National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

Modelling reactive nitrogen concentrations and **deposition on a local scale**

Comparison of eight models and their ensembles with three measurement campaigns

RIVM report 2026-0041
E. Kooi et al.

Modelling reactive nitrogen concentrations and deposition on a local scale

Comparison of eight models and their ensembles with three measurement campaigns

RIVM report 2026-0041

Colophon

© RIVM 2026
Parts of this publication may be reproduced, provided the source is referenced as follows: National Institute for Public Health and the Environment (RIVM), along with the title of the publication and the year it was published.

DOI 10.21945/RIVM-2026 0041

E. Kooi (author), RIVM
G. Thorkelsdottir (author), RIVM
P. Meijer (author), RIVM
J. Stocker (author), CERC
W. Lefebvre (author), VITO
A. Vigier (author), UKCEH
A.S. Lansø (author), Aarhus University
M. Krol (author), Wageningen University & Research
M. Sterk (author), RIVM
C. Jacobs (author), RIVM
S. Jonkers (author), RIVM
A. van Pul (author), RIVM
R. Wichink Kruit (author), RIVM

Contact:
Eelke Kooi
Milieu & Veiligheid/Centrum Veiligheid
eelke.kooi@rivm.nl

This investigation was performed by order, and for the account, of the Ministry of Agriculture, Fisheries, Food Security and Nature, within the framework of Work Package 2.2 of the SAGEN project (M/360092/01/BL) of the National Nitrogen Knowledge Programme (NKS).

Published by:
National Institute for Public Health and the Environment, RIVM
PO Box 1 | 3720 BA Bilthoven
The Netherlands
www.rivm.nl/en

Synopsis

Modelling reactive nitrogen concentrations and deposition at a local scale

Comparison of eight models and their ensembles with three measurement campaigns

Every year, RIVM maps the amount of nitrogen in the air (concentration) and in the soil (deposition) in the Netherlands. It uses models and measurements for this. The government uses the same scientific models to estimate how much a particular source contributes to nitrogen deposition in nature reserves. Such model calculations are necessary before the source can be issued with a permit. This report discusses these calculations.

In some cases, models calculate a concentration or deposition for a specific area that is too high or too low. RIVM therefore investigated whether simultaneous use of multiple models (ensemble modelling) would result in more accurate outcomes and provide greater insight into uncertainties. To this end, RIVM compared the outcomes of models and ensembles of models with measurements. It did not study the possible drawbacks of using ensemble modelling to decide on the issue of permits.

For this study, eight scientific models were used. These models are used at the national or international level to calculate the concentration and deposition close to the source. Furthermore, measurements from three measurement campaigns were used: the first campaign monitored the concentration and deposition in the vicinity of two poultry houses, the second the concentration near a motorway and the third the concentration in the vicinity of industry. Most monitoring locations were near the source; the greatest distance was 570 metres.

The comparison shows that calculated concentrations in the air more closely matched the measured results than calculated depositions did. There are two reasons for this. First, calculating concentrations is simpler than calculating depositions. Second, the concentration measurements used were more accurate than the deposition measurements used.

Combinations of models (ensembles) matched measurements as closely as the best individual models did. Additionally, ensemble modelling provided more insight into the uncertainties of outcomes.

Keywords: nitrogen, deposition, concentration, model, measurement, ensemble modelling

Publiekssamenvatting

Modellering van reactieve stikstofconcentraties en -depositie op lokale schaal

Vergelijking van acht modellen en hun ensembles met drie meetcampagnes

Het RIVM brengt elk jaar in beeld hoeveel stikstof er in Nederland in de lucht zit (concentratie) en op de bodem terechtkomt (depositie). Hiervoor gebruikt het RIVM rekenmodellen en metingen. De overheid gebruikt de wetenschappelijke rekenmodellen ook om in te schatten hoe groot de bijdrage van een enkele bron is aan de stikstofdepositie in natuurgebieden. Die berekeningen zijn nodig om een vergunning aan te vragen. Dit rapport gaat over deze berekeningen.

Modellen kunnen de concentratie of depositie op een bepaalde plek te hoog of juist te laag berekenen. Het RIVM onderzocht daarom of het gebruik van meerdere modellen tegelijk (ensemblemodellering) de uitkomsten nauwkeuriger maakt en beter inzicht geeft in onzekerheden. Hiervoor vergeleek het RIVM uitkomsten van rekenmodellen en groepen rekenmodellen met metingen. De mogelijke nadelen van het gebruik van ensemblemodellering bij het aanvragen van vergunningen zijn niet onderzocht.

Voor dit onderzoek zijn acht wetenschappelijke modellen gebruikt. Deze worden nationaal of internationaal gebruikt om concentratie en depositie dicht bij de bron te berekenen. Verder zijn meetgegevens uit drie meetcampagnes gebruikt: één met concentratie- en depositiemetingen in de buurt van twee kippenstallen, één met concentratiemetingen bij een autosnelweg en één met concentratiemetingen in de buurt van industrie. De meeste meetlocaties lagen dicht bij de bron; de grootste afstand was 570 meter.

De vergelijking laat zien dat berekende luchtconcentraties beter overeenkomen met metingen dan berekende deposities. Dit heeft twee oorzaken. Ten eerste is de berekening van concentraties eenvoudiger dan die van depositie. Ten tweede waren de gebruikte concentratiemetingen nauwkeuriger dan de gebruikte depositiemetingen.

Combinaties van modellen (ensembles) kwamen even goed overeen met metingen als de beste individuele rekenmodellen. Verder geeft ensemblemodellering meer inzicht in de onzekerheden in uitkomsten.

Kernwoorden: stikstof, depositie, concentratie, rekenmodel, meting, ensemblemodellering

Contents

Summary

Introduction

Nitrogenous gases, including nitrogen oxides and ammonia, are emitted into the air by sources such as industry, motorised traffic, and livestock. Eventually, these gases can deposit on soil and natural ecosystems, a process known as nitrogen deposition. Excessive nitrogen deposition has adverse environmental impacts, by favouring certain species over others, changing species composition, and altering soil chemistry, which impacts nutrient availability. For example, plants that grow well on nitrogen-rich soil, such as grass and nettle, can suppress rare plants that benefit from nutrient-poor soil. Nitrogen compounds can also cause soil acidification, which may accelerate leaching of nutrients and disappearance of certain plant species.

In order to define effective policy measures for protecting vulnerable nature areas, it is important to know how much nitrogen deposits on nature areas. In the Netherlands, such estimates are generated through a combination of model calculations, air concentration measurements, and direct deposition measurements. These estimates are not only used for policy support but also to evaluate permit applications for economic activities with emissions of nitrogen compounds.

The current research is part of the SAGEN project.[1] The aim is to improve the accuracy of the current Dutch measurement and modelling system. In work package 2 of the project, the accuracy of computational models is investigated by comparing them with measurements, and it is examined whether the use of ensemble modelling offers advantages. Ensemble modelling is the use of multiple computational models for estimating parameters such as concentration and deposition. The underlying idea is that the average outcome of multiple models is more accurate than the outcome of a single model. This is the case, for example, if some models provide estimates that are too high, and other models yield estimates that are too low. Another expected advantage of ensemble modelling is that the spread of outcomes of individual models provides insight into the uncertainty of the ensemble outcome. Furthermore, the use of ensemble modelling could also promote the exchange of knowledge between model developers, so that model improvements can be realised more quickly.

Work package 2 of the SAGEN project is split into two parts, one part relating to regional scale and another part to local scale. This report is about the second part. Local scale concerns distances ranging from one hundred meters to several kilometres from an individual source. This scale is especially important for permit applications for industrial and agricultural activities.

[1] SAGEN stands for: Research programme on the use of satellite data and ensemble modelling. It is part of the Dutch National Nitrogen Knowledge Programme.

Objective and scope of this study

For the development and implementation of policy, it is important that underlying calculation results are sufficiently accurate and that the uncertainty in the results is also known in sufficient detail. The programme plan of the SAGEN project expects model ensembles to provide more accurate outcomes than an individual model, and to provide more insight into the uncertainties in outcomes.

In this study, these two assumptions are tested by comparing calculated concentrations and deposition fluxes in the vicinity of an individual source with measurements. This results in the following research questions:

1. How do the results of individual calculation models for air quality and nitrogen deposition compare with measurements close to a source?
2. How do outcomes of model ensembles compare with those same measurements?
3. Can the uncertainty of ensemble outcomes be derived from differences between individual model outcomes?

The spatial scale is important for model calculations. Models that are accurate at short distances from the source are not necessarily accurate on larger scales, and vice versa. Therefore, the SAGEN project distinguishes between local scale and regional scale. This research focuses on the local scale, roughly from a hundred metres to a few kilometres away from an individual source, and the calculation models used for this scale.

The accuracy of models and model ensembles is determined by comparing them with measurements. The measurements must be sufficiently accurate for this. In this study, measurements in the immediate vicinity of a chicken farm, a motorway and a compressor station were used. The report does not provide insight into the accuracy of models and model ensembles for other types of emission sources, nor into the accuracy at more than a few hundred meters away. The underlying causes of differences between models have not been investigated due to budgetary constraints and a lack of detailed output data.

Selection of measurement campaigns

Data from three measurement campaigns was used to evaluate model outcomes:

(i) NH_3 concentration and NH_x deposition measurements around a chicken farm in Ringsted, Denmark,
(ii) NO_2 concentration measurements in the vicinity of a highway near Affligem, Belgium,
(iii) NO_2 and NO_x concentration measurements near a compressor station in Balko, United States.

For various reasons, no useful NH_x or NO_y deposition measurements near traffic sources and industrial sources were found.

Table A Summary of the selected measurement campaigns and available measurements

Measurement campaign	**Short description**
Ringsted, Denmark [6]	NH_3 concentration measurements at 15 separate locations in the vicinity of a farm with two chicken houses. Multiple measurement periods of 10 to 17 days. In total 27 individual measurements. Measurement locations mostly between 25 m and 200 m distance from the source. Furthest distance: 570 m. Indirect NH_x deposition measurements at 25 separate locations around the same source. Values for 16 locations used for model validation, the remaining 9 values were insufficiently accurate, according to the authors of the study [6].
Affligem, Belgium [16]	NO_2 concentration measurements at six separate locations near a highway. Distances ranging from 6 to 146 m. Weekly measurements for 36 consecutive weeks, translated by RIVM into average concentrations over the entire 36-week period.
Balko, Oklahoma, United States [20]	NO_2 and NO_x concentration measurements at four different locations around a compressor station. Distances ranging from 100 to 425 m from the dominant source. Hourly measurements for 13 consecutive months, translated by RIVM into average concentrations over the entire 13-month period.

Most measurement locations are between 5 and 200 meters away from the source (Table A). The deposition measurements in Ringsted have previously been used by Sommer et al. [6] to validate the deposition modelling in OML. Analyses during the current study indicate that the accuracy of these measurements was lower than expected.

Selection of models and ensembles

Eight atmospheric transport models were used for the study, all of which are regularly used for air quality and deposition calculations on a local scale: ADMS, AERMOD, IFDM, OML-Multi, OPS-ST, OPS-LT, SRM2, and STACKS-D. The first five models calculate the dispersion and deposition for each individual hour. The other three models were primarily developed to calculate annual average concentrations and deposition values, and do not calculate on an hourly basis. SRM2 was only used for comparison with the measurement campaign near the highway, while OML-Multi was not used for that campaign. The same eight models have previously been used in a model intercomparison study in which differences in outcomes were investigated [2].

To investigate the potential benefits of ensemble modelling, three model ensembles were defined for each measurement campaign:
group 1: three hourly models,
group 2: all hourly models,
group 3: all models.

This approach allowed investigating the influence of the size of the ensemble on the accuracy. The ensemble outcome is an average value of the outcomes of the underlying individual models. In this study, the arithmetic mean and the geometric mean were examined. The manner of averaging has little influence on the results.

Analysis of the accuracy of model results

Total period average concentrations were used for the Affligem and Balko campaigns, despite shorter duration measurements being available. The motivation is that policy decisions about deposition are generally based on annual average deposition. It is, then, important to know how accurately models calculate annual average deposition and concentrations. Relatedly, some models were not designed to provide hourly (Balko) or weekly (Affligem) average values.

The results of the eight models were compared with measurements at different locations. Then, differences were summarised using five model performance indicators:

- Fractional Bias (FB)
- Geometric Mean Bias (MG)
- Normalised Mean Square Error (NMSE)
- Geometric Variance (VG)
- Fraction of model outcomes within a factor of two of the measured values (FAC2).

FB and MG are measures of bias and indicate whether calculated values are on average higher or lower than measured values. NMSE and VG are values for the mean difference between calculated and measured values (the mean 'error'). Hanna & Chang's acceptance criteria [39] were used to assess the performance indicators. These are of limited use for the concentration measurements of this study. There are no acceptance criteria for deposition yet.

Results of the comparison with measurements

The results of the comparison with measurements are summarised in Figure A, Figure B and Table B. Results for OPS-LT relate to the 2024 version, unless otherwise stated.

Figure A shows the Geometric Mean Bias (MG) and the Geometric Variance (VG) for all models, ensembles, and datasets (measurement campaigns). This figure is frequently used to summarise results of model validation studies. MG represents the average bias of the model compared with the measurements. VG is a quadratic measure of the deviation between model results and measured values. The black line represents the contribution by systematic bias (MG) to VG. Points that lie well above the black line have a relatively large amount of 'random scatter' on top of the systematic bias.

Figure A Model performance in terms of Geometric Mean Bias (MG) and Geometric Variance (VG) regarding individual models (coloured points) and model ensembles (open symbols). Negative outcomes for Ringsted deposition are excluded. The result for STACKS-D regarding Ringsted deposition measurements is outside of the plot range. Dotted lines represent average deviations by a factor of two.

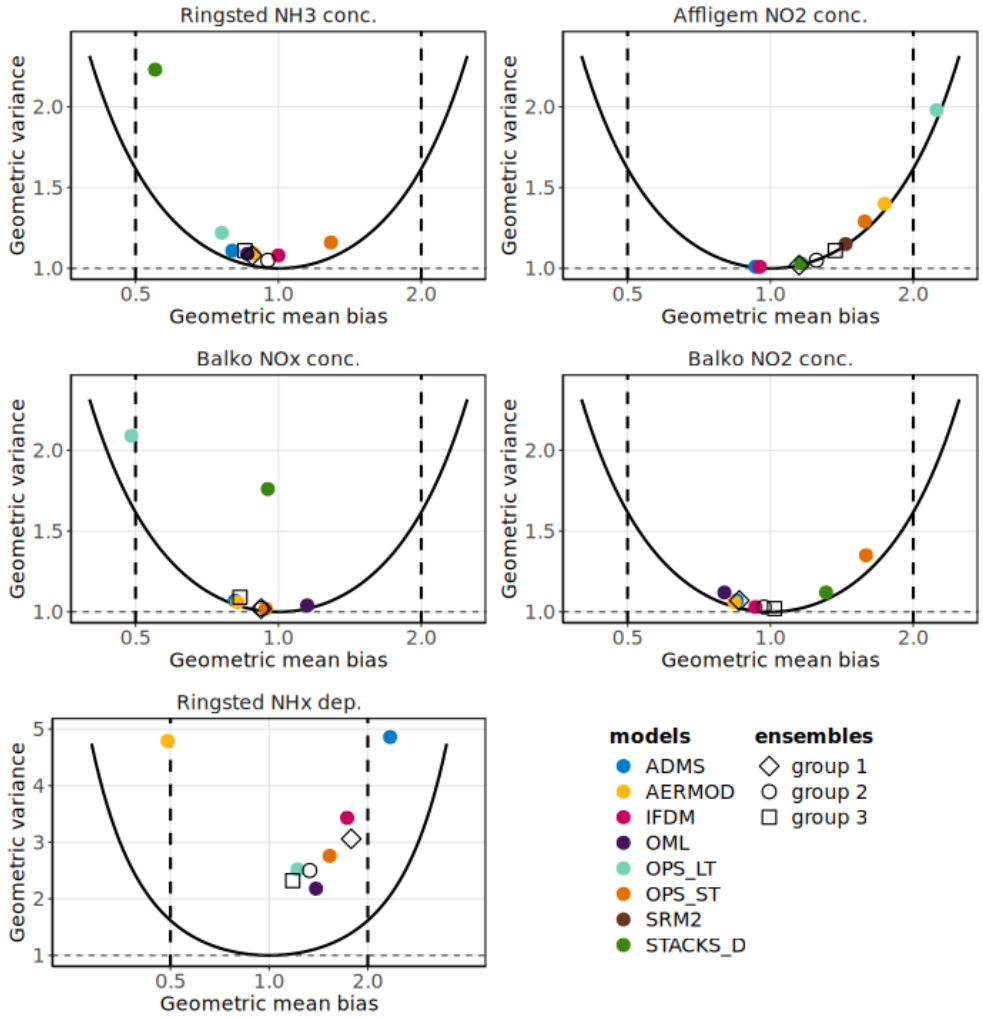
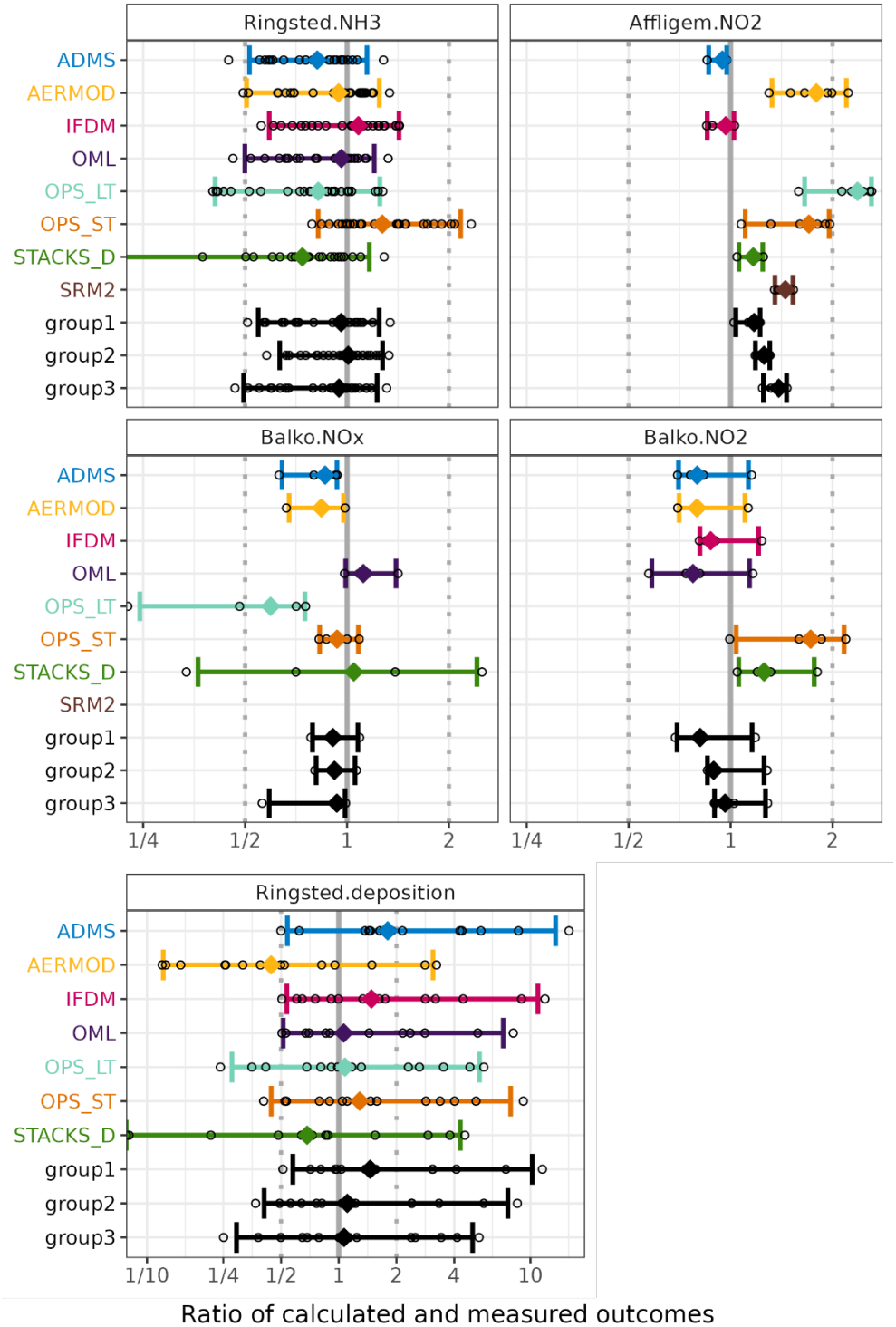


Figure B shows the ratio of calculated and measured values with an error bar ranging from the 2.5 to the 97.5 percentile. The ratios for individual measurements are indicated by open circles. The error bar reaches approximately from the smallest ratio to the largest ratio.

