clean-air-zones/.

[2] Aberdeen City Council (2016). Hydrogen Bus Project. http://www.aberdeencity.gov.uk/council_government/shaping_aberdeen/Shaping_Aberdeen_Hyrdogen_Bus.asp.

[3] Aberdeen City Council (2019). 2019 Air Quality Annual Progress Report. *LAQM Annual Progress Report*. https://www.aberdeencity.gov.uk/sites/default/files/2019-08/Air%20Quality%20Report%202019.pdf.

[4] Air Quality in Scotland (2018a). Cleaner air for Scotland (CAFS) strategy. https://www.gov.scot/publications/cleaner-air-scotland-air-quality-public-attitudes-behaviour-review-summary-report/.

[5] Air Quality in Scotland (2018b). Monitoring site summary. https://www.scottishairquality.scot/latest/summary.

[6] Air Quality in Scotland, A. Daily Air Quality Index (DAQI). https://www.scottishairquality.scot/air-quality/daqi.

[7] Air Quality in Scotland, A. Standards: Summary of objectives of the National Air Quality Strategy. http://www.scottishairquality.co.uk/air-quality/standards.

[8] Air Quality Scotland (2016). Monitoring. http://www.scottishairquality.co.uk/air-quality/monitoring.

[9] Alam, J. (2021). Case Studies in Data Optimization Using Python. *Chapman and Hall/CRC*.

[10] Alphasense Air: Sensors for air quality networks (2018). Nitrogen Dioxide Sensors (NO2). alphasense.com/products/nitrogen-dioxide/.

[11] Andrews, A. (2014). THE CLEAN AIR HANDBOOK: A practical guide to EU air quality law. *Client Earth*. https://www.clientearth.org/latest/documents/the-clean-air-handbook/.

[12] Andria, G. et al. (2008). Modelling study for assessment and forecasting variation of urban air pollution. *Measurement 41*.

[13] Ansari, T. U. et al. (2021). Temporally resolved sectoral and regional contributions to air pollution in Beijing: informing short-term emission controls. *Atmospheric Chemistry and Physics 88 (2)*.

[14] Anton, H. and C. Rorres (2011). Elementary linear algebra with supplemental applications. *Wiley 10th Edition*.

[15] Arvind, D. et al. (2016). The AirSpeck family of static and mobile wireless air quality monitors. *Digital System Design (DCD)*, 207–214. Euromicro Conference IEEE.

[16] Baker, T. (2022). Electric vehicles: New homes will be required to have EV charging stations from 2022, Boris Johnson to announce. *Sky News*.

[17] Banks, D. L. and M. B. Hooten (2021). Statistical Challenges in Agent-Based Modeling. *The American Statistician 0*(0), 1–8.

[18] Bastos, L. and A. O'Hagan (2009). Diagnostics for Gaussian Process Emulators. *Technometrics 51*(4), 425–438.

[19] Bayarri, M. et al. (2009). Using statistical and computer models to quantify volcanic hazards. *Technometrics 51 (4)*.

[20] BBC (2008). Q & A: London Low Emission Zone. *BBC News*. http://news.bbc.co.uk/1/mobile/england/london/7213360.stm.

[21] BBC (2016). Glasgow breaches WHO air pollution safety levels. *BBC News*. http://www.bbc.co.uk/news/uk-scotland-glasgow-west-36274801.

[22] BBC (2019). Traffic-free days begin in Edinburgh city centre. *BBC News*. https://www.bbc.co.uk/news/uk-scotland-edinburgh-east-fife-48139403.

[23] Bell, M., D. Davis, and T. Fletcher (2004). A Retrospective Assessment of Mortality from the London Smog Episode of 1952: The Role of Influenza and Pollution. *Environmental Health Perspectives 112*(1), 6–8.

[24] Benmouiza, K. and A. Cheknane (2016). Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models. *Theoretical and Applied Climatology 124-3*.

[25] Berrocal, V. J. et al. (2009). A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of Agriculture, Biological, and Environmental Statistics 15-2*.