Figure B Ratio of calculated and measured outcomes (concentration or deposition) for all individual models and model ensembles. The error bar ranges from 2.5 to 97.5 percentile. Open circles represent the ratio of calculated and measured values for individual measuring points. The diamond is the median of these ratios. The x-axis is logarithmic. A different scale was used for the Ringsted deposition. Negative deposition measurements for that campaign are excluded.



With regard to the three measurement campaigns with concentration measurements, the two figures show the following:

- IFDM, OML-Multi, and ADMS had consistently good scores across all considered concentration measurements.[2] These models, together with AERMOD, are the models that calculate air dispersion for each individual hour, and take hourly variations in emissions into account in great detail.
- SRM2, OPS-ST, STACKS-D, and OPS-LT had worse results than the other individual models across the four datasets. SRM2, STACKS-D, and OPS-LT were primarily developed for calculating annual average concentration and deposition patterns. The current study shows that the outcomes for these models are less accurate than the outcomes for hourly models if emission strengths vary greatly over time, if measurement durations are much shorter than one year, and/or if measurements are carried out at a short distance from a source.
- Across the three campaigns with concentration measurements, the ensemble of all hourly models (group 2) performed as well as IFDM and better than all other individual models. The other ensembles considered (group 1 and group 3) performed slightly worse, but still better than most individual models.

As regards OPS-LT, the modelling of NO_2 concentrations near roads has recently been updated. The latest version of 2025 uses distance-dependent NO/NO_2 ratios and takes into account the increased turbulence caused by highway traffic. The results of the new version of OPS-LT for the motorway measurement campaign were as good as the best other individual models for that campaign.

With regard to the deposition measurements in the Ringsted campaign, the following conclusions are drawn on the basis of the two figures:

- The discrepancies between measured and calculated outcomes were significantly larger than for the concentration measurements. This is expected to be because (i) deposition modelling is inherently more complex and (ii) these deposition measurements are (much) less accurate than the concentration measurements. Because the error in the measurements is not well known, it is not possible to indicate how large the influence is on calculated differences between measurements and model results.
- Among the individual models, OML-Multi and OPS-LT had the best results. The scores for the two larger model ensembles (groups 2 and 3) are similar to those of OML-Multi and OPS-LT. All three ensembles had better results than most individual models.

Calculated performance indicators are often compared in the literature with acceptance criteria from Hanna & Chang for measurements in rural or urban settings [39]. These are proposals by Hanna and Chang for certain controlled field studies with concentration measurements. The proposals have no official status. Furthermore, the suitability of the

[2] IFDM wasn't compared with Balko NO_x concentration measurements, and OML wasn't compared with Affligem NO_2 concentration measurements.

criteria for other types of measurements, including those used in this study, is unknown. In the absence of a better alternative, it was examined for the concentration measurements in this study, how many indicators meet the Hanna & Chang criteria for rural environments (Table B). The results do not indicate whether or not a model is suitable for regular model applications.

Table B Number of indicators out of a total of five that meet the acceptance criteria used

Model	**Ringsted**	**Affligem**	**Balko**
ADMS	4/5	5/5	5/5 for NO_x 5/5 for NO_2
AERMOD	5/5	2/5	5/5 for NO_x 5/5 for NO_2
IFDM	5/5	5/5	NO_x not run 5/5 for NO_2
OML-Multi	5/5	-	5/5 for NO_x 5/5 for NO_2
OPS-LT	4/5	1/5 (2024 version) 5/5 (2025 version)	2/5 for NO_x NO_2 not run
OPS-ST	4/5	3/5	5/5 for NO_x 2/5 for NO_2
SRM2	-	3/5	-
STACKS-D	2/5	5/5	4/5 for NO_x 4/5 for NO_2
Group 1: three hourly models	5/5	5/5	5/5 for NO_x 5/5 for NO_2
Group 2: all hourly models	5/5	5/5	5/5 for NO_x 5/5 for NO_2
Group 3: all models	4/5	3/5	4/5 for NO_x 5/5 for NO_2

Benefits of ensemble modelling

One important goal of this research is to determine whether the use of ensemble modelling offers advantages. Several potential benefits were listed in the SAGEN programme plan [1], based on experiences with ensemble modelling in air quality modelling. The most important potential benefits were investigated in this study.

Accuracy
Model ensembles may be more accurate than results from individual models. In specific concentration measurement campaigns, the model ensembles often performed about as well as the better individual models (IFDM, OML-Multi and ADMS). For the Ringsted deposition measurements, the results for the two larger ensembles (groups 2 and 3) were similar to those of OML-Multi and OPS-LT. Overall, then, the two larger ensembles seem to perform better than most individual models, but not better than each individual model.

Insight into uncertainties
Ensemble modelling may provide insight into uncertainties. This assumes that the uncertainty in the ensemble outcome can be derived from the variation in the individual model outcomes. For the two larger

ensembles, deviations from measurements were indeed about as large as the differences between individual model outcomes. For the smallest ensemble, however, discrepancies with measurements were sometimes much larger than differences between individual model outcomes; the latter thus not being a good measure of uncertainty.

Exchange of knowledge
Ensemble modelling may lead to more exchange of knowledge between model developers. Model improvements can therefore be realised more quickly. The experience of this project is that the exchange of knowledge is mostly contingent on available time. It requires that causes of differences can be investigated and addressed. Causes of differences can only be found by comparing model formulations and by studying outcomes, including intermediate results, at a detailed level. As such, an automatic interface for running models in parallel is not sufficient to identify and understand differences between models. However, ensemble modelling could be a way to formalise collaborations between institutions and obtain structural funding for knowledge exchange and model development. Exchange of knowledge can also be stimulated by regularly conducting benchmark studies involving multiple organisations.

Possible disadvantages
Possible disadvantages of ensemble odelling were not investigated in this study. These include the added financial costs for aligning input, developing a software interface and keeping that up-to-date, and the additionally required computer power (or longer durations of model runs).

Conclusions

In this study, outcomes of individual models and model ensembles were compared with measurements in three measurement campaigns. Finding useful deposition measurements in the immediate vicinity of a source proved to be particularly difficult: most deposition measurements are carried out in nature reserves at a (too) great distance from individual sources. When measurements are carried out in the vicinity of a source, the duration of the measurements is often limited, or the contribution of dry deposition is not properly considered. In the Ringsted campaign, deposition in the vicinity of a chicken farm was determined on the basis of isotope measurements in biomonitors. However, these (indirect) deposition measurements were less accurate than had been expected.

For the concentration measurements in Ringsted, Affligem, and Balko, IFDM, OML-Multi, and ADMS had consistently good results. The results for these models also met all or almost all of the acceptance criteria that Hanna & Chang proposed for concentration measurements in rural campaigns. The other models had good results for some campaigns and less good results for others. In addition, acceptance criteria for individual model performance indicators proposed by Hanna & Chang were more often not fulfilled.

For the Ringsted deposition measurements, the differences between model results and measurements were greater than for concentration calculations. This is probably due to (i) deposition modelling being inherently more complex than concentration modelling and (ii) these deposition measurements being (much) less accurate than the concentration measurements that were used in this study. Out of the individual models, OML-Multi and OPS-LT had the best agreement with measured values. However, the measurements were not sufficiently accurate to draw strong conclusions.

Overall, model ensembles performed equally well as the better individual models for individual campaigns, but not better. Between the considered ensembles, the ensemble of five hourly models received the best results. Furthermore, ensemble modelling allows for estimating the uncertainty in the outcomes from the spread in underlying individual model outcomes, if the ensemble is not too small.[3] Potential disadvantages of ensemble modelling have not been investigated.

The above conclusions relate to the measurements used in this study. In particular, most of the measurement locations were 5 to 200 m away from the source. Therefore, this research does not provide insight into the accuracy of the models and model ensembles at (much) greater distances.

Recommendations

This research showed that some models correspond better with certain measurements than other models. RIVM recommends that the causes of these differences be further analysed in a follow-up study. This can be done by looking at model formulations and intermediate model results in detail. Then, the models that deviate the most from the measurements can be improved using the knowledge gained. The measurement campaigns from the current study and the underlying input data for the model calculations are suited to be a basis for such research.

The study also showed that there are hardly any reliable measurements available to validate calculated deposition fluxes in the vicinity of a source. This is because measurement methods, particularly for dry deposition, have important requirements. RIVM recommends investigating what is needed to set up a reliable measurement campaign with deposition measurements near an individual source.

[3] For the smallest ensemble, consisting of three models, the spread of model outcomes was not always a good indicator of ensemble outcome uncertainty.

Samenvatting

Inleiding

Verschillende bronnen zoals industrie, gemotoriseerd verkeer en vee, stoten stikstofhoudende gassen uit naar de lucht. Hiervan komen stikstofoxiden en ammoniak uiteindelijk weer op de grond neer. Dit noemen we stikstofdepositie. Te veel stikstofdepositie is slecht voor de natuur. Planten die goed groeien op stikstofrijke grond, zoals grassen en brandnetels, kunnen zeldzame planten verdringen die juist een voedingsarme bodem nodig hebben. Ook kunnen stikstofverbindingen zorgen voor verzuring van de bodem waardoor voedingsstoffen sneller uitspoelen en bepaalde soorten planten kunnen verdwijnen.

Het is belangrijk om te weten hoeveel stikstofdepositie er is. Dat is nodig om effectieve beleidsmaatregelen te nemen die kwetsbare natuurgebieden beschermen. In Nederland bepalen we de stikstofdepositie met een combinatie van modelberekeningen, luchtconcentratiemetingen en directe depositiemetingen. De uitkomsten worden gebruikt voor beleidsondersteuning en de beoordeling van vergunningaanvragen voor economische activiteiten die stikstofverbindingen uitstoten.

Het huidige onderzoek is onderdeel van het SAGEN-project.[4] Het doel daarvan is de nauwkeurigheid van het huidige Nederlandse systeem van meten en modelleren te verbeteren. In werkpakket 2 wordt de nauwkeurigheid van rekenmodellen onderzocht door ze te vergelijken met metingen. Ook wordt bekeken of het gebruik van ensemblemodellering voordelen biedt. Ensemblemodellering is het gebruik van meerdere rekenmodellen voor het schatten van parameters zoals concentratie en depositie. Het achterliggende idee is dat de gemiddelde uitkomst van meerdere modellen nauwkeuriger is dan de uitkomst van een enkel model. Dit is bijvoorbeeld het geval als een deel van de modellen te hoge schattingen geeft en een ander deel te lage schattingen. Een ander verwacht voordeel van ensemblemodellering is dat de spreiding van de uitkomsten van individuele rekenmodellen inzicht geeft in de onzekerheid van het ensembleresultaat. Verder zou het toepassen van ensemblemodellering mogelijk ook het uitwisselen van kennis tussen modelontwikkelaars bevorderen, zodat modelverbeteringen sneller gerealiseerd kunnen worden.

Werkpakket 2 van het SAGEN-project is opgesplitst in twee deelonderzoeken: een voor de regionale schaal en een voor de lokale schaal. Dit rapport gaat over het tweede deelonderzoek. Lokale schaal betreft afstanden van honderd meter tot enkele kilometers vanaf een individuele bron. Deze schaal is vooral van belang bij vergunning-aanvragen voor industriële en agrarische activiteiten.

[4] SAGEN staat voor Satellietgebruik en Ensemble en is onderdeel van het Nationaal Kennisprogramma Stikstof (NKS).

Doelstelling en reikwijdte van dit onderzoek

Voor het ontwikkelen en uitvoeren van beleid is het van belang dat onderliggende rekenresultaten voldoende nauwkeurig zijn en dat ook de onzekerheid in de uitkomsten in voldoende mate bekend is. In het programmaplan van het SAGEN-project is de verwachting uitgesproken dat modelensembles nauwkeurigere uitkomsten geven dan een individueel model, en meer inzicht bieden in de onzekerheden in uitkomsten.

In dit onderzoek worden deze twee aannames getoetst door berekende concentraties en depositiefluxen dicht bij een individuele bron te vergelijken met metingen. Dat leidt tot de volgende onderzoeksvragen:

1. Hoe verhouden uitkomsten van individuele rekenmodellen voor luchtkwaliteit en stikstofdepositie zich tot metingen dicht bij een bron?
2. Hoe verhouden uitkomsten van modelensembles zich tot metingen dicht bij een bron?
3. Kan de onzekerheid van ensemble-uitkomsten worden afgeleid uit de verschillen tussen individuele modeluitkomsten?

De ruimtelijke schaal is voor modelberekeningen belangrijk. Modellen die op korte afstanden van de bron nauwkeurig zijn, zijn dat niet per se op grotere schaal en andersom. Daarom maakt het SAGEN-project onderscheid tussen lokale en regionale schaal. Dit onderzoek richt zich op de lokale schaal, ruwweg van honderd meter tot enkele kilometers afstand van een individuele bron, en de rekenmodellen die voor deze schaal gebruikt worden.

De nauwkeurigheid van modellen en modelensembles wordt bepaald door ze te vergelijken met metingen. De metingen moeten hiervoor wel voldoende nauwkeurig zijn. In deze studie zijn metingen in de directe omgeving van een kippenboerderij, een autosnelweg en een compressorstation gebruikt. Het rapport geeft geen inzicht in de nauwkeurigheid van modellen en modelensembles voor andere soorten emissiebronnen en ook niet in de nauwkeurigheid op meer dan enkele honderden meters afstand. De onderliggende oorzaken van verschillen tussen modellen zijn vanwege budgettaire beperkingen en een gebrek aan gedetailleerde uitvoergegevens niet onderzocht.

Selectie van meetcampagnes

Gegevens uit drie meetcampagnes zijn gebruikt om de modelresultaten te evalueren:

(i) NH_3 concentratie- en NH_x depositiemetingen rond een kippenboerderij in Ringsted, Denemarken,
(ii) NO_2-concentratiemetingen in de buurt van een snelweg bij Affligem, België,
(iii) NO_2- en NO_x-metingen in de buurt van een compressorstation in Balko, Verenigde Staten.

Om verschillende redenen werden er geen bruikbare NH_x- of NO_y-depositiemetingen in de buurt van verkeersbronnen en industriële bronnen gevonden.

Tabel A Samenvatting van de geselecteerde meetcampagnes en de beschikbare metingen

Meetcampagne	**Korte omschrijving**
Ringsted, Denemarken [6]	NH3 concentratiemetingen op 15 verschillende locaties in de nabijheid van een boerderij met twee kippenstallen. Meerdere meetperiodes van 10 tot 17 dagen. In totaal 27 individuele metingen. Meetlocaties hoofdzakelijk tussen 25 m en 200 m afstand tot de bron. Verste afstand: 570 m. Indirecte NH_x depositiemetingen op 25 verschillende locaties rond dezelfde bron. Waarden voor 16 locaties zijn gebruikt voor modelvalidatie, de overige 9 waarden waren volgens de auteurs van de studie [6] onvoldoende nauwkeurig.
Affligem, België [16]	NO_2-concentratiemetingen op zes verschillende locaties nabij een snelweg. Afstanden variërend van 6 tot 146 m. Wekelijkse metingen gedurende 36 opeenvolgende weken, door het RIVM vertaald naar gemiddelde concentraties over de hele periode van 36 weken.
Balko, Oklahoma, Verenigde Staten [20]	NO_2- en NO_x-concentratiemetingen op vier verschillende locaties rond een compressorstation. Afstanden variërend van 100 tot 425 m van de dominante bron. Uurlijkse metingen gedurende 13 opeenvolgende maanden, door het RIVM vertaald naar gemiddelde concentraties over de hele periode van 13 maanden.

De meeste meetlocaties liggen tussen 5 en 200 meter afstand van de bron (Tabel A). De depositiemetingen in Ringsted zijn eerder door Sommer et al. [6] gebruikt om de depositiemodellering in OML te valideren. Analyses tijdens het huidige onderzoek wijzen erop dat de nauwkeurigheid van deze metingen lager was dan verwacht.

Selectie van modellen en ensembles

Voor het onderzoek zijn acht luchttransportmodellen gebruikt die regelmatig gebruikt worden voor luchtkwaliteits- en depositie-berekeningen op lokale schaal: ADMS, AERMOD, IFDM, OML-Multi, OPS-ST, OPS-LT, SRM2 en STACKS-D. De eerste vijf modellen berekenen de verspreiding en depositie voor elk afzonderlijk uur. De overige drie modellen zijn primair ontwikkeld om jaargemiddelde concentraties en depositiewaarden te berekenen. Deze rekenen niet van uur tot uur. SRM2 is alleen gebruikt voor vergelijking met de meetcampagne bij de snelweg, terwijl OML-Multi voor die meetcampagne juist niet is gebruikt. Dezelfde acht modellen zijn eerder gebruikt in een modelvergelijkings-studie waarbij onderlinge verschillen in uitkomsten in kaart zijn gebracht [2].

Om de mogelijke voordelen van ensemblemodellering te onderzoeken zijn voor elke meetcampagne drie modelensembles gedefinieerd:

Groep 1: drie uurlijkse modellen;
Groep 2: alle uurlijkse modellen
Groep 3: alle modellen

Deze aanpak maakte het mogelijk om te onderzoeken welke invloed de grootte van het ensemble heeft op de nauwkeurigheid. De ensemble-uitkomst is een gemiddelde waarde van de uitkomsten van de individuele modellen. In dit onderzoek zijn het rekenkundig gemiddelde en het meetkundig gemiddelde bekeken. De manier van middelen heeft weinig invloed op de resultaten.

Analyse van de nauwkeurigheid van modeluitkomsten

Voor de meetcampagnes in Affligem en Balko is gebruikgemaakt van de gemiddelde concentraties over de totale duur van de campagne, ook al waren er metingen voor een kortere tijdsduur beschikbaar. De motivatie hiervoor is dat beleidsbeslissingen over depositie doorgaans gebaseerd zijn op de gemiddelde jaarlijkse depositie. Het is dan belangrijk om te weten hoe nauwkeurig de berekening is van de gemiddelde jaarlijkse depositie en concentraties. Een deel van de gebruikte modellen is ook niet ontworpen om gemiddelde waarden per uur (Balko) of per week (Affligem) te leveren.

De uitkomsten van de acht rekenmodellen zijn per meetlocatie vergeleken met de metingen. Vervolgens zijn de verschillen samengevat met vijf modelprestatie-indicatoren:

- Fractional Bias (FB)
- Geometric Mean Bias (MG)
- Normalised Mean Square Error (NMSE)
- Geometric Variance (VG)
- Fractie rekenresultaten binnen een factor twee van de meting (FAC2).

FB en MG zijn maten voor de ‘bias’ en geven aan of berekende waarden gemiddeld hoger of lager zijn dan meetwaarden. NMSE en VG zijn waarden voor het gemiddelde verschil tussen berekende en gemeten waarden (de gemiddelde ‘fout’). Voor het beoordelen van de prestatie-indicatoren zijn acceptatiecriteria van Hanna & Chang [39] gebruikt. Deze zijn beperkt toepasbaar voor de concentratiemetingen van dit onderzoek. Voor depositie bestaan nog geen acceptatiecriteria.

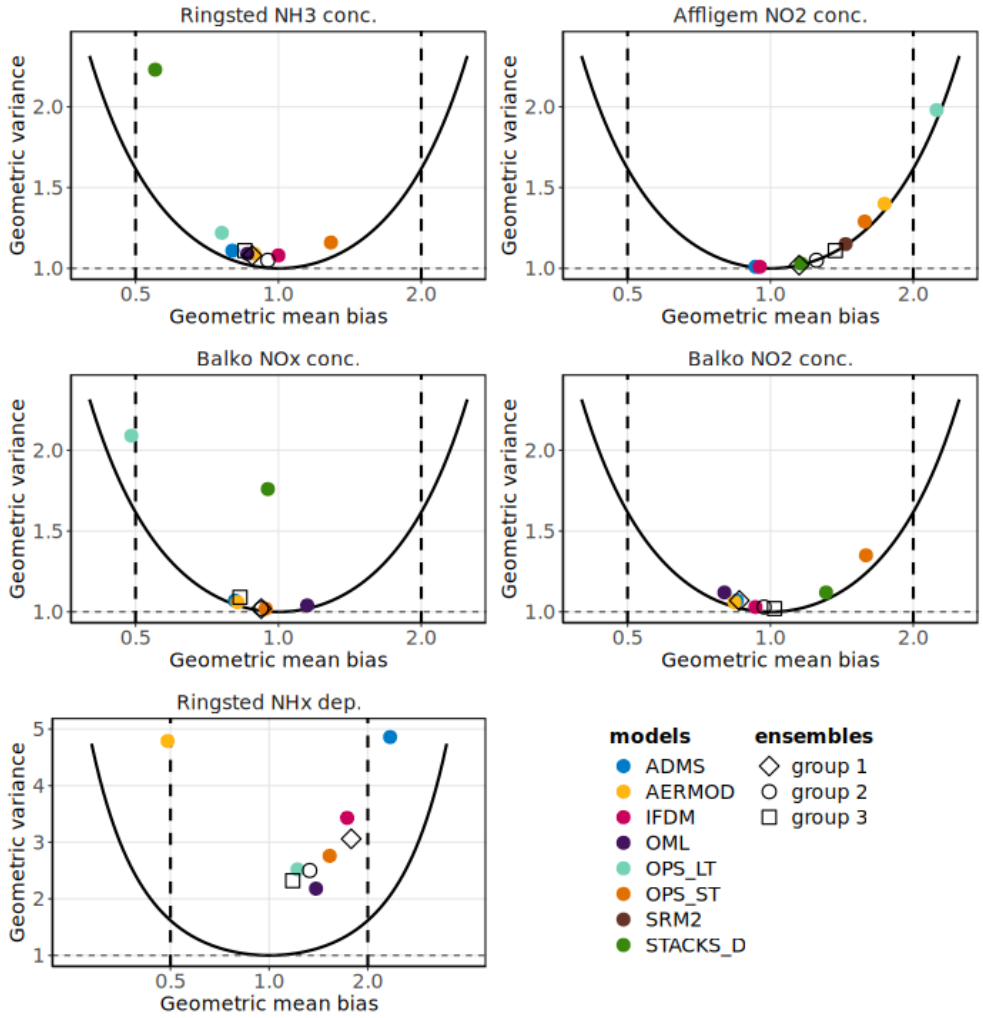
Resultaten van de vergelijking met metingen

De resultaten van de vergelijking met metingen zijn samengevat in Figuur A, Figuur B en Tabel B. De resultaten voor OPS-LT hebben betrekking op de versie van 2024, tenzij anders is vermeld.

Figuur A geeft de Geometric Mean Bias (MG) en de Geometric Variance (VG) weer voor alle modellen, ensembles en meetcampagnes. Deze figuur wordt vaak gebruikt om resultaten van modelvalidatiestudies samen te vatten. MG is de geometrisch gemiddelde bias van het model ten opzichte van de metingen. VG is een kwadratische maat voor de afwijking tussen modeluitkomsten en meetwaarden. De zwarte lijn laat

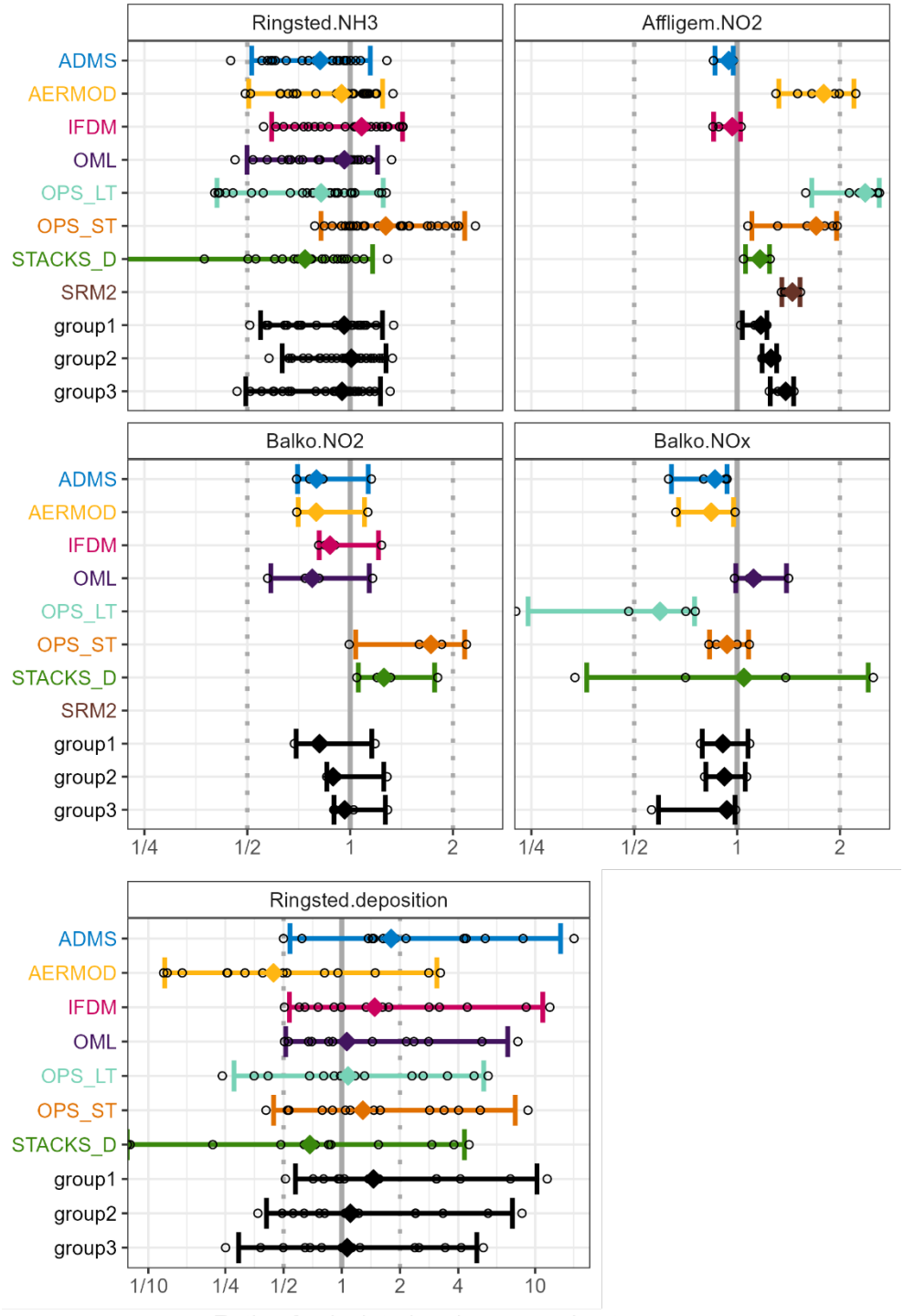
de bijdrage van de systematische afwijking (MG) aan de totale afwijking (VG) zien. Punten die ruim boven de zwarte lijn liggen, hebben verhoudingsgewijs veel ‘random scatter’ boven op de systematische afwijking.

Figuur A Modelprestaties in termen van Geometric Mean Bias (MG) en Geometric Variance (VG) wat betreft de individuele rekenmodellen (gekleurde rondjes) en modelensembles (open symbolen). Voor de Ringsted-depositie zijn negatieve metingen buiten beschouwing gelaten. Voor STACKS-D valt het resultaat voor de Ringsted-depositiemetingen buiten beeld. Gestippelde lijnen geven een gemiddelde afwijking weer van een factor twee.



Figuur B geeft de verhouding van berekende en gemeten waarden weer met een foutbalk die loopt van 2,5 tot 97,5 percentiel. De verhoudingen voor individuele meetpunten zijn weergegeven met open rondjes. De foutbalk reikt bij benadering van de kleinste verhouding tot de grootste verhouding.

Figuur B Verhouding van berekende en gemeten uitkomsten (concentratie of depositie) voor alle individuele modellen en modelensembles. De foutbalk loopt van 2,5 tot 97,5 percentiel. Open rondjes geven de verhouding weer van berekende en gemeten waarden voor individuele meetpunten. De ruit is de mediaan van deze verhoudingen. De x-as is logaritmisch. Voor Ringsted-depositie is een andere schaalverdeling gebruikt. Negatieve depositiemetingen in die campagne zijn niet meegenomen.



Met betrekking tot de drie meetcampagnes met concentratiemetingen laten de twee figuren het volgende zien:

- IFDM, OML-Multi en ADMS hadden consistent goede scores voor alle beschouwde concentratiemetingen.[5] Samen met AERMOD, zijn dit ook de modellen die de luchtverspreiding voor elk uur specifiek berekenen, en uurlijkse variatie van emissies in hoog detail meenemen.
- SRM2, OPS-ST, STACKS-D en OPS-LT hadden over de drie meetcampagnes heen beschouwd slechtere resultaten dan de andere individuele modellen. SRM2, STACKS-D en OPS-LT zijn in de eerste plaats ontwikkeld voor het berekenen van jaarlijkse gemiddelde concentratie- en depositiepatronen. Uit het huidige onderzoek blijkt dat, als emissies sterk variëren in de tijd, meetperiodes veel korter zijn dan een jaar en/of de metingen op korte afstand van een bron zijn gedaan, de resultaten van deze modellen minder nauwkeurig zijn dan van uurlijkse modellen.
- Over de drie meetcampagnes met concentratiemetingen tezamen presteerde het ensemble van alle uurlijkse modellen (groep 2) even goed als IFDM en beter dan alle andere individuele modellen. De andere beschouwde ensembles (groep 1 en groep 3) presteerden iets slechter, maar nog wel beter dan de meeste individuele modellen.

Wat betreft OPS-LT is de modellering van NO_2-concentraties nabij wegen onlangs bijgewerkt. De nieuwste versie van 2025 gebruikt afstandsafhankelijke NO/NO_2-verhoudingen en houdt rekening met de toegenomen turbulentie door snelwegverkeer. De resultaten van de nieuwe versie van OPS-LT waren voor de meetcampagne bij de snelweg net zo goed als de beste andere individuele modellen voor die campagne.

Met betrekking tot de depositiemetingen in de Ringsted-campagne wordt op basis van de twee figuren het volgende geconcludeerd:

- De afwijkingen tussen gemeten en berekende uitkomsten waren beduidend groter dan voor de concentratiemetingen. Dit komt naar verwachting doordat (i) depositiemodellering inherent complexer is en (ii) deze depositiemetingen (veel) minder nauwkeurig zijn dan de concentratiemetingen. Omdat de fout in de metingen niet goed bekend is, kan niet aangegeven worden hoe groot de invloed daarvan is op berekende verschillen tussen metingen en modeluitkomsten.
- Van de individuele modellen hadden OML-Multi en OPS-LT de beste resultaten. De scores voor de twee grotere modelensembles (groepen 2 en 3) zijn vergelijkbaar met die van OML-Multi en OPS-LT. Alle drie de ensembles hadden betere resultaten dan de meeste individuele modellen.

Berekende prestatie-indicatoren worden in de literatuur vaak vergeleken met acceptatiecriteria van Hanna en Chang voor metingen in landelijke of stedelijke omgevingen [39]. Het betreft voorstellen van Hanna en Chang voor bepaalde gecontroleerde veldstudies met concentratie-

[5] IFDM is niet vergeleken met Balko NO_x concentratiemetingen, en OML is niet vergeleken met Affligem NO_2 concentratie metingen.

metingen. De voorstellen hebben geen officiële status. Verder is de geschiktheid van de criteria voor andere soorten metingen, inclusief de metingen die in deze studie gebruikt zijn, onbekend. Bij gebrek aan een beter alternatief is voor de concentratiemetingen in deze studie bekeken hoeveel indicatoren voldoen aan de criteria van Hanna en Chang voor landelijke omgevingen (Tabel B). De uitkomsten geven niet aan of een model wel of niet geschikt is voor reguliere modeltoepassingen.

Tabel B Aantal indicatoren op een totaal van vijf dat aan de gebruikte acceptatiecriteria voldoet

Model	**Ringsted**	**Affligem**	**Balko**
ADMS	4/5	5/5	5/5 voor NO_x 5/5 voor NO_2
AERMOD	5/5	2/5	5/5 voor NO_x 5/5 voor NO_2
IFDM	5/5	5/5	NO_x niet doorgerekend 5/5 voor NO_2
OML-Multi	5/5	-	5/5 voor NO_x 5/5 voor NO_2
OPS-LT	4/5	1/5 (2024 versie) 5/5 (2025 versie)	2/5 voor NO_x NO_2 niet doorgerekend
OPS-ST	4/5	3/5	5/5 voor NO_x 2/5 voor NO_2
SRM2	-	3/5	-
STACKS-D	2/5	5/5	4/5 voor NO_x 4/5 voor NO_2
Groep 1: drie uurlijkse modellen	5/5	5/5	5/5 voor NO_x 5/5 voor NO_2
Groep 2: alle uurlijkse modellen	5/5	5/5	5/5 voor NO_x 5/5 voor NO_2
Group 3: alle modellen	4/5	3/5	4/5 voor NO_x 5/5 voor NO_2

Voordelen van ensemblemodellering

Een belangrijk doel van dit onderzoek is om te bepalen of het gebruik van ensemblemodellering voordelen biedt. In het programmaplan voor SAGEN [1] worden verschillende mogelijke voordelen genoemd op basis van eerdere ervaringen met ensemblemodellering voor luchtkwaliteit. De belangrijkste mogelijke voordelen zijn in deze studie onderzocht.

Nauwkeurigheid

Modelensembles zijn mogelijk nauwkeuriger dan uitkomsten van individuele modellen. In specifieke concentratiemeetcampagnes presteerden de modelensembles vaak ongeveer even goed als de betere individuele modellen (bijvoorbeeld IFDM, OML-Multi en ADMS). Voor de Ringsted-depositiemetingen waren de resultaten voor de twee grotere ensembles (groep 2 en 3) vergelijkbaar met die van OML-Multi en OPS-LT. Over het algemeen lijken de twee grotere ensembles dus beter te presteren dan de meeste individuele modellen, maar niet beter dan elk afzonderlijk model.

Inzicht in onzekerheden
Ensemblemodellering geeft mogelijk inzicht in onzekerheden. Dit gaat ervan uit dat de onzekerheid in de ensemble-uitkomst kan worden afgeleid uit de variatie in de individuele modeluitkomsten. Voor de twee grotere ensembles waren de afwijkingen ten opzichte van de metingen inderdaad ongeveer even groot als de verschillen in de individuele modelresultaten. Voor het kleinste ensemble waren de afwijkingen ten opzichte van de metingen echter soms veel groter dan de verschillen in de individuele modelresultaten; deze laatste zijn voor dat ensemble dus geen goede maatstaf voor onzekerheid.

Uitwisseling van kennis
Ensemblemodellering leidt mogelijk tot meer uitwisseling van kennis tussen modelontwikkelaars. Modelverbeteringen kunnen daardoor sneller gerealiseerd worden. In dit onderzoek hebben we ervaren dat het uitwisselen van kennis vooral een kwestie van beschikbare tijd is.
Het vereist dat oorzaken van verschillen kunnen worden onderzocht en aangepakt. Dat is alleen mogelijk door modelformuleringen te vergelijken en uitkomsten, inclusief tussenresultaten, op detailniveau te bestuderen. Een automatische interface voor het parallel uitvoeren van modellen alleen is niet voldoende om verschillen tussen modellen te identificeren en te begrijpen. Wel zou ensemblemodellering een manier kunnen zijn om samenwerkingen tussen instituten te formaliseren en structurele financiering te verkrijgen voor kennisuitwisseling en modelontwikkeling. Het uitwisselen van kennis kan ook gestimuleerd worden door regelmatig benchmarkstudies uit te voeren met meerdere organisaties.

Mogelijke nadelen
Mogelijke nadelen van ensemblemodellering zijn in deze studie niet onderzocht. Hierbij moet gedacht worden aan de extra financiële kosten voor het op één lijn brengen van invoergegevens, het ontwikkelen en up-to-date houden van een software-interface, en de grotere benodigde computerkracht (of langere duur van modelruns).

Conclusies

Dit onderzoek vergeleek de uitkomsten van individuele modellen en modelensembles met metingen in drie meetcampagnes. Het vinden van bruikbare depositiemetingen in de directe omgeving van een bron bleek bijzonder lastig: de meeste depositiemetingen worden uitgevoerd in natuurgebieden op (te) grote afstand van individuele bronnen. Wanneer wel in de nabijheid van een bron wordt gemeten, is de duur van de metingen vaak beperkt, of wordt de bijdrage van droge depositie niet goed meegenomen. In de Ringsted-campagne is de depositie in de omgeving van een kippenboerderij bepaald aan de hand van isotoopmetingen in biomonitors. Deze (indirecte) depositiemetingen waren minder nauwkeurig dan vooraf werd verwacht.