[26] Blanc, E. (2017). Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models. *Agricultural and Forest Meteorology 236*.

[27] Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet 327*, 307–310.

[28] Bowman, V. and D. C. Woods (2016). Emulation of multivariate simulators using thin-plate splines with application to atmospheric dispersion. *SIAM-ASA JOURNAL ON UNCERTAINTY QUANTIFICATION 4 (1)*.

[29] Callaghan, D. (2014). Glasgow City Council Detailed Assessment. https://www.glasgow.gov.uk/CHttpHandler.ashx?id=32459&p=0.

[30] Campbell, M. J. and M. J. Gardner (1988). Calculating confidence intervals for some non-parametric analyses. *Statistics in Medicine 296*, 1454–1456.

[31] Cape, J. N. (2005). Review of the use of passive diffusion tubes for measuring concentrations of nitrogen dioxide in air. *Centre for Ecology and Hydrology (Edinburgh Research Station)*.

[32] Carroll, R. J. et al. (1997). Ozone Exposure and Population Density in Harris County, Texas. *Journal of the American Statistical Association (92:438)*.

[33] Carruthers, D. (2006). Intercomparison of five modelling methods including ADMS-Airport and EDMS for predicting air quality at London Heathrow Airport. *CERC, UK*.

[34] Carslaw, D. and K. Ropkins (2019). Wind and Pollution Roses. *The Openair Project*. https://bookdown.org/david_carslaw/openair/sec-windRose.html.

[35] Carslaw, D. and K. Ropkins (2021). openair: Tools for the Analysis of Air Pollution Data. *The Comprehensive R Archive Network*.

[36] CERC (2017). ADMS-Urban: Urban Air Quality Management System. *User Guide 4.1.1*. https://www.cerc.co.uk/environmental-software/user-guides.html.

[37] CERC (2018). ADMS-Urban. *Environmental Software and Service*. http://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html.

[38] CERC (2020a). ADMS 5. *Environmental Software and Service*. https://www.cerc.co.uk/environmental-software/ADMS-model.html.

[39] CERC (2020b). ADMS-Airport. *Environmental Software and Service*. https://www.cerc.co.uk/environmental-software/ADMS-Airport-model.html.

[40] CERC (2020c). ADMS-Roads. *Environmental Software and Service*. https://www.cerc.co.uk/environmental-software/ADMS-Roads-model.html.

[41] CERC (2020d). ADMS-Screen. *Environmental Software and Service*. https://www.cerc.co.uk/environmental-software/ADMS-Screen-model.html.

[42] CERC (2020e). ADMS-Urban. *Environmental Software and Service*. https://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html.

[43] CERC (2020f). Environmental software. *Environmental Software and Service*. https://www.cerc.co.uk/environmental-software.html.

[44] Chatfield, C. (2005). The Analysis of Time Series: An Introduction. *Chapman&Hall/CRC Sixth Edition*, 58–59.

[45] City of York Council (2015). Go Ultra Low York. *iTravel York*.

[46] Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association 74 (368)*, 829–836.

[47] Cohen, I., Y. Huang, J. Chen, and J. Benesty (2009). Pearson Correlation Coefficient. *Noise Reduction in Speech Processing 5*, 37–40.

[48] COMEAP (2016). Long-term Exposure to Air Pollution and Chronic Bronchitis. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/541745/COMEAP_chronic_bronchitis_report_2016__rev_07-16_.pdf.

[49] Conti, S. and A. O'Hagan (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference 140*, 640–651.

[50] Cressie, N. (1993). Statistics for Spatial Data. *Wiley Series in Probability and Statistics*.

[51] Cressie, N. and C. K. Wikle (2011). Statistics for Spatio-Temporal Data. *Wiley Series in Probability and Statistics*.

[52] Currin, C. et al. (1991). Bayesian Prediction of Deterministic Functions With Applications to the Design and Analysis of Computer Experiments. *Journal of the American Statistical Association 86*.