Voor de concentratiemetingen in Ringsted, Affligem en Balko hadden IFDM, OML-Multi en ADMS consistent goede resultaten. De uitkomsten van deze modellen voldeden ook aan (bijna) alle acceptatiecriteria die Hanna en Chang hebben voorgesteld voor concentratiemetingen in landelijk gebied. De overige modellen hadden goede resultaten voor

sommige campagnes en minder goede resultaten voor andere campagnes. Daarbij werden acceptatiecriteria voor individuele modelprestatie-indicatoren van Hanna en Chang vaker niet gehaald.

Voor de Ringsted-depositiemetingen geldt dat de verschillen tussen modeluitkomsten en metingen groter waren dan voor concentratie-berekeningen. Dit komt waarschijnlijk doordat (i) depositiemodellering inherent complexer is dan concentratiemodellering en (ii) deze depositiemetingen (veel) minder nauwkeurig zijn dan de gebruikte concentratiemetingen. Tussen de individuele modellen hadden OML-Multi en OPS-LT de meeste overeenstemming met meetwaarden. De metingen waren echter niet nauwkeurig genoeg om sterke conclusies te trekken.

Modelensembles presteerden over het algemeen even goed als de betere individuele modellen voor individuele meetcampagnes, maar niet beter. Van de onderzochte ensembles gaf het ensemble van vijf uurlijkse modellen de beste resultaten. Ensemblemodellering maakt het verder mogelijk om de onzekerheid in de uitkomsten te schatten op basis van de spreiding in individuele modeluitkomsten. Voorwaarde is wel dat het ensemble niet te klein is.[6] Mogelijke nadelen van ensemblemodellering zijn niet onderzocht.

De bovenstaande conclusies hebben betrekking op de metingen die in deze studie zijn gebruikt. De meeste meetlocaties lagen op 5 tot 200 meter afstand van de bron. Dit onderzoek gaf daarom geen inzicht in de nauwkeurigheid van de modellen en modelensembles op (veel) grotere afstand.

Aanbevelingen

Uit dit onderzoek bleek dat sommige modellen beter overeenkomen met bepaalde metingen dan andere modellen. Het RIVM beveelt aan om de oorzaken van deze verschillen nader te analyseren in een vervolgstudie. Dat kan door modelformuleringen en tussenresultaten van modellen in detail te bekijken. De modellen die het meest afwijken van de metingen kunnen vervolgens met de opgedane kennis worden verbeterd. De meetcampagnes uit het huidige onderzoek en de onderliggende invoergegevens voor de modelberekeningen, zijn voldoende geschikt als basis voor dergelijk onderzoek.

Het onderzoek liet verder zien dat er nauwelijks betrouwbare metingen beschikbaar zijn om berekende depositiefluxen in de buurt van een bron te valideren. Dat komt doordat meetmethoden voor met name de droge depositie belangrijke vereisten hebben. Het RIVM beveelt aan om te onderzoeken wat er nodig is om een betrouwbare meetcampagne met depositiemetingen nabij een individuele bron op te zetten.

[6] Voor het kleinste ensemble, dat drie modellen omvatte, was de spreiding van de modelresultaten niet altijd een goede indicator voor de onzekerheid in de ensemble uitkomst.

1 Introduction

1.1 Context of the work

In the Netherlands and other countries, nitrogenous gases such as ammonia and nitrogen oxides are emitted by industrial sources, motorised traffic, and livestock, among others. A large part of the emissions deposit on the soil and in nature. Too high deposition has negative effects on the environment. First, nitrogen deposition affects the species composition in a nature area because some species benefit from high nitrogen deposition, supplanting other species. Second, nitrogen deposition affects the chemical composition of the soil and thus the presence of nutrients in the soil.

In order to develop and implement policy plans to protect vulnerable nature areas, it is necessary to know with sufficient precision how large the nitrogen deposition is. In the Netherlands, this estimate is made on the basis of a combination of model calculations, air concentration measurements, and deposition measurements. Model calculations also play a role in assessing permit applications for various economic activities.

To serve the above assessments, the models and the results must be sufficiently accurate. Accuracy can be identified in several ways, for example by comparing the results of calculations with measurements, by mirroring the design of models with current scientific knowledge, by comparing the results for various models, and by performing sensitivity analyses for input parameters and model parameters. Each of these ways has its own merits and complements the others.

The current work was undertaken as part of work package 2.2 of the SAGEN project [1]. The overall aim of the SAGEN project is to reduce uncertainties in determining the deposition of reactive nitrogen. Work package 2.2 aims to reduce uncertainties in calculations of nitrogen deposition at local scale, where local scale is defined as ‘from a hundred metres to a few kilometres’. This scale is important for permit application.

Uncertainties can be reduced by acquiring new knowledge and implementing that knowledge into the calculation models. Furthermore, it can be expected that using data from not one but several calculation models may reduce uncertainties. The uncertainty in the average outcome of models would, then, be smaller than the uncertainty in the individual outcome of one model.

The work in work package 2.2 of the SAGEN project is divided into five parts:

1. A comparison of differences in outcomes of operational models for calculating concentrations and deposition at local scale.
2. An analysis of the circumstances that bring about the largest differences between models.

3. A comparison of concentration and deposition outcomes of these models with observations at local scale.
4. An analysis of possible benefits of ensemble modelling, for example, whether ensemble modelling can provide higher accuracy and/or better insight into uncertainties.
5. An investigation on whether runs with a Large Eddy Simulation (LES) model can identify improvements to the operational models.

The current report describes items 3 and 4 in the list above. Results for items 1 and 2 have been reported recently in Kooi et al. [2]. The runs with the LES model (item 5) are being carried out by Wageningen University & Research during 2025 and 2026 and will be reported by Wageningen University & Research following full analysis of these results.

1.2 Expected benefits of ensemble modelling

Ensemble modelling involves the use of a number of models to derive estimates for quantities of concern, such as concentrations or deposition fluxes. Benefits of ensemble modelling have been demonstrated in various scientific fields, and have, for example, resulted in the development of fully automated forecasting systems for weather, climate, air quality, and flood risk.

According to the SAGEN programme plan [1], ensemble modelling offers multiple benefits compared with the use of one single model: it is supposed to result in more accurate estimates, to provide more insight into uncertainties in outcomes, to prevent users from using the most optimistic or pessimistic model, and to promote the exchange of knowledge between model developers resulting in faster model improvements.

In a way, ensemble modelling resembles the 'wisdom of the crowd': the mean of a wide spectrum of different opinions tends to be quite accurate. And hearing different opinions helps to reduce the risk of tunnel vision, that is, focusing too much on one outcome and ignoring possible other outcomes. Similar to the crowd, different models have different expertise (strengths) as well as blind spots (weaknesses) and use different sources of information. The resulting variability of outcomes has value. Moreover, ensembles need to find a good balance in weighing different perspectives. Without proper balance, ensembles may not perform better than state-of-the-art, or even fit-for-purpose individual models.

1.3 Aim and scope of this study

This study discusses two types of model outcomes: deposition fluxes and concentrations in ambient air. Deposition is directly relevant for the quality of ecosystems and nature areas. Concentrations in ambient air are indirectly relevant, via its influence on deposition.

The two main goals of this study are to identify how individual model outcomes compare with local-scale observations, for realistic sources of ammonia and nitrogen oxides emissions, and to identify how ensemble

outcomes compare with the same data. The larger goal is to obtain more detailed knowledge of the precision of individual model outcomes and outcomes for model ensembles. Policies for nature protection normally use annual average deposition as input. Therefore, annual average outcomes get more attention than outcomes for shorter time periods (hours, weeks, etc.).

The most relevant emission sources are livestock housing, industries, and motorway traffic. The relevant spatial range was defined in the programme plan [1] and involves distances ranging from hundreds of metres to a few kilometres from the source. Larger distances are investigated in WP2.1 of the SAGEN programme [3].

The report for the intercomparison of model outcomes [2] discussed to what extent outcomes of individual models vary; the 'spread of results' between models. Outcomes in the current study will be used to verify the conclusions from the intercomparison study.

1.4 Selection of models and validation campaigns

The available funding enabled us to select eight models and three measurement campaigns.

The selected models are presented in Table 1.1. All these models are regularly used for air quality and deposition calculations at local scale. The same set of models was previously used for intercomparing outcomes of individual models [2]. That report also contains short descriptions of these models.

Table 1.1 Selected models, the user and the role of the user

Model	**User**	**Role of user**
ADMS	CERC	Owner, developer, and user
AERMOD	UKCEH	User
IFDM	VITO	Owner, developer, and user
OML-Multi	Aarhus University	Owner, developer, and user
OPS-LT	RIVM	Owner, developer, and user
OPS-ST	RIVM	Owner, developer, and user
SRM2	RIVM	Developer and user
STACKS-D	RIVM	User

Outcomes of models and model ensembles should be compared with measurements near relevant emission sources. Finding campaigns with deposition measurements was treated as the highest priority but proved to be difficult, as only one campaign with deposition measurements was identified as useful for the current study. For this campaign, concentrations and deposition in the vicinity of a poultry farm in Denmark were measured during an eight-week period. The two other measurement campaigns identified for modelling have concentration measurements only, one relating to NO_x emissions from traffic on a motorway and the other relating to NO_x emissions from an industrial source (a compressor station). The search for measurement data and the results are expounded in in section 2.1 from Chapter 2.

1.5 Assessing model performance

In this study, the performance of models is assessed by comparing model outcomes with measurements. Performance scores then depend on the magnitudes of discrepancies between modelled outcomes and measurements. These deviations depend on the accuracy of the model, the accuracy of the input parameters for the model calculations, and the accuracy of the measurements. Therefore, performance scores only reflect 'intrinsic' model accuracy if the input parameters are known with sufficient accuracy and if the measurements are sufficiently accurate. A literature review was conducted to investigate which indicators of model performance are commonly used and to explore which outcomes are considered to be acceptable for models.

1.6 Limitations of the work

The scope of the study was limited by the available amount of time and funding and by the availability of measurements. Ideally, models are validated with observational data from a large number of campaigns that differs in terms of emissions characteristics and environment. The available budget only stretched to using three measurement campaigns for validating the selected models. Most models had already been validated against different types of measurements from other campaigns in previous studies.

Significant efforts were undertaken to find publicly available campaigns in which deposition was measured with sufficient precision at distances ranging from hundreds of meters to a few kilometres from a dominant known source, during a sufficiently long time period (ideally one year or more). Such campaigns proved to be scarce. Eventually only one deposition measurement campaign was believed to be suited for use in this study. Therefore, conclusions about the accuracy of model outcomes for deposition heavily depend on the scope, characteristics and precision of this single measurement campaign, and to the measurement locations in this campaign, which were mostly close to the source (the largest distance being 260 m).

Regarding concentrations, model outcomes could be compared with measured concentrations in three different campaigns, using measurements around a Danish chicken farm, a Belgian motorway and a compressor station in the USA, all during normal operations. An important limitation is that these measurements were also carried out relatively close to the source; mostly at distances between 5 and 200 m from the source.

1.7 Organisation of the study

The work in this study was coordinated by RIVM, using funding from the Dutch Ministry of Agriculture, Fisheries, Food Security, and Nature. Four foreign partners of RIVM (Aarhus University, CERC, UKCEH, and VITO) agreed to contribute to this study, using a small part of the funding from the ministry and additional external and/or internal funding.

This report is a joint product by RIVM and the partners mentioned above. RIVM made a proposal for the measurement campaigns to use, also using input and feedback from the partners, and taking into account

the available funding. This proposal was approved by the group. Then, RIVM provided the partners with more detailed descriptions of the measurement campaigns to such an extent that each partner was able to define which settings and input values to use in their model. The analysis of model outcomes, including the calculation of performance scores, was undertaken by RIVM, and subsequently verified by the partners. Lastly, RIVM wrote this report using input and feedback from the various partners.

Data processing and data visualisation were carried out with the computer language R [4].

2 Selection of data and methods

2.1 Selection of measurement campaigns

The aim of this study is to use observations from field campaigns to investigate the accuracy of modelled concentrations and deposition fluxes, regarding emissions of reactive nitrogen from a single source, in the vicinity of that source. Validating deposition outcomes had higher priority than validating concentration outcomes, but proved to be difficult. Livestock farms, industries, and motorway traffic were believed to be the most important source types. Distances to the source should ideally be between a hundred metres and several kilometres.

Prior to this study, all selected models had already been validated against concentration measurements (see Appendix 6 of [2] for details). The number of measurement campaigns used to validate concentration outcomes was particularly large for ADMS and AERMOD, although these models had limited evaluation of performance in terms of deposition modelling. OML had previously been compared with local-scale deposition measurements near a chicken farm in Ringsted, Denmark [6]. Deposition outcomes of OPS-LT had previously been compared with deposition measurements at regional scale (i.e. not close to a strong individual source) [5], but not at local scale.

2.1.1 Campaigns with deposition measurements

The primary focus of this work was to find useful measurement data for validating calculated deposition rates close to a source (also referred to as 'local scale'). A particular feature of this scale is that concentrations can be much higher than regional average concentrations, at least part of the time (when the wind is blowing towards the location of concern). The related research question concerns the impact of relatively large atmospheric concentrations on deposition rates. In terms of ammonia, plants can become saturated when exposed to high concentrations for a sufficiently long time. Saturation limits further ammonia uptake and may even result in re-emission at later times. This means that model validation requires measurements in which plants were exposed to comparable concentration levels for part of the time. Ideally, deposition measurements are carried out close to a source, at locations and in habitats similar to those that models are expected to represent.

The availability of measurement data was investigated by studying literature, by exploring websites for air quality model validation, and by using existing knowledge of the various partners in the project. In total, more than 150 research papers and reports were scanned for potential relevance to this project. About twenty papers and reports were analysed in great detail.

Following the analysis of literature, it appears that the majority of deposition measurements can be grouped into one of the following three categories:

1. Deposition measurements in regional or national networks of deposition monitoring stations. These campaigns measure spatial

and temporal variance of deposition and make it possible to study correlations between deposition rates, concentrations, season/climate, and ecosystem types. These stationary monitoring stations are mostly located in nature areas and away from strong individual emission sources.
2. Ad hoc campaigns that aim to identify daily and seasonal trends in deposition for a specific type of ecosystem or plant, including the influence of ambient concentrations on deposition rates. These temporary campaigns are normally carried out in nature areas.
3. Campaigns that measure vertical concentration fluxes in exposed fields with the aim of identifying effective emission strengths from these fields. Such measurements are, for example, carried out on fields with grazing cattle or on fields where manure or fertiliser has been applied. The measured flux is the sum of an upward flux (emission) and a downward flux (deposition) and the result is an effective emission strength. The time duration of such experiments is usually short (several hours or several days).

The above classification shows that the majority of the deposition measurements had either been carried out far from a dominant source, or that the measurement duration was short. Measurements away from dominant sources were believed to be unsuited to the current study for the following reasons:

1. The concentrations to which vegetation in these 'regional studies' is exposed, are typically lower than the concentrations near a source. The deposition rates found in regional studies may not be representative of the deposition rates at higher concentration levels.
2. The concentrations at these locations originate from many sources, and at least some are located at a considerable distance from the measurement location. Thus, comparing these measurements with model outcomes shows how capable models are of calculating contributions from far-away sources, rather than their capability to calculate concentrations and deposition rates near a source. Only the latter falls within the scope of this study.

Short-duration campaigns were excluded from this study because:

1. Policy plans and decisions regarding deposition of reactive nitrogen components to nature areas generally use annual average deposition for input.
2. Relatedly, OPS-LT, SRM2 and STACKS-D are primarily designed to calculate annual average concentrations and deposition fluxes. The aim of the current study is to identify which models are suitable for policy applications. Models may not need to represent processes to the highest level of detail in order to generate sufficiently accurate period-average concentration and deposition values.

Thus, the challenge for the current study was to find deposition measurements that allow validation of modelled deposition fluxes close to a source during a prolonged period (the scope of this study). This

resulted in the following list of requirements for selecting deposition measurements:

1. Deposition measurements were carried out near a dominant local source. The preferred distance range between the source and the measurement location is 50 to 500 m for low sources (animal houses, road traffic, and exposed fields), and perhaps slightly larger for high sources (industrial stacks).
2. Emissions include reactive nitrogen: ammonia and/or nitrogen oxides. The emission source is representative of one of the preferred source types (livestock farms, exposed fields, industrial stacks, and motorway traffic).
3. Emission rates and other relevant emission characteristics are known in sufficient detail.
4. The main features of the campaign have been published in a report or scientific paper. That report or paper describes which measurement techniques were used and provides information on the quality/accuracy/reliability of the observations.
5. Measurement data is publicly available or can be made publicly available.
6. Either total deposition was measured or dry and wet deposition were measured separately. Dry deposition should relate to plants (vegetation) and should be measured with sufficient accuracy. The latter is important because close to a source, dry deposition is believed to be more important than wet deposition.
7. The duration of deposition measurements was at least four weeks. This minimum campaign duration is needed to get reasonable estimates from models that are designed to calculate annual-average deposition fluxes.

Between all deposition measurement campaigns that were explored in this study, only one campaign fulfilled the above requirements. This campaign is hereafter referred to as the 'Ringsted campaign'. Even for this campaign, the accuracy of the deposition measurements was not as high as had been hoped for. All other campaigns had one or more limitations that made them unsuited to validation deposition outcomes at local scale. The most relevant campaigns are described in more detail in Appendix 1 (sections 12.3.3 and 12.3.4), including reasons for selecting them (Ringsted) or not (all others).

Precise reasons why accurate deposition measurements near a source and with sufficient duration are scarce have not been investigated. We conceive that the following factors may play are role:

- The main focus of reactive nitrogen deposition studies is the impact of deposition on the quality of ecosystems. Most studies are, therefore, carried out in nature areas. These areas are usually remote; not near strong emission sources of reactive nitrogen.
- Measuring reactive nitrogen deposition close to a source appears to be challenging, in particular regarding dry deposition (which is expected to be the dominant contribution close to a source).

More information on limitations of different measurement techniques is provided in section 12.3.1 of Appendix 1.

2.1.2 *Additional campaigns with concentration measurements*

The idea was to select measurement campaigns for a livestock farm, a motorway, and an industrial source. No suited deposition measurements were found for the latter two source types (see previous subsection). As a result, only model outcomes for concentrations could be validated with measurements for these two source types. The criteria for selecting such measurements were similar to the criteria used for selecting deposition measurements. Most importantly, the measurements had to be carried out close to a known source, for a prolonged period (at least four weeks) and without significant contributions from other, more distant, sources. The search criteria and the concentration measurement campaigns that were further investigated, are discussed in Appendix 1 (in particular in section 12.4).

Regarding concentration measurements around industrial sources, more than thirty datasets were found that had been used for validating air quality models in the past. In particular, the websites from EPA [17], HARMO [22], and ADMLC [23], provide good overviews of these datasets. Following consideration of the benefits and limitations of the various datasets, a measurement campaign around a compressor station in Balko, Oklahoma (USA) [20], was regarded as being most useful. In contrast to many other datasets, this dataset involved NO_x and NO_2 measurements and related to a source during normal operation. The duration of this campaign (13 months) was much longer than usual, and the quality of the measurements was believed to be above average. The main restriction of this dataset was the low number of measurement locations (only four).