[53] Damblin, G. and A. Ghione (2021). Adaptive use of replicated Latin Hypercube Designs for computing Sobol' sensitivity indices. *Reliability Engineering & System Safety 212*.

[54] David G, N. and D. Leith (2010). Use of passive diffusion tubes to monitor air pollutants. *Journal of the Air & Waste Management Association 60 (2)*.

[55] Dawid, A. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika 68*, 265–274.

[56] de Boor, C. (1970). On Calculating with B-Splines. *Journal of Approximation Theory 6*.

[57] Dedele, A. and A. Miskinyte (2015). The statistical evaluation and comparison of ADMS-Urban model for the prediction of nitrogen dioxide with air quality monitoring network. *Environmental Monitoring and Assessment 187*, 578.

[58] DEFRA (2007). The Air Quality Strategy for England, Scotland, Wales and Northern Ireland. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/69337/pb12670-air-quality-strategy-vol2-070712.pdf.

[59] DEFRA (2011). Monitoring Networks. https://uk-air.defra.gov.uk/networks/network-info?view=automatic.

[60] DEFRA (2017). Air quality plan for nitrogen dioxide (NO2) in UK (2017). *Policy paper*. https://uk-air.defra.gov.uk/library/no2ten/.

[61] DEFRA (2017). UK plan for tackling roadside nitrogen dioxide concentrations. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633269/air-quality-plan-overview.pdf.

[62] DEFRA (2020). Air modelling for DEFRA. *UK Air*. https://uk-air.defra.gov.uk/research/air-quality-modelling?view=modelling#:~:text=CMAQ%20is%20an%20open%20source,km%20squares%20for%20the%20UK.

[63] Diggle, P. J. and P. J. R. Jr. (2007). Model-based Geostatistics. *Springer*.

[64] Diggle, P. J., R. Menezes, and T. li Sun (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59 (2)*.

[65] Dobson, A. J. (2002). An Introduction to Generalised Linear Models. *Chapman&Hall/CRC Second Edition*.

[66] Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association 56*, 52–64.

[67] Eddelbuettel, D. (2013). Seamless R and C++ Integration with Rcpp. *Springer*.

[68] Environment and Forestry Directorate (2015). Cleaner air for Scotland: the road to a healthier future. *Environment and climate change*. https://www.gov.scot/publications/cleaner-air-scotland-road-healthier-future/.

[69] Environment and Forestry Directorate (2019). Cleaner Air for Scotland strategy: independent review. *Environment and climate change*. https://www.gov.scot/publications/cleaner-air-scotland-strategy-independent-review/.

[70] EPA (2020). Managing Air Quality - Ambient Air Monitoring. https://www.epa.gov/air-quality-management-process/managing-air-quality-ambient-air-monitoring.

[71] European Commission (2018). A Europe that protects: Clean air for all. https://ec.europa.eu/environment/pubs/pdf/factsheets/air/en.pdf.

[72] European Commission (2020). The Mobility Packages 1,2,3. *IRU*. https://www.iru.org/where-we-work/europe/europe-overview/european-commission-mobility-package.

[73] European Environment Agency (2019a). Air quality in Europe. *Publications Office of the European Union*. https://www.eea.europa.eu/publications/air-quality-in-europe-2019.

[74] European Environment Agency (2019b). European Union emission inventory report 1990-2017 under the UNECE Convention on Long-range Transboundary Air Pollution (LRTAP). *Publications Office of the European Union*. https://www.eea.europa.eu/publications/european-union-emission-inventory-report-1990-2018.

[75] European Environment Agency (2020). Air quality in Europe - 2020 report. *Publications Office of the European Union*. https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report.

[76] European Parliament (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union*. https://www.eea.europa.eu/policy-documents/directive-2008-50-ec-of.

[77] European Parliament (2010). Directive 2010/75/EU of the European Parliament and of the Council of 24 November 2010 on industrial emissions (integrated pollution prevention and control). *Official Journal of the European Union*. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:334:0017:0119:en:PDF.

[78] Faraway, J. (2006). Extending the Linear Model with R.

[79] Faraway, J. (2009). Linear Models with R.

[80] Fassó, A. and F. Finazzi (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics 22*, 735–748.

[81] Finazzi, F. and A. Fassó (2014). D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. *Journal of Statistical Software 62*, 1–29.

[82] Finazzi, F., Y. Napier, M. Scott, A. Hills, and M. Cameletti (2019). A statistical emulator for multivariate model outputs with missing values. *Atmospheric Environment 199*, 415–422.

[83] Finazzi, F. and L. Paci (2019). Quantifying personal exposure to air pollution from smartphone?based location data. *Biometrics: Journal of the International Biometric Society 75:4*, 1356–1366.

[84] Finch, P. (1960). On the covariance determinants of moving-average and autoregressive models. *Biometrika 47*.

[85] Franck, C. T. and R. B. Gramacy (2020). Assessing Bayes Factor Surfaces Using Interactive Visualization and Computer Surrogate Modeling. *The American Statistician 74 (4)*.

[86] Fricker, T., J. Oakley, and N. Urban (2013). Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics 55 (1)*.

[87] Gardner, M. and D. Altman (1989). Statistics with Confidence: Confidence Intervals and Statistical Guidelines. *British Medical Journal 1*.

[88] Gelman, A. et al. (2003). Bayesian Data Analysis. *London: Chapman and Hall Second Edition*.

[89] Gianniotis, S. (2018). Empirical Study on Bayesian and Frequentist Model Calibration of Computer Models.

[90] Glasgow City Council (2019). Avenues. https://www.glasgow.gov.uk/avenues.

[91] Glasgow City Council (2021). Glasgow's Low Emission Zone (LEZ). https://www.glasgow.gov.uk/LEZ.

[92] Google Maps (2021). Google Maps online. https://www.google.com/maps.

[93] Government, U. (2022a). Electric Vehicle Homecharge Scheme: customer guidance.

[94] Government, U. (2022b). Plug-in taxi grants: eligibility and applications.

[95] Gradshteyn, I. S. and I. M. Ryzhik (2000). Hessian Determinants. *CA: Academic Press*.

[96] Gramacy, R. B. and D. W. Apley (2015). Local Gaussian Process Approximation for Large Computer Experiments. *Journal of Computational and Graphical Statistics*, 561–578.

[97] Guillas, S., N. Glover, and L. Malki-Epshtein (2014). Bayesian calibration of the constants of the k-ϵ turbulence model for a CFD model of street canyon flow. *Computer Methods in Applied Mechanics and Engineering 279.*

[98] Haining, R. (1990). Spatial data analysis in the social and environmental sciences. *Cambridge University Press First Edition.*

[99] Hooten, M., C. Wikle, and M. Schwob (2020). Statistical Implementations of Agent-Based Demographic Models. *International Statistical Review 88 (2).*

[100] Hu, X.-M. et al. (2021). Multisensor and Multimodel Monitoring and Investigation of a Wintertime Air Pollution Event Ahead of a Cold Front Over Eastern China. *Journal of Geographical Research: Atmospheres 126 (10).*

[101] Huang, H. et al. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic Methods in Accident Research 14*, 10–21.

[102] Iman, R. L. and J. C. Helton (1988). An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models. *Risk Analysis 8 (1).*

[103] Jack, E. (2019). Estimating the changes in health inequalities across Scotland over time. *PhD Thesis University of Glasgow*.

[104] James, A. T. (1964). Distributions of Matrix Variates and Latent Roots Derived from Normal Samples. *The Annals of Mathematical Statistics 35 (2).*

[105] James, G., D. Witten, T. Hastie, and R. Tibshirani (2017). An Introduction to Statistical Learning with Applications in R. *Springer e-Book.*

[106] Javier, W. R. and A. K. Gupta (1985). On matric variate-T distribution. *Communications in Statistics - Theory and Methods 4*, 1413–1425.