Regarding concentrations measurements near motorways, three potentially suited datasets were investigated in detail (see section 12.4.4). Out of these three, the Life+ ATMOSYS highway campaign near Affligem, Belgium [16] was believed to be the most useful. In this campaign, NO_2 concentrations had been measured at 6 locations close to the motorway during a period of 36 consecutive weeks.

2.1.3 *Summary*

In summary, only one measurement campaign, the Ringsted campaign, was found in which both concentrations and deposition fluxes were measured near a local NH_3 or NO_x source, for a sufficiently long duration and with adequate spatial coverage. The campaign was selected for validating model outcomes in this study. Further details about the campaign can be found in Chapter 3.

Two additional campaigns were selected in which concentrations were measured near a motorway (Affligem) and near an industrial source (Balko). Both have emissions of NO_x. The Affligem campaign is described in Chapter 4 and the Balko campaign in Chapter 5.

2.2 Methods for model validation

Model validation involves the comparison of model outcomes with measurements. A prerequisite for validating models with measurements is that the measurements must be sufficiently accurate.

A considerable amount of literature is available on the various ways in which model outcomes are compared with measurements, including papers and reports from Chang & Hanna [26], [27], Dennis, Fox & Fuentes [28] and Thunis, Pederzoli & Pernigotti [29]. In general, the comparison is carried out on the basis of visualisations and statistical parameters that summarise the differences between model outcomes and measurements. A recurring statement in literature is that different types of visualisations and statistical parameters each have specific advantages and disadvantages. For a detailed understanding of differences between model outcomes and measurements, multiple visualisations and statistical parameters must be considered together. The right choice depends on the focus of the validation exercise and the available data.

The performance of models is normally investigated with a set of statistical parameters: referred to as performance 'measures', 'metrics', 'indicators', or 'indices' (hereafter: indicators). More than ten separate indicators are used in validation studies for air quality models; some more frequently than others. In this report, we discriminate between indicators that do not take measurement uncertainty as input (section 2.2.1), and indicators that do consider measurement uncertainty (section 2.2.2). The first type of indicators have been frequently used since the early 1990s, in particular as a response to early work by Hanna, Chang, and Olesen [26], [27], [31]. The latter type of indicators have been developed by the EU Joint Research Centre and the FAIRMODE consortium for evaluating air quality models [29], [32], [33]. All performance indicators depend on (i) intrinsic model quality, (ii) input accuracy, and (iii) measurement precision.

Performance criteria are criteria that define which outcomes for performance indicators are deemed acceptable for specific types of model uses.

2.2.1 Performance indicators that do not directly consider measurement uncertainty

A paper by Chang & Hanna that was published in 2004 [26] provides an overview of various model validation techniques and their advantages and limitations. Six separate performance indicators (in their terminology: performance measures) were discussed in that paper: Fractional Bias (FB), Geometric Mean Bias (MG), Normalized Mean Square Error (NMSE), Geometric Variance (VG), Correlation Coefficient (R), and fraction of predictions within a factor of two of observations (FAC2). These same six indicators were used in the BOOT software for validating air dispersion models [27] and have been widely used ever since [37]. All these indicators depend on measured and modelled outcomes. Measurement accuracy is not input for any of these indicators.

Some related indicators have later been proposed by other authors (see [36] for more details), including Average Normalised Absolute Bias (ANB), Normalised Standard Deviation (NSD), Index of Agreement (IoA), and Hit Rate (H). These indicators are used less frequently and appear to have limited added value for the current study. The same

applies to the Normalised Absolute Deviation (NAD), which is discussed in [39].

The six performance measures (metrics/indices/indicators) from Chang & Hanna [26] and the BOOT software for model evaluation ([31],[27]) are discussed below. The symbol *O* refers to the set of observations, and *M* to the corresponding set of model outcomes.[7]

- The Fractional Bias (FB) is a measure that indicates whether, on average, the model overpredicts or underpredicts. Chang & Hanna [26] use a positive sign for FB if modelled outcomes are, on average, lower than measured outcomes.

  $$FB = \frac{(\bar{O} - \bar{M})}{\frac{1}{2} \cdot (\bar{O} + \bar{M})} \quad (1)$$

  In this study, we prefer reversing the sign of the outcome; positive values for bias (FB), then, indicate average overprediction by models, while negative values indicate average model underprediction. The (modified) definition of Fractional Basis that we will use in this report is:

  $$FB = \frac{(\bar{M} - \bar{O})}{\frac{1}{2} \cdot (\bar{M} + \bar{O})} \quad (2)$$

  The term 'bias' suggests that FB=0.1 means that calculated outcomes are on average 10% larger than observed values. This is, however, only approximately true for low FB. For example, 50% larger calculated values gives FB=0.4, and 50% lower calculated values gives FB=-0.67.

- The Geometric Mean Bias (MG) is an alternative to FB and also indicates whether the model overpredicts or underpredicts. The difference is that the evaluation now depends on factor deviations rather than absolute deviations. Simply put, a factor of two underprediction is believed to be as good or as bad as a factor of two overprediction. Factor deviation can be evaluated by taking the logarithm of deviations, because ln(a)=-ln(1/a).

  $$MG = e^{\overline{\ln(O)} - \overline{\ln(M)}} = e^{\overline{\ln(O)}} / e^{\overline{\ln(M)}} \quad (3)$$

  Similar to FB, we prefer, however, to associate values larger than 1 with model overprediction, and values smaller than 1 with model underprediction. Therefore, in this study, we will use the following formula for Geometric Mean Bias:

  $$MG = e^{\overline{\ln(M)} - \overline{\ln(O)}} = e^{\overline{\ln(M)}} / e^{\overline{\ln(O)}} \quad (4)$$

[7] In their paper, Chang & Hanna use C_p for model outcomes (predictions) and C_o for measured outcomes (observations), where C stands for 'concentration'.

MG corresponds better to intuitive bias than FB: if all calculated outcomes M are 50% larger than corresponding measured outcomes O, then MG would be 1+0.5=1.5. conversely, a 50% lower values gives MG = 1-0.5=0.5. The same applies to all other percentages.

- The Normalised Mean Square Error (NMSE) is a dimensionless (=normalised) metric for the average square deviation between model outcomes and measured outcomes:[8]

$$NMSE = \frac{1}{\bar{O} \cdot \bar{M}} \cdot \overline{(O - M)^2} \quad (5)$$

If all (absolute) deviations are of similar magnitude, NMSE is a typical value for these (absolute) deviations. If deviations are both small and large, the larger (absolute) deviations influence the NMSE stronger than the smaller values.

- The Geometric Variance (VG) evaluates differences between model outcomes and measurements by order of magnitude (i.e. by taking the logarithm of modelled and measured outcomes). It is an alternative to NMSE.

$$VG = e^{\overline{(\ln(O) - \ln(M))^2}} = e^{\overline{(\ln(O/M))^2}} \quad (6)$$

The interpretation of VG is more difficult than that of NMSE; it is not a logarithmic evaluation of mean deviation, but a logarithmic evaluation of mean square deviation. VG can, however, be transposed into a Geometric Standard Deviation using $GSD = e^{\sqrt{ln(VG)}}$. Then, GSD is a measure for the average factor deviation between observations and measurements. In absence of random scatter, all deviations are defined by model bias only, and then GSD is equal to MG (average ratio of calculated outcomes and measured outcomes).

- Correlation (R) is an indicator for proportionality (whether high/low trends in observations are reflected by high/low trends in predictions and vice versa:

$$R = \overline{(O - \bar{O}) \cdot (M - \bar{M})} / (\sigma_m \cdot \sigma_o) \quad (7)$$

- FAC2 is the fraction of model outcomes that is within a factor 2 of the corresponding observations. FAC2 has a value between 0 (imperfection) and 1 (perfection).

If all model outcomes are identical to the corresponding observations, MG, VG, R, and FAC2 would be 1 and FB and NMSE would be 0.

The two 'arithmetic' indicators, FB and NMSE, depend on absolute differences between measured and modelled outcomes, whereas the two 'geometric' indicators, MG and VG, depend on relative deviations

[8] Normalised Mean Square Deviation would have been a better name. We do not know to what extent modelled and measured outcomes are right or wrong: we measure the differences between the two.

(deviations as a fraction of the measured outcome). This difference is important when measured and modelled outcomes vary significantly[9] in space or time. Outcomes for FB and NMSE are, then, dominated by the largest measured and calculated outcomes, for instance, close to the (dominant) source. For such studies, the geometric indicators MG and VG normally provide a more balanced outcome for the full dataset ([26], [30], [37]). A concern for the use of MG and VG is that these indicators can be strongly affected by large relative deviations for extremely low outcomes (concentrations or deposition fluxes), for instance. near or below instrument thresholds ([26], [30]) or model cut-offs.

Chang & Hanna [26] used the experiences from at least eight separate model evaluation and validation studies to deduce 'typical magnitudes' of good model performance. In each of these studies, data from multiple measurement campaigns had been used to evaluate or validate models. The total number of campaigns used in all these studies together appears to be larger than 30 and includes various 'research grade field experiments with good instruments and uncomplicated terrain and simple source scenarios', and some mesoscale tracer studies or measurements from routine monitoring networks. In most cases, the arc-wise maximum concentrations were compared across models and measurements, at different distances and for different times. The analysis of these studies resulted in the following proposed criteria for well-performing air quality models:

- the fraction of model predictions within a factor of two of observations is about 50% (or higher);
- the mean bias is within ±30% of the mean;
- the random scatter is about a factor of two to three of the mean, or less.

The first criterion implies FAC2≥0.5, the second criterion implies |FB|<0.3 or 0.7≤MG≤1.3, and the third criterion implies[10] NMSE<4 or VG<1.6 when allowing a factor of two for scatter, and NMSE<9 or VG<3.3 when allowing a factor of three for scatter.

The above criteria were further tested by Hanna & Chang in the period 2000–2010, using four new urban tracer studies, and five rural campaigns [39].[11] From this additional experience, they proposed new criteria for rural and urban concentration measurement campaigns (Table 2.1). The criteria for urban campaigns should be somewhat looser than the first due to the 'larger variability introduced by buildings and differing land use'. The criteria relate to arc-wise maximum concentrations.

Furthermore, Hanna & Chang [39] introduced a new performance index: the Normalised Absolute Deviation (NAD), with two substantially different definitions. In the first definition, $NAD = \overline{|O - M|}/(\bar{O} + \bar{M})$, NAD is a normalised mean absolute deviation between observations and

[9] For example, one or several orders of magnitude, depending on the numbers of high versus low outcomes.
[10] The translations into performance measures are from the BOOT software User Guide [27]. The last criterium was incorrectly translated into NMSE < 1.5 and VG < 4 (i.e. criteria for NMSE and VG reversed). These wrong conditions have been used in several research papers since (e.g. Theobald et al. [41] and Patiño and Duong [47]). The right translations were provided in Schatzmann et al. [38].
[11] The paper refers to 'five rural field experiments' but does not specify these campaigns. It is also unclear whether these rural campaigns had already been considered in [26], or not.

measurements. In the second definition, also referred to as 'threshold-based normalized absolute difference' and 'fractional area for errors', NAD defines the overlap between measured and observed concentrations above a certain concentration threshold.[12] This parameter can be used to identify to what extent calculated concentration footprints overlap with measured concentration footprints. It appears to be most useful when comparing arc-wise concentration measurements with model outcomes regarding field experiments with orderly spaced sensors (e.g. concentric circles).

Table 2.1 compares the new (2012) criteria for rural and urban campaigns from Hanna & Chang with their previous (2004) criteria.

- For rural campaigns, the tolerance for scatter (NMSE) was lowered, compared with 2004. In addition, two extra criteria were defined; one for (threshold-based) NAD and one for tolerance. The tolerance criterion states that the four criteria (FB, NMSE, FAC2, and NAD) 'should be met over half the time, on average, at all field experiments tested'.
- For urban campaigns, the tolerance for FB was increased to 0.67, and the tolerance for NMSE was set to 6. Again, two extra criteria were added; one for (threshold-based) NAD and one tolerance criterion. The tolerance criterion, 'at least half of the performance criteria are met for at least half of the field experiments considered', is less stringent than for rural campaigns.

Table 2.1 Overview of criteria proposed by Chang and Hanna

Chang&Hanna 2004	**Hanna&Chang 2012 Rural campaigns**	**Hanna&Chang 2012 Urban campaigns**
|FB| < 0.3	|FB| < 0.3	|FB| < 0.67
0.7 ≤ MG ≤ 1.3	See text below	See text below
NMSE <4 or NMSE <9	NMSE < 3	NMSE < 6
VG <1.6 or VG <3.3	See text below	See text below
FAC2 ≥ 0.5	FAC2 ≥ 0.5	FAC2 ≥ 0.3
	NAD ≤ 0.3	NAD ≤ 0.5
	+ tolerance condition	+ tolerance condition

The new tolerance conditions for rural and urban campaigns are formulated in different ways, although Hanna & Chang refer to these criteria as being the same. The data in Table 6 of their article [39] and the accompanying text make clear what is actually meant. All performance indicators for separate sets of measurements should be grouped together, and the fraction of indicators within acceptable limits should be larger than 0.5.[13]

[12] The second definition depends on modelled/predicted concentrations (C_p), observed concentrations (C_O) and a user-defined threshold concentration (C_T), and then defines three types of outcomes: false positives ($C_p>C_T$ & $C_o<C_T$), false negatives ($C_p<C_T$ & $C_o>C_T$) and overlaps ($C_p>C_T$ & $C_o>C_T$). Then, the NAD, now also referred to as the 'fractional area for errors', is equal to the ratio of (i) the mean number of false positives and false negatives, and (ii) the total sum of the false positives, false negatives and overlaps.

[13] Table 6 in [39] provides a summary of 60 experiments for which 4×60=240 indicator outcomes were calculated. 141 of these outcomes were within acceptable limits, that is, 59%. Hanna & Chang conclude: 'Overall, more than half of the time the acceptance criteria proposed above were met (the actual fraction is 0.59). As a result, the comprehensive acceptance measure exceeds 0.50 and it is concluded that JEM's performance is acceptable for urban applications.'

The criteria for mean bias were expressed in terms of FB. In terms of MG, the mean bias criterion would be 0.7≤MG≤1.3 for rural campaigns, and 0.5≤MG≤2 for urban campaigns. Similarly, scatter criteria in terms of VG would be VG<1.35 for rural campaigns and VG<2.2 for urban campaigns.[14] Furthermore, it is worth noting that the default criterion for MG, (0.7≤MG≤1.3), is more tolerant for underpredictions than for overpredictions. Alternative formulations: 0.77≤MG≤1.3 or 0.7≤MG≤1.42 treat under- and overpredictions on an equal footing.

Several authors notice that the above criteria for air quality models should be used as a guidance but not be applied too strictly. For example, Chang & Hanna [39] mention that 'some trial and error was necessary because of the desire to set the acceptance criteria at a practical level, such that they can realistically discriminate between models that perform well and those that do not perform well', and also that 'the proposed rural and urban model acceptance criteria themselves are somewhat arbitrary, and a more valid and widely recognized set is expected to result from further testing of these acceptance criteria with more models and field data sets by a wide array of research groups.' In a related paper [40], they state that the proposed criteria 'are intended to represent a start to the discussion and it is hoped that further studies can help develop acceptance criteria for other applications and types of available data'. According to Schatzmann et al. [38], the criteria were proposed for comparisons of maximum concentrations on arcs and should be used with prudence when comparing 'point by point' outcomes at specific locations.

The above indicators have been widely used from the 1990s onwards. Some examples of how various authors use these indicators to evaluate model performance are provided below. In all of these papers, performance indicators are calculated and used to further discuss model performance. In about half these papers, indicator outcomes are also compared with acceptance criteria proposed by Chang and Hanna, with some differences in the choice of appropriate criteria.

- Theobald et al. [41] compared concentration measurements near two pig farms with outcomes for four dispersion models. It appears that annual average concentrations (not weekly measurements) were used for these comparisons. Five indicators (FB, MG, NMSE, VG, and FAC2) were calculated for each model and for both campaigns and were compared with the initial criteria from Chang & Hanna [26]. Inspired by later work by Chang and Hanna, model performance was defined as being acceptable if the model 'meets the criteria for at least half of the performance tests', here meaning that at least five out of the ten tests (i.e. five tests for two campaigns) should be fulfilled.[15] Probably due to a mistake in the User Guide for the BOOT software (see footnote 10), the acceptance criterion for NMSE was confused with the acceptance criterion for VG, resulting in some false positive and some false negative indicator outcomes.
- Heist et al. [42] compared concentration measurements close to two roads with outcomes of models. In terms of performance

[14] If *a* is the factor random scatter, then $NMSE = a^2$ and $VG = e^{ln(a)^2} = e^{ln(\sqrt{NMSE})^2}$.
[15] This tolerance criterion is, however, slightly different from Hanna & Chang's tolerance criteria for rural and urban campaigns expressed in [39].

indicators, they only discuss FB, NMSE, R, and FAC2. The outcomes were not compared with acceptability criteria.
- In Walker et al. [43], concentration measurements downwind from a poultry farm were compared with outcomes of a single model. FB, MG, NMSE, VG, and FAC2 were calculated and compared with the initial acceptance criteria proposed by Chang & Hanna [26].
- In Stocker et al. [44], concentration measurements in three campaigns near poultry farms were compared with calculated concentrations. FB, NMSE, FAC2, and IoA were used to get an impression of the performance of two models and various input options for each model.
- Patiño & Duong [47] compared concentration outcomes of three models with measurements downwind from an industrial stack. The model performance was measured with FB, NMSE, MG, VG, FAC2, and NAD, and corresponding outcomes were compared with the acceptance criteria for urban campaigns from [39] regarding FAC2, FB, NMSE, and NAD, and to original acceptance criteria from the BOOT software [27] regarding MG and VG (including the erroneous use of VG≤4 instead of VG≤2.2, see footnote 10).
- In a paper by Mazzola et al. [48], outcomes of seventeen dispersion models are compared with measured concentrations in nine field trials with short-duration chlorine releases. The concentrations were measured at 200 m to 11 km distance from the release source. Model outcomes were compared with measured data as a function of distance. Four performance indicators, VG, MG, FAC2, and FAC5,[16] were used to summarise model performance. Acceptance criteria for rural and urban campaigns proposed by Hanna & Chang were believed to be unsuited to these dense gas experiments.

Regarding deposition, no studies were found in which performance indicators and criteria from Hanna and Chang have been used to validate deposition model outcomes against deposition measurements. From a formal point of view, there are no great restrictions for using the indicators from Chang and Hanna to investigate deposition model performance.[17] The performance criteria from Chang and Hanna were derived from validation studies for modelled concentrations. The purpose of these criteria was to distinguish between better performing and worse performing models, using experience of what can be typically expected from models in terms of concentration modelling accuracy. For deposition modelling, such experience is currently being developed. Acceptance criteria for deposition modelling can only be proposed once sufficient experience is available.

2.2.2 Performance indicators from the FAIRMODE framework for air quality monitoring

Some new performance indicators and acceptance criteria were proposed by the EU Joint Research Centre and the FAIRMODE consortium [29], [32], [33], [34]. The motivation of this work was to

[16] The fraction of predicted concentrations within a factor of five from observed concentrations.
[17] As for any type of measurement, it will still be important to investigate measurement accuracy and limitations in measurement precision (e.g. threshold levels for reliable values).

define which models may be used for monitoring air quality in the context of the EU Air Quality Directive (AQD).

In this new framework, model performance is assessed with a 'Model Quality Indicator' (MQI). This MQI is the ratio of the average difference between model outcomes and observations, and the product of the measurement uncertainty and a scaling factor. The latter two (measurement uncertainty and scaling factor) have only been defined for a limited number of standardised air quality measurements (see [33]).

Although the FAIRMODE framework is an accepted framework in the context of air quality monitoring, we decided not to use it in the current study for the following reasons:

- The current study focuses on deposition modelling, which falls outside of the scope of the FAIRMODE framework.
- The framework can only be applied if both the measurement uncertainty and the scaling factor are known. Guidance is available for some standard applications but not for the cases considered in this report.
- Only limited information on measurement uncertainties was available for the considered cases.
- The available amount of time in this project was too limited for detailed analyses of measurement uncertainties and appropriate values for the scaling factor.

2.3 Time durations to be used for the evaluation of models

The goal is to determine to which extent models are fit for purpose regarding their application to policies for deposition of reactive nitrogen. These policies normally use annual average deposition as input. This is why the model performance for annual average outcomes is measured rather than shorter-term (hourly or weekly) outcomes.

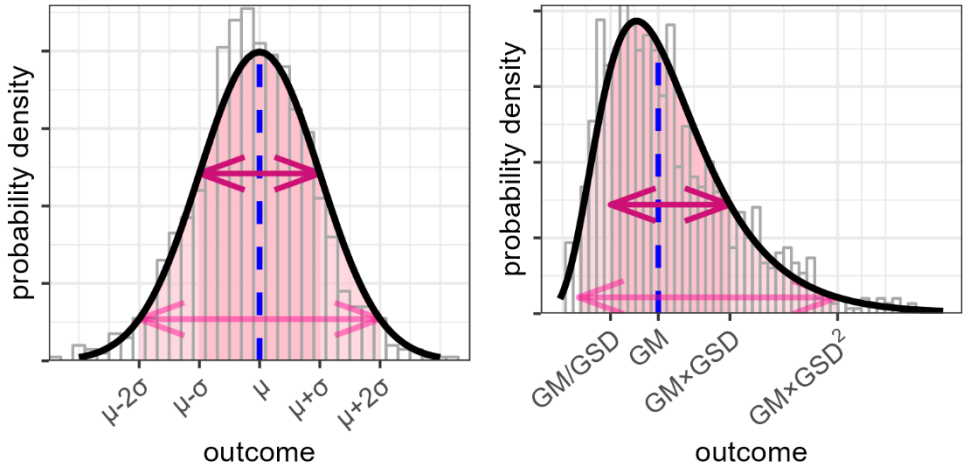
2.4 Metrics for defining average outcomes and spread of outcomes

In different chapters of this report, we calculate average outcomes and spread of outcomes. For model ensembles (Chapter 6), the 'ensemble outcome' is taken to represent an average of the underlying individual model outcomes. The uncertainty in the ensemble outcome (section 6.6) is assumed to correlate with the spread between individual model outcomes. For individual models, we calculate mean deviations with measured outcomes (bias) and spread of outcomes around the mean bias.

In general, we assume that outcomes (including deviations) are either normally (Gaussian) or lognormally spread around a common mean (see Figure 2.1). The normal distribution is defined by its mean value (μ) and its standard deviation (σ). The standard deviation is a measure for the width of the distribution and can be used to define probability intervals: for instance, 68% of the data points have a value between $\mu-\sigma$ and $\mu+\sigma$, and 95% have a value between $\mu-2\cdot\sigma$ and $\mu+2\cdot\sigma$. Similarly, the lognormal distribution is defined by its Geometric Mean and its Geometric Standard Deviation. In this distribution, 68% of the data

points have a value between GM/GSD and GM·GSD, and 95% have a value between GM/GSD^2 and GM/GSD^2.

Figure 2.1 Probability densities for the normal (left) and lognormal (right) distributions



If model outcomes are spread normally around their common mean, the mean of the distribution (μ) is equal to the Arithmetic Mean (AM) of the individual model outcomes. The AM is the normal average of the set of outcomes:

$$AM = \frac{1}{N} \cdot \sum_i M_i \qquad 8$$

The standard deviation of the distribution (σ) is equal to the Standard Deviation of the set of outcomes:

$$SD = \left(\frac{\sum_i (M_i - \bar{M})^2}{(N-1)} \right)^{1/2} \qquad 9$$

Then, the Relative Standard Deviation is defined as the ratio of the Standard Deviation (SD) and the Arithmetic Mean (AM):

$$RSD = \frac{SD}{AM} \qquad 10$$

The symbols μ and σ refer to the true (perfect) normal distribution, AM and SD refer to the sets of model outcomes, which can only be normally distributed by approximation.

If model outcomes resemble a lognormal distribution, the Geometric Mean (GM) is defined as:

$$GM = \left(\prod_i M_i \right)^{1/n} = exp\left[\frac{1}{n} \sum_i \ln (M_i) \right] \qquad 11$$

and the Geometric Standard Deviation as:

$$GSD = exp\sqrt{\left(\frac{1}{N-1}\sum_i (ln(M_i) - ln\,(GM))^2\right)} = exp\sqrt{\frac{1}{N-1}\cdot\sum_i \left(ln\left(\frac{M_i}{GM}\right)\right)^2} \qquad 12$$

An important difference between SD and GSD is that SD has the same units as the quantity of concern, while GSD is dimensionless. For example, a set of concentration values can have an average value (AM) of 200 mg/m^3 and a standard deviation of 40 mg/m^3. GSD is a typical multiplicative factor for deviation. If GM = 180 mg/m^3 and GSD is 1.4, the majority of the concentrations have values between 180/1.4=129 mg/m^3 and 180·1.4=252 mg/m^3.

The normal distribution is symmetric around its mean and goes from -∞ to +∞. The lognormal has a lower limit of zero. If outcomes have a lower limit (e.g. 0), they will only resemble the normal distribution if their standard deviation is much smaller than their arithmetic mean. If the standard deviation is not small, the distribution of outcomes may be closer to the lognormal distribution.

When we assume certain distributions of outcomes (normal or lognormal), we will also test whether that assumption is true for the data used.

2.5 Analysis of meteorological conditions

For each campaign, measurements of meteorological conditions were provided to RIVM. The model intercomparison study [2] showed that atmospheric stability is an important parameter for model results. In this report, a set of nine stability classes (see Table 2.2) was used to classify atmospheric stability for each hour. The stability class is derived from the Monin-Obukhov length, which was calculated with OPS-ST. The classification is derived from a paper by Gryning et al. [35], except for the extremely unstable and extremely stable classes. These were added by RIVM in order to allow classification of all hours. The classification of stability was only used to provide more insight into the meteorological conditions during the campaigns. It was not used as input for any of the models.

Table 2.2 Classification of stability (only used to group outcomes when discussing the relevance of atmospheric stability)

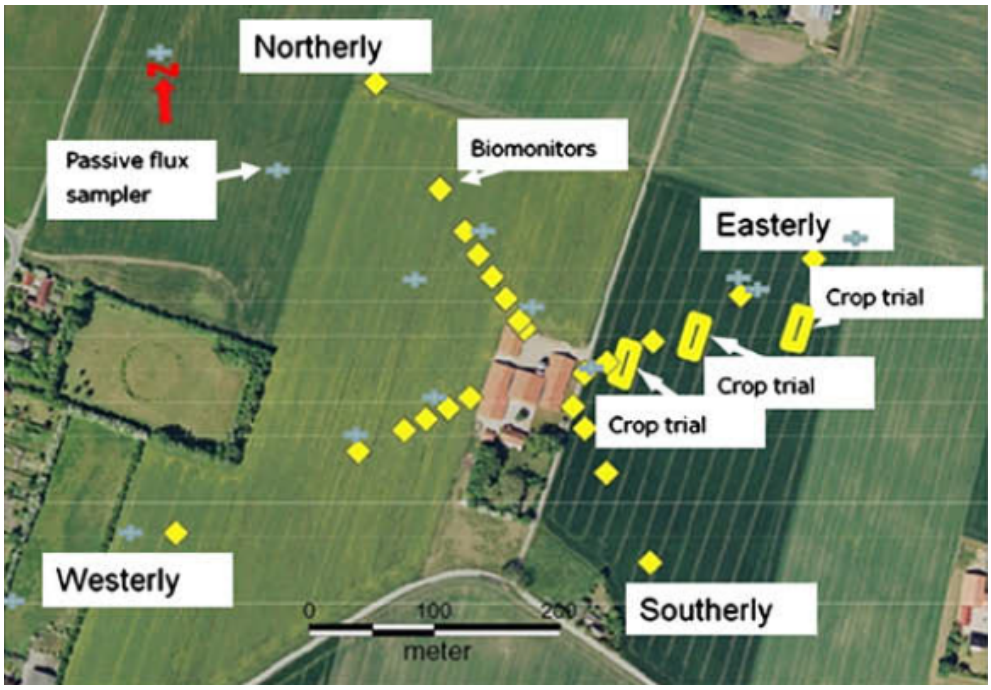
Stability class	**Monin-Obukhov length (L_b)**
Extremely unstable (xu)	$-50 \le L_b < 0$
Very unstable (vu)	$-100 \le L_b < -50$
Unstable (u)	$-200 \le L_b < -100$
Near-neutral unstable (nnu)	$-500 \le L_b < -200$
Neutral (n)	$L_b < -500$ or $L_b > 500$
Near-neutral stable (nns)	$200 < L_b \le 500$
Stable (s)	$50 < L_b \le 200$
Very stable (vs)	$10 < L_b \le 50$
Extremely stable (xs)	$0 < L_b \le 10$

3 Ringsted measurement campaign

3.1 Description of the measurement campaign

In the Ringsted measurement campaign, ammonia concentrations and ammonia deposition fluxes were measured in the vicinity of a chicken farm near Ringsted, Denmark. The aim of the study was to validate the dispersion and deposition modelling in OML. The campaign is described in reasonable detail in a paper by Sommer et al. [6]. This paper also describes how concentration and deposition outcomes for OML compare with the observations. Relevant input data for atmospheric dispersion modelling (e.g. meteorology, measurement locations and measurement outcomes) was provided to RIVM by Aarhus University. The concentration and deposition measurements were carried out in September and October 2005.

Figure 3.1 Aerial picture of the Ringsted campaign location (image copied from [6]). Grey plusses: locations of concentration samplers (not all locations visible). Yellow diamonds: locations of biomonitors.



The farm was located in a mostly flat area with fields and crops (Figure 3.1). The farm consisted of four buildings, including two mechanically ventilated chicken houses. The ammonia emissions from the chicken houses were calculated from air exchange rates and ammonia concentrations in the inlets and outlets of the chicken houses. Ambient concentrations of ammonia around the farm were measured with passive samplers at fifteen separate locations. For most locations, concentrations were measured during two periods, with durations ranging between ten and seventeen days each. Deposition was measured using pots of rye grass that was harvested and analysed after 54 days of exposure. These plots were placed in triplets at 25 separate

locations. Some of these locations were affected by ammonia emissions from an additional nearby source of unknown emission strength.

3.1.1 Sources and emission

Information on the ammonia emissions from the poultry farm was provided to RIVM by Aarhus University. Ammonia was emitted from eight exhausts in two chicken houses, see Table 3.1. Emissions (emission rate, temperature, and flow rate) varied across the exhausts and also per hour. Ammonia release rates were calculated from measured air exchange rates and inlet and outlet concentrations. The inlet and outlet concentrations were measured at every second inlet/outlet of the farm and then interpolated for the other inlets/outlets [6].

The poultry farm consisted of four buildings (Figure 3.2), including two chicken houses (B1 and B2). Both chicken houses had four outlets located 40 cm above the roof and spread out along the length of the buildings. These outlets are the emission sources. The heights of the buildings[18] were provided to RIVM by Aarhus University, while the widths, lengths and orientations were estimated by RIVM using an aerial picture of the farm (Figure 3.2). The resulting building characteristics are reported in Table 3.2.

Figure 3.2 The poultry farm at Ringsted, On the left: Aerial picture of the farm taken from Google earth (accessed: 11-7-2024). The source locations have been marked by blue pins and the buildings have been marked from B1 – B4. On the right: a picture taken from the measurement campaign of the farm from the south-east. Provided to RIVM by Aarhus University.

[18] It is unknown whether these heights are average heights or maximum heights.

Table 3.1 Locations and heights of the source outlets of the chicken houses

Source	UTM-E [m]	UTM-N [m]	Outlet height [m]	Outlet diameter [m]	Building height [m]	Average emission rate [gN/s]
1	670368	6150415	6.6	0.90	6.2	0.0122
2	670365	6150405	6.6	0.90	6.2	0.0122
3	670361	6150396	6.6	0.90	6.2	0.0124
4	670359	6150388	6.6	0.90	6.2	0.011
5	670405	6150394	6.6	0.90	6.2	0.0067
6	670408	6150402	6.6	0.90	6.2	0.0063
7	670411	6150411	6.6	0.90	6.2	0.0058
8	670415	6150421	6.6	0.90	6.2	0.0064

The emission strengths gradually decreased between 5 September and 25 October (Figure 3.3, left). The emission data also shows a strong daily cycle (Figure 3.3, right).

Figure 3.3 Variation of emission strength during the measurement period

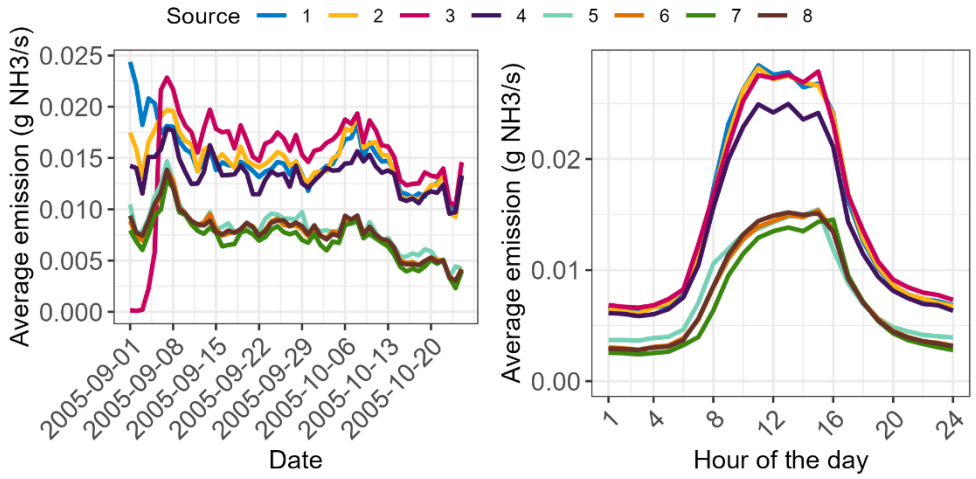


Table 3.2 Characteristics of the buildings at the Ringsted poultry farm

Building	UTM-E [m]	UTM-N [m]	Length [m]	Width [m]	Height [m]	Angle from north
B1	670367	6150405	43.4	21.0	6.2	19°
B2	670412	6150411	43.4	21.0	6.2	19°
B3	670389	6150407	21.8	19.0	6.2	19°
B4	670379	6150444	21.3	17.0	6.2	19°

3.1.2 *Meteorology*

Meteorological parameters were measured hourly from 1 September until 25 October 2005 at a local station close to the farm that was specifically set-up for the Ringsted campaign. The measured variables were:

- Temperature
- Wind speed
- Wind direction
- Relative humidity

- Sensible heat flux
- Friction velocity (derived from measurement data)
- Monin–Obukhov length (derived from measurement data)

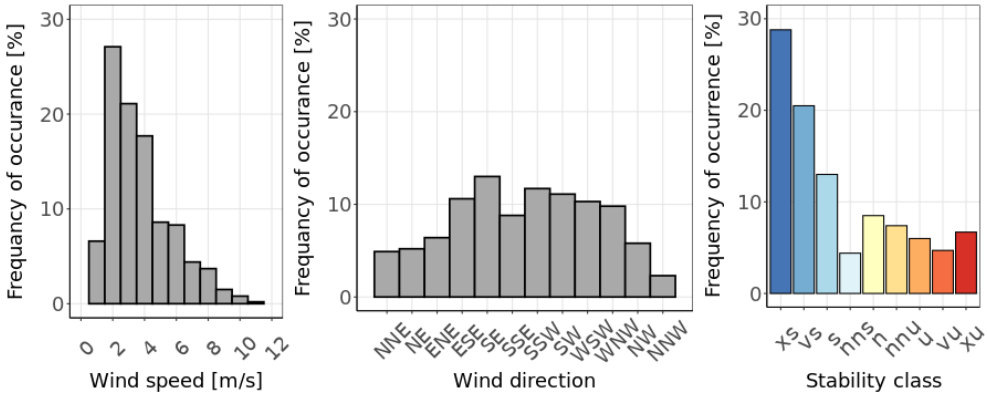
According to meteorological data, provided to RIVM by Aarhus University, mostly concerning observations from a local sonic anemometer with measurements at an altitude of 7.2 m. This data was supplemented with some interpolated measurements from another, undefined, station and with wind speed measurements from an 11 m high distant mast. The received data has not been converted to standard height (10 m).

RIVM retrieved additional ERA5 hourly meteorological data for the entire campaign duration from the Copernicus Climate Change Service [18], using the campaign location (55.46N, 11.69E). These additional variables are:

- Dewpoint temperature
- Total precipitation
- Surface pressure
- Total cloud cover
- Global radiation
- Snow fall

Over the course of the campaign, the wind was blowing mostly from eastern, southern and western directions. The average wind speed was 3.7 m/s. The stability of weather conditions was derived from model outputs from the OPS-ST. Weather conditions are classified as extremely stable (xs) for almost 30% of the time, and as very stable (vs) for 20% of the time (Figure 3.4).

Figure 3.4 Summary of meteorological variables from the Ringsted poultry farm campaign. The data on wind speed and wind direction is *from a local measurement station. Data on the stability classes come from the output from OPS-ST.*



3.1.3 *Land use and surface roughness length*

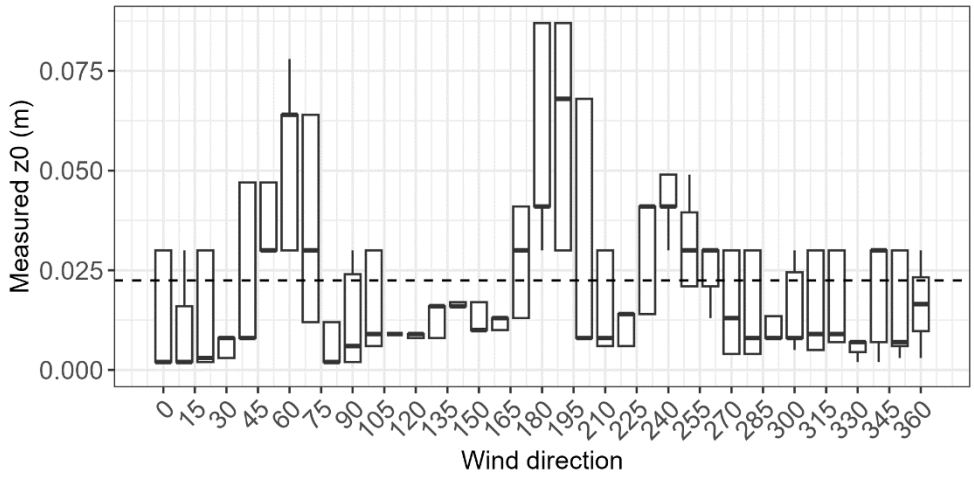
The farm was surrounded by fields, with crops on the east of the farm and grass fields on the west of it. Aarhus University provided RIVM with a dataset used by Sommer et al. [6].This dataset provides the total surface area of 16 separate types of land use in cells of 100×100 m.