[107] Jian, L. et al. (2012). An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. *Science of The Total Environment 426.*

[108] Jin, J.-Q. et al. (2019). Using Bayesian spatio-temporal model to determine the socio-economic and meteorological factors influencing ambient $PM_{2.5}$ levels in 109 Chinese cities. *Environmental Pollution 254-A.*

[109] Johnson, M. E., L. M. Moore, and D. Ylvisaker (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference 26 (2)*.

[110] Kac, M., W. Murdock, and G. Szeg (1953). On the Eigen-Values of Certain Hermitian Forms. *Indiana University Mathematics Journal 4*.

[111] Kennedy, M. and A. O'Hagan (1998). Bayesian Calibration of Complex Computer Models. *Technical Report 98-10*.

[112] Kennedy, M. and A. O'Hagan (2001). Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc. B 63*, 425–464.

[113] Lee, D. (2020). Spatial Notes 2020. *University of Glasgow*.

[114] Lee, D., C. Ferguson, and R. Mitchell (2009). Air pollution and health in Scotland: a multicity study. *Biostatistics 10-3*.

[115] Lee, D., C. Miller, and M. Scott (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174-1*.

[116] Lewis, A. et al. (2015). Evaluating the performance of low cost chemical sensors for air pollution research. *Faraday Discussions 189*, 85–103.

[117] Lewis, A. and P. Edwards (2016). Validate personal air-pollution sensors. *Nature 535*, 7610.

[118] Lin, C. et al. (2017). Practical Field Calibration of Portable Monitors for Mobile Measurements of Multiple Air Pollutants. *Atmosphere 8(12)*.

[119] Lin, Y.-C. et al. (2021). Air pollution diffusion simulation and seasonal spatial risk analysis for industrial areas. *Environmental Research 94*.

[120] Mallet, V. et al. (2018). Meta-modelling of ADMS-Urban by dimension reduction and emulation. *Atmospheric Environment 184*, 37–46.

[121] Matheron, G. (1963). Principles of geostatistics. *Economic Geology 58*, 1246–1266.

[122] McCullagh, P. and J.A.Nelder (1989). Generalised Linear Models. *Second Edition*.

[123] McDonald, M. (2011). Hybrid, Electric, Plug-In; What's the difference? *Top Speed*. https://www.topspeed.com/cars/car-news/hybrid-electric-plug-in-what-s-the-difference-ar106626.html.

[124] McHugh, C., D. Carruthers, and H. Edmunds (1997). ADMS-Urban: an air quality management system for traffic, domestic and industrial pollution. *International Journal of Environment and Pollution 8*, 666–674.

[125] McKay, M., R. Beckham, and W. Conover (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics 21 (2)*.

[126] McLachlan, G. J. and T. Krishnan (1996). The EM Algorithm and Extensions.

[127] McMillan, N. J. et al. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics 21-1*.

[128] Mead, M. et al. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment 70*, 186–203.

[129] Met Office (2015). The Great Smog of 1952. http://www.metoffice.gov.uk/learning/learn-about-the-weather/weather-phenomena/case-studies/great-smog.

[130] Mohebbi, M. et al. (2011). A poisson regression approach for modelling spatial autocorrelation between geographically referenced observations. *BMC Medical Research Methodology*.

[131] Mooibroek, D. and J. Wesseling (2009). Geinterpoleerde meteorologie voor SRM-1: toepassing in 2007. *National Institute for Public Health and the Environment*.

[132] Morris, M. D. and T. J. Mitchell (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference 43*.

[133] M.Urban, N. and T. E.Fricker (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Computers & Geosciences 36 (6)*, 746–755.

[134] Nguyen, P. and J. Wesseling (2013). TModelleren van scheepvaartemissie fase 1 : Harmonisatie tussen SRM3 en SRM2 voor NOx. *National Institute for Public Health and the Environment*.

[135] NOAA-NASA. Aerosol Optical Depth: Quick Guide. http://cimss.ssec.wisc.edu/goes/OCLOFactSheetPDFs/ABIQuickGuide_BaselineAerosolOpticalDepth.pdf.