Within a 1 km range from the farm, the dominant land use is arable land.

The dataset also contained representative numbers for the surface roughness length of each land use type in the dataset. Using these numbers and the surface area fractions of each land use type, the average surface roughness length for the area within 1 km distance from the farm was assumed to be in the order of 0.1 m.[19]

The surface roughness was also measured with a local sonic anemometer. The outcomes of these measurements depend on wind direction (Figure 3.5). The average observed surface roughness length during the measurement period (from 1 September until 25 October) is 0.02 m.

Figure 3.5 Surface roughness length observations from sonic anemometer



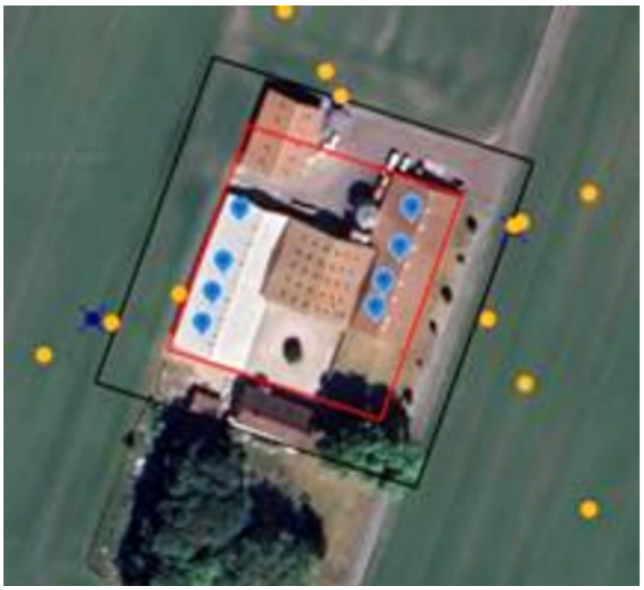
The deposition was measured with ryegrass in controlled pots that were placed around the farm (section 3.1.4.2). Therefore, the vegetation at the deposition measurement locations is assumed to be grass.

3.1.4 Concentrations and deposition measurements

Aarhus University provided RIVM with measurement locations and measurement outcomes (both for concentration and for deposition). The measurement locations closest to the source are shown in Figure 3.6.

[19] This outcome is not sensitive to the distance range: roughly the same outcomes are found for any distance between 100 m and 2.5 km.

Figure 3.6 The measurement locations at the Ringsted farm closest to the building. A hypothetical rectangular building around the sources is shown in red. The receptor locations are shown as orange dots (deposition) and blue crosses (concentration). A 30 m buffer zone around the hypothetical building is shown in black in order to give an impression of dimensions.



3.1.4.1 Concentration measurements

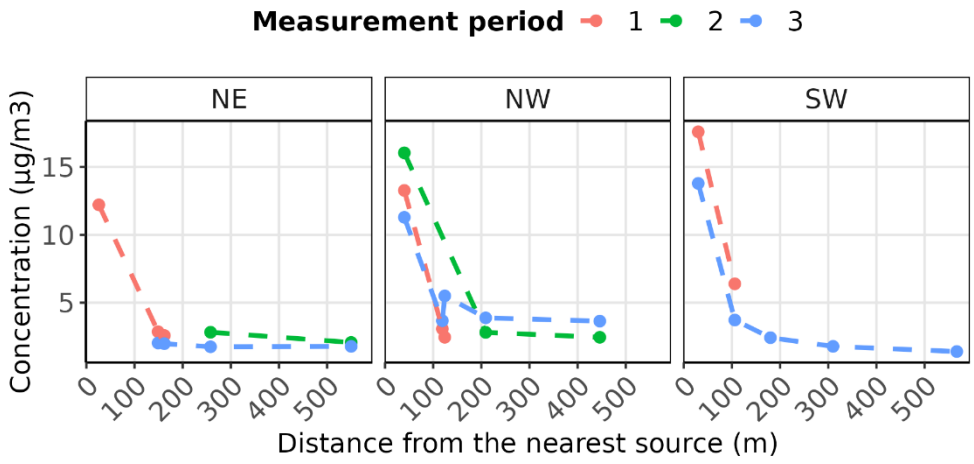
NH_3 concentrations were measured with passive samplers between 5 September and 17 October 2005 (7 weeks). There were fifteen stations spread across three main directions. The concentration was measured at 2 m above the surface. Concentrations were measured during 3 periods of 10 to 17 days each. For most locations, only measurements for one or two of these periods were available (see Table 3.3). The average time coverage of all locations was 57%.

Table 3.3 Available concentration measurements in the Ringsted campaign

Receptor	Direction	Distance to nearest source	Available measure-ments	Total days and time fraction
54_NOP	NE	26	1	10 (24%)
179_NO	NE	149	2	24 (57%)
192_NOP	NE	162	2	24 (57%)
289_NO	NE	258	2	32 (76%)
580_NO	NE	550	2	32 (76%)
47_NVP	NW	40	3	38 (90%)
132_NVP	NW	119	2	24 (57%)
143_NV	NW	124	2	24 (57%)
228_NV	NW	209	2	32 (76%)
465_NV	NW	446	2	32 (76%)
601_SV	SW	30	1	14 (33%)
140_SVP	SW	106	2	24 (57%)
215_SV	SW	180	1	14 (33%)
344_SV	SW	310	1	14 (33%)
60_SVP	SW	567	2	24 (57%)

Figure 3.7 shows how measured concentrations depend on distance. The plot shows rapidly reducing concentrations (with distance) close to the source, and much smaller reductions further away. For many locations, concentration measurements are only available for one or two measurements periods.

Figure 3.7 Measured concentrations around the Ringsted farm during three periods (colours) and in three directions (frames)



3.1.4.2 Deposition measurements

Combined ammonia and ammonium deposition was measured by planting rye grass in controlled pots. The plants were planted on 1 August in nitrogen-free sand and placed in a glass house. They were watered with a nitrogen-free nutrition solution and subsequently placed outside, at locations around the chicken farm, on the 1 September. On days 10 and 25 after sowing, the plants were provided with a KNO_3 solution with 2.786% ^{15}N excess. The rye grass was eventually harvested on 25 October, after 54 days of exposure, and the total nitrogen content in the plant's roots and leaves was measured, as well as the total dry mass and the ^{15}N isotope fraction.

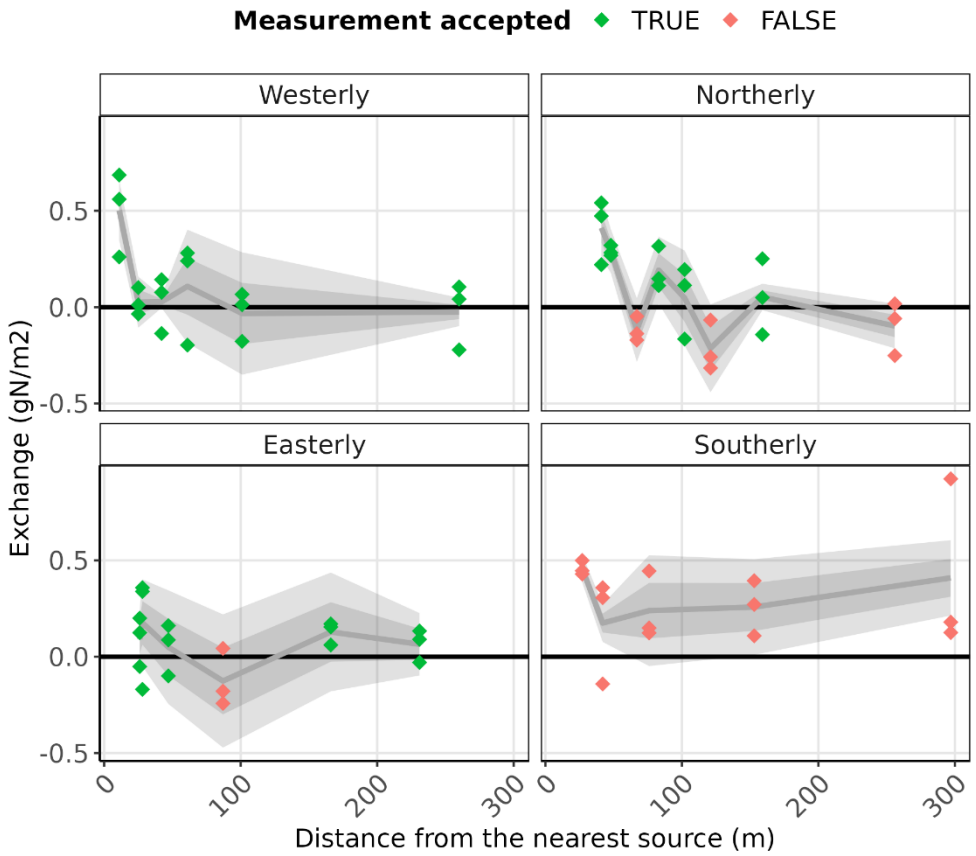
Three pots were placed at each of the 25 locations. The locations were divided across four directions (south-west, north-west, north-east, and south-east). The distances to the centre of the farm ranged from 15 m to 320 m.

The deposition fluxes to different pots relate to a contribution from the source and a contribution from ambient NH_x and NO_y background concentrations. The background contribution was set equal to the deposition at a 'control location' at 320 m distance in WSW direction (240°). According to Sommer et al. [6], the 'biomonitors at this position were affected mainly by background deposition of NH_4^+ and NO_3^- (…) because there were few occasions when the wind was blowing from the farm towards this location, that is, from the 60° direction.' Source contributions to deposition in different pots were subsequently determined by comparing the ^{15}N concentrations in the pots with the average ^{15}N concentration at the control location, where lower ^{15}N concentrations indicate larger source contributions. The resulting

source contributions were averaged over the three pots per location in order to give an overall estimate for the source contribution to deposition per location. The error in these estimates per location was assumed to be equal to the standard deviation of the three outcomes per location divided by √3. [20,21] Possible systematic errors were not included in these estimates for measurement error.

The results are shown in Figure 3.8. Calculated measurement errors are indicated by grey shades (dark grey for 1 standard error and light grey for 2 standard errors).

Figure 3.8 Measured deposition in the Ringsted campaign. Observed deposition for individual pots is indicated by diamonds. Grey lines connect the means of three outcomes. Shades reflect 1 standard error (dark grey) and 2 standard error bandwidths (light grey).



Measurements for nine locations (indicated by red diamonds in Figure 3.8) were not used by Sommer for validation of OML. Five locations in southerly direction were assumed to have been affected by a non-specified ammonia source south of the farm. Four locations with negative average deposition were not considered for the following

[20] Standard error is a measure for the (un)certainty that the mean value of observations approximates the mean value of the population. se = sd/√n. See, for example, https://users.physics.unc.edu/~deardorf/uncertainty/UNCguide.pdf for more information.
[21] The errors shown in Figure 3 of Sommer's paper [6], are believed to have been calculated with the wrong formulae (wrong copying of an initially correct formula to subsequent rows in an underlying Excel spreadsheet). The figures in this report use recalculated errors.

reasons: *'These observations of loss of N from the biomonitors is here considered as outliers, since the atmospheric NH_3 concentrations are high and above the compensation points of healthy plants (Sutton et al., 1993). Thus we speculate that the loss could have been caused either by plant disease or by damage from wild animals eating the plants.'* [6].

The resulting source contribution to deposition at the control location (320 m west) was not zero but slightly negative. This appears to be contradictory to the approach discussed in Sommer's paper [6]. A second location in westerly direction also has a calculated average source contribution to deposition smaller than 0. Both locations were nonetheless still used for validation of OML, and are also used in the current study. Sommer's claim that 'there were few occasions when the wind was blowing from the farm towards this location; that is from the 60° direction' is debatable (see Figure 3.4). Hence, the estimate for the background contribution to total deposition, which affects the calculated source contribution to deposition, does not appear to be accurate.

The reduction of deposition in Figure 3.8 is not as strong as the reduction of concentration (Figure 3.7). Relatedly, lines are not as smooth as before. The calculated absolute measurement error differs per location and does not appear to correlate with either measured exchange or distance. The relative error (the ratio of absolute error in the estimate and the estimate) is, therefore, largest for the smallest deposition outcomes. On average, the relative error in these measurements is slightly over 100%, even in the absence of any systematic errors (see before).[22] Therefore, the Ringsted deposition measurements do not appear to be as accurate as the concentration measurements used in this study.

3.1.5 Background concentrations

The average NH_3 background concentration was derived from measurements at the furthest distance from the chicken farm. During the measurement period, the background level was 1.4 µg NH_3-N/m^3 on average [6], which converts to 1.7 µg NH_3 /m^3.

3.2 Model validation results

The Ringsted campaign was analysed with seven models: ADMS, AERMOD, IFDM, OML-Multi, OPS-LT, OPS-ST, and STACKS-D. The model performance for concentrations is discussed in section 3.2.1, while the performance for deposition fluxes is discussed in section 3.2.2. The model performance is analysed using five performance indicators (see section 2.2). These indicators do not account for measurement uncertainty, and therefore, only quantify model performance if the measurements are sufficiently accurate.

3.2.1 Concentrations

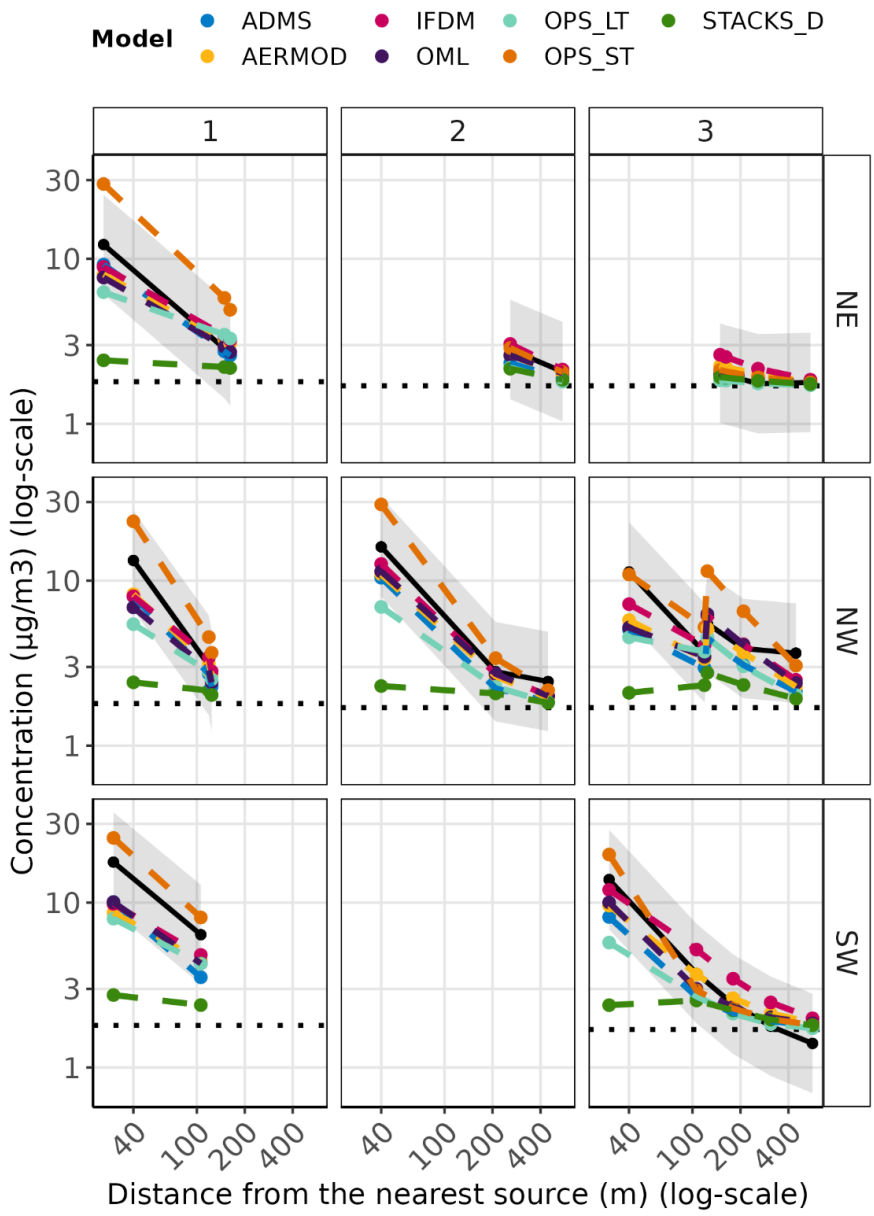
During the Ringsted campaign, concentrations were measured at 15 separate locations (section 3.1.4). Measurements were spread over three periods, each with a duration between 10 and 17 days. For many

[22] The relative standard error is defined as the ratio of the standard error and the exchange (=mean value). The average value of the relative standard error is 106% if all 25 locations are considered and 137% if only the 16 accepted measurements are considered.

receptors, only one or two measurements were available. Total period average concentrations could not be derived from the measurements. Therefore, the comparison of model outcomes with observations is carried out for individual measurements (durations between 10 and 17 days). All relevant outcomes are reported in Appendix 2.

The comparison of model outcomes with measured outcomes is shown in Figure 3.9. The various frames represent different measurement periods (columns) and directions (rows). Modelled concentrations at different distances are indicated by coloured dashed lines, while measured concentrations are indicated by a solid black line. The grey shade represents a factor of two bandwidth around the measured concentrations. The discontinuities for measured and modelled concentrations in NW direction (period 3) relate to the grouping of two slightly different 'NW' directions (see also Figure 3.1). Additional plots per individual model are available in Appendix 2. Appendix 2 also discusses which receptors contribute the most to total NMSE and VG (see below).

Figure 3.9 Comparison of measured concentrations (black points and lines) with model outcomes (coloured points and lines). Dotted black line: assumed background concentration. Grey shade: a factor of two bandwidth around measured outcomes.



The following can be observed from Figure 3.9 and the additional plots in Appendix 2:

- Modelled concentrations are mostly within a factor of two from measured concentrations. Many deviations are much smaller than a factor of 2 (see further below).
- Closest to the source, measured concentrations are mostly higher than modelled concentrations. Between the models, STACKS-D stands out at this distance, providing the lowest concentrations (much lower than measured). OPS-ST is the only model that

calculates higher concentrations than measured, close to the source.
- At larger distances, modelled and measured concentrations are closer together, both in absolute and in relative terms.
- The plot for period 3 (right-hand column) and NW direction (middle row) show discontinuities in measured and modelled outcomes between 100 and 150 m distance. These discontinuities relate to grouping two slightly different directions ('NØ' and 'NØP', see Appendix 3). Models capture the influence of direction on concentrations.