[136] NOAA-NASA (2012). GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document For Suspended Matter/Aerosol Optical Depth andAerosolSize Parameter. https://www.goes-r.gov/education/docs/Factsheet_ABI.pdf.

[137] Nocedal, J. and S. Wright (2006). Numerical Optimization. *Springer*, 136–143.

[138] Oakley, J. (1999). Bayesian uncertainty analysis for complex computer codes. *University of Sheffield Ph.D. thesis*.

[139] Oakley, J. E. and A. O'Hagan (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society. Series B (Methodological) 66 (3)*.

[140] O'Hagan, A. (2004). Bayesian Analysis of Computer Code Outputs: A Tutorial.

[141] O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety 91*, 1290–1300.

[142] Onishchik, A. (2018). Encyclopedia of Mathematics. *Springer / The European Mathematical Society*.

[143] Overstall, A. and D. Woods (2016a). Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. *Journal of the Royal Statistical Society: Applied Statistics Series C 65*, 483–505.

[144] Overstall, A. and D. Woods (2016b). Supplementary Material for "Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model". *Journal of the Royal Statistical Society: Applied Statistics Series C 65*.

[145] Pang, X. et al. (2017). Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air quality monitoring. *Sensors and Actuators B: Chemical 340*, 829–837.

[146] Pannullo, F. et al. (2015). Improving spatial nitrogen dioxide prediction using diffusion tubes: A case study in West Central Scotland. *Atmospheric Environment 118*.

[147] Parliament of the United Kingdom (1993). Clean Air Act. *The National Archives*. https://www.legislation.gov.uk/ukpga/1993/11/contents.

[148] Pebesma, E. and B. Graeler (2019a). Introduction to Spatio-Temporal Variography.

[149] Pebesma, E. and B. Graeler (2019b). Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation.

[150] Pinheiro, J. et al. (2020). nlme: Linear and Nonlinear Mixed Effects Models. *CRAN*. https://cran.r-project.org/web/packages/nlme/index.html.

[151] Pinto, J. A. et al. (2020). Kriging method application and traffic behavior profiles from local radar network database: a proposal to support traffic solutions and air pollution control strategies. *Sustainable Cities and Society*.

[152] Pirani, M., J. Gulliver, G. W. Fuller, and M. Blangiardo (2014). Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. *Journal of Exposure Science and Environmental Epidemiology 24*, 319–327.

[153] Pope, C. A., M. Ezzati, and D. W. Dockery (2009). Fine-Particulate Air Pollution and Life Expectancy in the United States. *The New England Journal of Medicine 360*(4), 376–386.

[154] Popescu, F. and I. Ionel (2010). Anthropogenic air pollution sources. *Air Quality 1*.

[155] Popoola, O. A. et al. (2018). Use of networks of low cost air quality sensors to quantify air quality in urban settings. *Atmospheric Environment 194*.

[156] Qian, P. Z. G., H. Wu, and C. F. J. Wu (2008). Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors. *Technometrics 50 (3)*.

[157] R Core Team (2013). R: A Language and Environment for Statistical Computing.

[158] Rasmussen, C. E. and C. K. I. Williams (2006). Gaussian Processes for Machine Learning. *The MIT Press*.

[159] Righi, S., P. Lucialli, and E. Pollini (2009). Statistical and diagnostic evaluation of the ADMS-Urban model compared with an urban air quality monitoring network. *Atmospheric Environment 43*, 3850–3857.

[160] Rincon, P. (2015). Court orders UK to cut NO_2 air pollution. *BBC News*. https://www.bbc.co.uk/news/science-environment-32512152.

[161] Rincon, P. (2016). Green group wins air pollution court battle. *BBC News*. https://www.bbc.co.uk/news/science-environment-37847787.

[162] Rivera, J. P. et al. (2015). An Emulator Toolbox to Approximate Radiative Transfer Models with Statistical Learning. *Remote sensing 7 (7)*, 9347–9370.

[163] Rougier, J. (2007). Lightweight emulators for multivariate deterministic functions. *Technical Report*.

[164] Rouhani, S. and H. Wackernagel (1990). Multivariate geostatistical approach to space?time data analysis. *Water Resources Research 26 (4)*.

[165] Roustant, O., D. Ginsbourger, and Y. Deville (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software 51*(1), 1–55.

[166] Royal College of Physicians (2016). Every breath we take: The lifelong impact of air pollution. https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution.

[167] Royal College of Physicians (2018). Reducing air pollution in the UK: Progress Report 2018. https://www.rcplondon.ac.uk/news/reducing-air-pollution-uk-progress-report-2018.

[168] S. Girard, V. Mallet, I. K. and A. Mathieu (2016). Emulation and Sobol' sensitivity analysis of an atmospheric dispersion model applied to the Fukushima nuclear accident. *Journal of Geophysical Research 121*.

[169] Sacks, J. et al. (1989). Design and Analysis of Computer Experiments. *Statistical Science 4*, 409–423.

[170] Sacks, J. D. et al. (2014). Influence of Urbanicity and County Characteristics on the Association between Ozone and Asthma Emergency Department Visits in North Carolina. *Environmental Health Perspectives 122(5)*.

[171] Schuurman, N. K., R. P. P. P. Grasman, and E. L. Hamaker (2016). A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models. *Multivariate Behavioral Research 51*, 185–206.

[172] Schwartz, J. and A. Marcus (1990). Mortality and air pollution in London: A Time Series Analysis. *American Journal of Epidiomology 131-1*.

[173] Sebastian Burhenne, D. J. and G. P. Henze (2011). Sampling based on Sobol' sequences for Monte Carlo techniques applied to building simulations. *Conference of International Building Performance Simulation Association, Sydney 12*.

[174] SEPA (2020). Environment - Air. https://www.sepa.org.uk/environment/air/.

[175] Shaddick, G. et al. (2018). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 67:1*, 231–253.

[176] Shaddick, G. and J. Wakefield (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 51*(3), 351–372.

[177] Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52 (3-4)*, 591?611.

[178] Shen, J. et al. (2019). Block Design-Based Key Agreement for Group Data Sharing in Cloud Computing. *IEEE Transactions on Dependable and Secure Computing 16 (6)*, 996–1010.

[179] Sherman, M. (2011). Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties. *Wiley Series in Probability and Statistics*.

[180] Shields, M. D. and J. Zhang (2016). The generalization of Latin hypercube sampling. *Reliability Engineering and System Safety 148*.

[181] Smyth, S. C. et al. (2006). Evaluation of CMAQ O_3 and $PM_{2.5}$ performance using Pacific 2001 measurement data. *Atmospheric Environment 189*.

[182] Sobol, I. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitelnoi Matematiki i Matematicheskoi Fiziki 7*.

[183] Stratton, S. (2014). Air Quality Further Assessment for Musselburgh. *East Lothian Council*. https://www.eastlothian.gov.uk/downloads/file/23471/air_quality_further_assessment_2014_-_further_assessment_of_air_quality_musselburgh.

[184] Targa, J. and A. Loader. Diffusion Tubes for Ambient NO_2 Monitoring: Practical Guidance for Laboratories and Users. *Report to DEFRA and the Devolved Administrations 1a*.

[185] The European Commission (2012). Regulation (EU) of the European Parliament and of the Council of 17 April 2019 setting CO_2 emission performance standards for new passenger cars and for new light commercial vehicles, and repealing Regulations (EC) No 443/2009 and (EU) No 510/2011. *Official Journal of the European Union*. https://www.eea.europa.eu/policy-documents/443-2009.

[186] The National Institute for Occupational Safety and Health (NIOSH) (2016). Ozone. *Centers for Disease Control and Prevention*. https://www.cdc.gov/niosh/npg/npgd0476.html.

[187] The Scottish Government (2015). CLEANER AIR FOR SCOTLAND: The Road To A Healthier Future. https://www.gov.scot/publications/cleaner-air-scotland-road-healthier-future/.

[188] The Scottish Government (2016). Air Quality in Scotland. http://www.gov.scot/Topics/Environment/waste-and-pollution/Pollution-1/16215.

[189] The Scottish Parliament (2020). What are the powers of the Scottish Parliament? *Visit & Learn: Scottish Parliament*. https://www.parliament.scot/about/how-parliament-works/powers-of-the-scottish-parliament.

[190] Titsias, M. K. (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. *International Conference on Artificial Intelligence and Statistics*, 567?574.

[191] Tonellato, S. F. (2002). A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50-2*.

[192] Transport & Environment (2022). Electric cars.

[193] Transport for London (2016). Improving buses. https://tfl.gov.uk/modes/buses/improving-buses.

[194] Tukey, J. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics 5*, 99–114.

[195] UN environment assembly (2017). Towards a Pollution-Free Planet: Background report. https://wedocs.unep.org/bitstream/handle/20.500.11822/21213/Towards_a_pollution_free_planet_advance%20version.pdf?sequence=2&isAllowed=y.

[196] Unpingco, J. (2019). Python for Probability, Statistics, and Machine Learning. *Springer*.

[197] U.S. Department of Commerce (2016). SURFRAD Aerosol Optical Depth. *National Oceanic and Atmospheric Administration*. http://www.esrl.noaa.gov/gmd/grad/surfrad/aod/.

[198] US Environmental Protection Agency (2016). Summary of the Clean Air Act. *US EPA*. https://www.epa.gov/laws-regulations/summary-clean-air-act.

[199] Vernon, I. (2016). Bayesian Statistics Applied to Complex Models of Physical Systems. *Washington University Workshop*.

[200] W. Date (2017). Views sought on "inner London" ULEZ. *airqualitynews.com*. https://airqualitynews.com/2014/12/23/views-sought-on-air-quality-planning-guidance/.

[201] Wald, L. (1999). Some Terms of Reference in Data Fusion. *IEEE Transactions on Geoscience and Remote Sensing 37:3*, 1190 – 1193.

[202] Wall, M. (2015). Hydrogen, hydrogen everywhere. *BBC News*. https://www.bbc.co.uk/news/business-31926995#:~:text=Hydrogen%20is%20the%20most%20abundant,to%20be%20the%20perfect%20fuel.

[203] Warren, J. et al. (2012). Bayesian Spatial-temporal Model for Cardiac Congenital Anomalies and Ambient Air Pollution Risk Assessment. *Environmetrics 23-8*.

[204] Wesseling, J. and K. van Velze (2015). Technische beschrijving van standaardrekenmethode 2 (SRM-2) voor luchtkwaliteitsberekeningen. *National Institute for Public Health and the Environment*.

[205] White, F. E. (1991). Data Fusion Lexicon. *Data Fusion Panel Joint Directories of Laboratories Technical Panel for C3*.

[206] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.

[207] Wikle, C. K., L. M. Berliner, and N. Cressie (1999). Hierarchical Bayesian space-timemodels. *Environmental and Ecological Statistics 5*.

[208] World Health Organization: Department of Public Health, Environmental and Social Determinants of Health (PHE) (2016). Ambient air pollution: A global assessment of exposure and burden of disease. https://apps.who.int/iris/handle/10665/250141.

[209] Wyatt, D. W., H. Li, and J. E. Tate (2014). The impact of road grade on carbon dioxide (CO_2) emission of a passenger vehicle in real-world driving. *Transportation Research Part D: Transport and Environment 32*, 160–170.

[210] Xie, W. et al. (2015). Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality. *Heart 101*.

[211] Yang, Y. et al. (2020). Impacts of aerosol-radiation interaction on meteorological forecasts over northern China by offline coupling of the WRF-Chem-simulated aerosol optical depth into WRF: a case study during a heavy pollution event. *Atmospheric Chemistry and Physics 20*, 12527–12547.

[212] Zhang, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *American Statistical Assosiation 99 (465)*, 250–261.