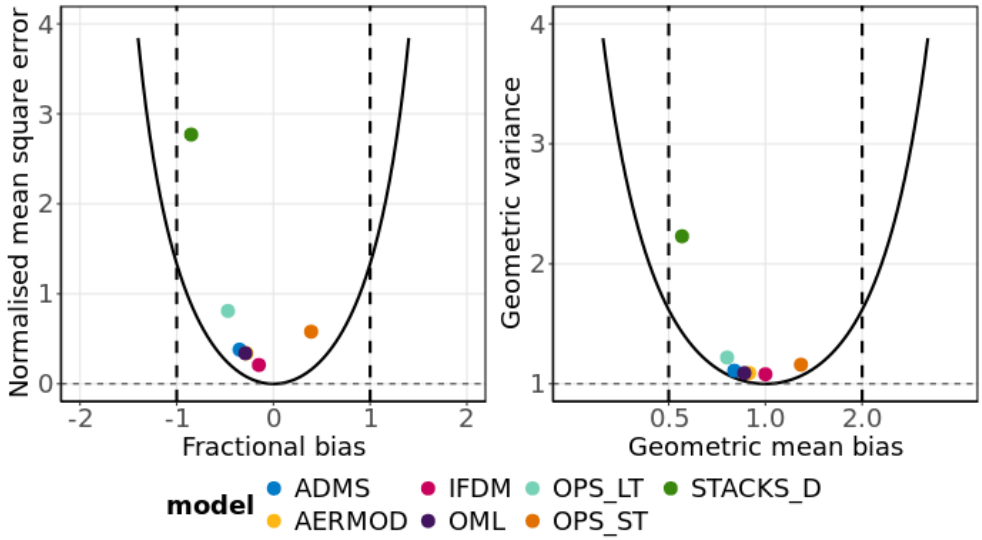
The performance of models regarding 'biweekly' concentration measurements for different locations (27 measurements in total) is summarised in terms of Fractional Bias (FB), Normalised Mean Square Error (NMSE), Geometric Mean Bias (MG), and Geometric Variance (VG). The outcomes are shown in Table 3.4 and Figure 3.10. Positive bias (FB>0 or MG>1) means that model outcomes are larger than observed outcomes (note: this convention is opposite to the convention used by Chang & Hanna [26]). Deviations are a combined result of model bias and random fluctuations. The solid line in Figure 3.10 shows the error associated with mean bias only.[23]

Table 3.4 Model performance for the Ringsted biweekly concentration measurements

Model	**FB**	**NMSE**	**MG**	**VG**	**FAC2**
ADMS	-0.35	0.38	0.80	1.1	0.96
AERMOD	-0.28	0.34	0.89	1.1	0.96
IFDM	-0.15	0.21	1.0	1.1	1.0
OML-Multi	-0.29	0.34	0.86	1.1	0.96
OPS-LT	-0.47	0.81	0.76	1.2	0.81
OPS-ST	0.39	0.58	1.3	1.2	0.89
STACKS-D	-0.85	2.8	0.55	2.2	0.74

[23] The vertical distance of outcomes to the solid line is a measure for the number of random fluctuations (scatter) on top of bias. A large amount of scatter can be a sign of model inaccuracy but can also be a warning that observations are inaccurate.

Figure 3.10 Deviations and biases of biweekly concentration outcomes of the Ringsted campaign. Left pane: FB and NMSE, right pane: MG and VG. Visually, ADMS is partly hidden behind OML, and AERMOD behind IFDM and OML.



Further analysis of the performance outcomes (Table 3.4 and Figure 3.10) and more detailed outcomes in Appendix 2 yield the following observations:

- Mean deviations (either NMSE or VG) are largest for STACKS-D. This is mostly caused by a significant underprediction of measured concentrations close to the source. Other models are quite close together, with outcomes for NMSE typically between 0.2 and 0.8 and outcomes for VG between 1.1 and 1.2.
- Most models have a tendency to calculate lower concentrations than measured. OPS-ST is the only model that on average calculates higher concentrations than measured.
- The fraction of calculated outcomes within a factor of 2 of the measured outcomes is smallest for STACKS-D (0.74) and largest for IFDM (1).
- The three locations closest to the source (26 m NE, 30 m SW and 40 m NW) determine to a large extent the total NMSE. Together, these three locations contribute 89% to 97% to the total NMSE (see Appendix 2). Together, the other 13 locations contribute between 3% and 11%.

The performance indicators of Table 3.5 are compared with the performance criteria from Hanna & Chang (2012) for rural terrain [39]. As was noted in section 2.2.1, these criteria are intended as guidance; some failures to meet conditions are tolerated. ADMS, OPS-ST, and OPS-LT have too large biases in terms of Fractional Bias (FB), but not in terms of Geometric Mean Bias (MG). STACKS has too large values for both bias indicators, and also for Geometric Variance (VG).

Table 3.5 Comparison of calculated performance indicators with criteria from [39]

Test	**Models that fail the test (ordered from lesser to larger violations of the criterion)**
|FB| < 0.3	ADMS, OPS-ST, OPS-LT and STACKS-D
0.7 ≤ MG ≤ 1.3	STACKS-D
NMSE < 3	
VG < 1.35	STACKS-D
FAC2 ≥ 0.5	

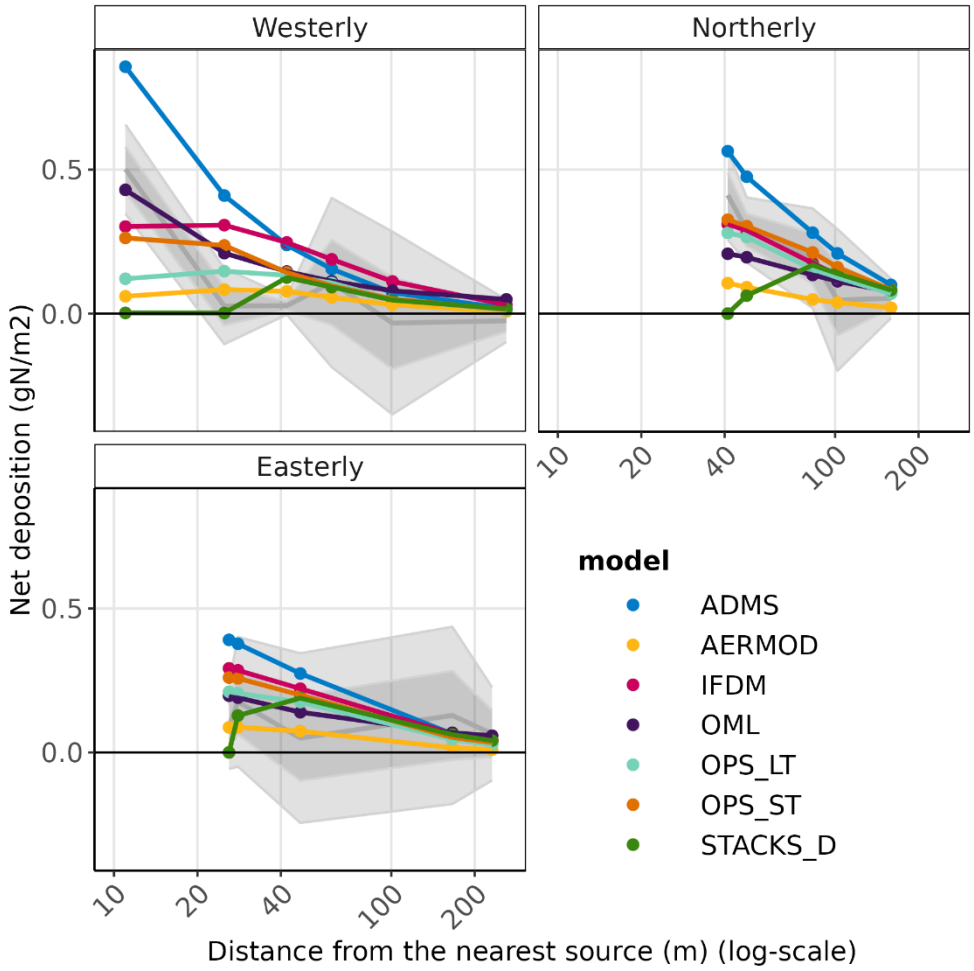
3.2.2 Deposition

Deposition was measured at 25 locations, using 3 samples for each location (see section 3.1.4). Nine locations were discarded by Sommer et al. [6], either due to the presence of a local source south of the poultry farm (five locations) or because the measured outcomes were considered to be outliers (four locations).

In our analysis, we use the same (sixteen) measurement locations as the ones used in Sommer et al. They include two locations where the measured deposition flux is negative, even though negative values seem unrealistic (from a biochemistry point of view and considering the duration of the measurement). These two locations were excluded when calculating MG and VG, because logarithms of negatives values are invalid. All relevant outcomes are reported in Appendix 2.

Figure 3.11 shows for each model how model outcomes compare with measured outcomes. Outcomes in different directions are shown in different rows. The unit for deposition on the y-axis is the same as the unit used by Sommer et al. (gN/m^2). Model outcomes are visualised by means of coloured points and connected by coloured lines as a visual aid. The mean measured outcomes are visualised by means of a solid black line and the grey shades represent one (dark grey) or two (light grey) standard errors around the mean measured value. Detailed graphs per individual model are available in Appendix 2.

Figure 3.11 Comparison of modelled and measured deposition. See text for description



The following observations can be made from Figure 3.11 and the more detailed graphs in Appendix 2:

- The calculated standard error, which is derived from three measured outcomes in a triplet, is subject to random fluctuations: it can be small at one location and large at the neighbouring location.
- Mean measured deposition fluctuates with distance. The fluctuations are largest in westerly and easterly directions. It is likely that they are the result of measurement inaccuracy. Model outcomes generally show a smoother decrease with distance.
- The largest absolute differences between modelled and measured outcomes occur close to the source: in westerly direction at 11 and 25 m distance from the nearest source and in northerly direction at 41 and 48 m distance from the source.

Outcomes for the key performance indicators (FB, NMSE, MG, VG, and FAC2) are reported in Table 3.6. Model outcomes were compared with the average of the three (triplet) measurements at each location. For the geometric indicators (VG and MG), two receptors with negative measured deposition flux had to be excluded. For STACKS-D, two

further receptors with zero predicted deposition also had to be excluded when calculating VG and MG.

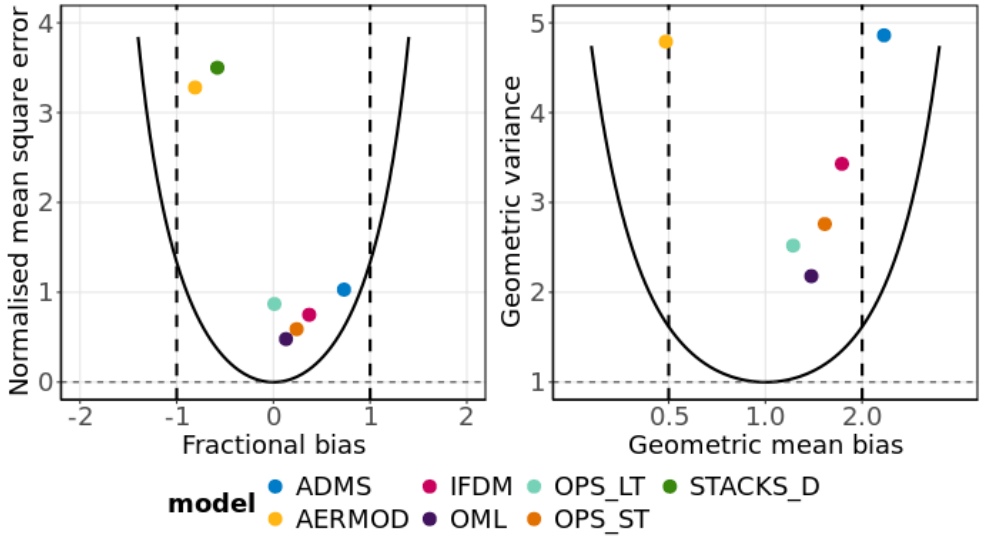
Table 3.6 Model performance for the Ringsted deposition measurements

Model	FB	NMSE	MG	VG	FAC2
ADMS	0.73	1.03	2.3	4.9	0.44
AERMOD	-0.81	3.28	0.49	4.8	0.25
IFDM	0.37	0.75	1.7	3.4	0.56
OML-Multi	0.13	0.48	1.4	2.2	0.56
OPS-LT	0.01	0.87	1.2	2.5	0.38
OPS-ST	0.24	0.59	1.5	2.8	0.50
STACKS-D	-0.58	3.5	0.57 *	38 *	0.31

* Two receptors with outcome 0 had to be excluded when calculating VG and MG for STACKS-D.

Figure 3.12 shows the outcomes for deviation and bias (left-hand pane: Normalised Mean Square Error (NMSE) and Fractional Bias (FB), right-hand pane: Geometric Mean Bias (MG) and Geometric Variance (VG)).). Deviations are a combined result of model bias and random fluctuations. The solid line shows the deviation associated with bias only.[24] Outcomes for STACKS-D are not visible in the right-hand pane because the Geometric Variance (VG) is outside of the range of the plot.

Figure 3.12 Deviations and biases of deposition outcomes for the Ringsted campaign. Left-hand *pane: FB and NMSE, right*-hand *pane: MG and VG. In the right*-hand *pane, STACKS-D is outside the visible range.*



Analysis of the performance outcomes (Table 3.6 and Figure 3.12) results in the following observations:

- STACKS-D and AERMOD have the largest Normalised Mean Square Error (NMSE). STACKS-D has some very low outcomes (compared to measurements and other models) very close to the source. AERMOD has lower than average outcomes at all distances.

- STACKS-D has by far the largest Geometric Variance (VG). This is the result of the significant underprediction of two deposition fluxes close to the source (see previous item).
- In terms of Geometric Variance, OML-Multi, OPS-LT, and OPS-ST have the best results, followed by IFDM, AERMOD, ADMS and, at a distance, STACKS-D (outside the visible range of the plot).
- While STACKS-D and AERMOD have a tendency to underpredict, other models have a tendency to overpredict measured deposition fluxes, with ADMS yielding the largest overprediction.
- The fraction of calculated deposition fluxes that is within a factor of 2 of the average observed deposition flux, is lowest for AERMOD (25%) and highest for IFDM and OML-Multi (56%). Fractions for other models lie within that range.
- In Figure 3.12, vertical distances from model outcomes to the solid line are relatively high (more so than in other validation studies). These distances represent random fluctuations independent of model bias. Large vertical distances can be a sign of measurement inaccuracies.

It should be noted that the above outcomes for model performance use the mean measurement outcome as input. Several figures (e.g. Figure 3.8 and Figure 3.11) suggested that the accuracy of the deposition measurements in the Ringsted campaign is limited.[25]

The calculated performance indicators have not been compared with performance criteria from Chang & Hanna, because these criteria were proposed for concentration measurements, not for deposition measurements.

3.2.3 Comparing calculated biases for concentration and deposition

Model predictions for deposition depend on modelled concentrations as well as on inputs for and formulations of deposition processes. Models that accurately represent the deposition processes would, therefore, predict consistent model outcomes for both concentration and deposition; in other words, if the concentration was being under-predicted, the deposition should also be under-predicted, and vice versa.

A comparison of the calculated biases for modelled concentrations and modelled deposition is, however, believed to be useful only if the measurement locations and durations are similar between concentration and deposition measurements and if both measurements are sufficiently accurate. For the Ringsted campaign, concentration and deposition measurements differ with regard to their number, locations and durations. In addition, the accuracy of the deposition measurements appeared to be limited. Graphs in which the Fractional Biases (FB) and the Geometric Mean Biases (MG) of the models are compared, are available in Appendix 3 (section 13.3). Given the above limitations, we do not recommend drawing strong conclusions from these graphs.

[25] Nonetheless, we still believe that it is the best available dataset.

4 Affligem motorway case

4.1 Description of the measurement campaign

In the Life+ ATMOSYS campaign [16], concentrations of various substances near a highway were measured. NO_2 measurements were carried out between 20 April 2012 and 28 December 2012 (total duration 36 weeks). The project was carried out by the Flemish Environmental Agency (VMM). VMM provided the measurement data to RIVM for use in the current study. VITO provided emission data for traffic on this motorway.

The location of the measurements was near Affligem (Belgium), roughly 10 km west of the outskirts of Brussels and about 20 km southeast of Ghent. At this location, the motorway has 6 lanes (3 in each direction). One fixed (AF07) and one mobile (L800) sensor were placed south of the motorway. Eight additional sensors were placed north of the motorway. The maximum distance to the motorway was 146 m. The sensors north of the motorway were mostly placed on the side of a local dead-end road perpendicular to the motorway (see Figure 4.1).

Figure 4.1 Measurement locations for the Life+/ATMOSYS campaign. Image copied from [16].

4.1.1 *Emissions*

VITO provided a table with XY coordinates of 56 motorway segments, along with an average emission rate in kg/km/hour for each segment. These emissions relate to the total emissions of nitrogen oxides (NO_x) in NO_2-equivalent mass. VITO estimates that the average NO_2/NO_x ratio in these emissions was 29.8%. VMM provided results of traffic counts for different road sections at different times. This data includes the number of vehicles for different vehicle types and average velocity of those vehicles.

Hourly varying emissions can be derived by multiplying the average emission rates by 3 corrections factors:

- f1: a correction factor for different hours in the day;
- f2: a correction factor for different days in the week (weekdays);
- f3: a correction factor for different months in the year.

The hour correction factors (Table 4.1) are different for summer and winter months.[26] November, December, January, February, and March are assumed to be 'winter months'. Months from April until October are assumed to be 'summer months'.

Table 4.1 Hour correction factors (f1)

Hour	**Hour correction factor (f1) – winter months**	**Hour correction factor (f1) – summer months**
0-1h UTC	0.236	0.191
1-2h UTC	0.191	0.21
2-3h UTC	0.21	0.307
3-4h UTC	0.307	0.574
4-5h UTC	0.574	0.985
5-6h UTC	0.985	1.24
6-7h UTC	1.24	1.335
7-8h UTC	1.335	1.359
8-9h UTC	1.359	1.438
9-10h UTC	1.438	1.465
10-11h UTC	1.465	1.453
11-12h UTC	1.453	1.477
12-13h UTC	1.477	1.529
13-14h UTC	1.529	1.543
14-15h UTC	1.543	1.603
15-16h UTC	1.603	1.604
16-17h UTC	1.604	1.446
17-18h UTC	1.446	1.172
18-19h UTC	1.172	0.89
19-20h UTC	0.89	0.683
20-21h UTC	0.683	0.541
21-22h UTC	0.541	0.41
22-23h UTC	0.41	0.311
23-24h UTC	0.311	0.236

Weekday correction factors (f2) are shown in Table 4.2.

[26] These hourly factors are shifted by one hour between summer and winter. If expressed in local time (either 'summer time' or 'winter time'), the factors would have been identical.

Table 4.2 Weekday correction factors (f2)

Weekday	**Weekday correction factor (f2)**
Monday	1.03
Tuesday	1.05
Wednesday	1.07
Thursday	1.06
Friday	1.12
Saturday	0.87
Sunday	0.8

Month correction factors (f3) are shown in Table 4.3.

Table 4.3 Month correction factors (f3)

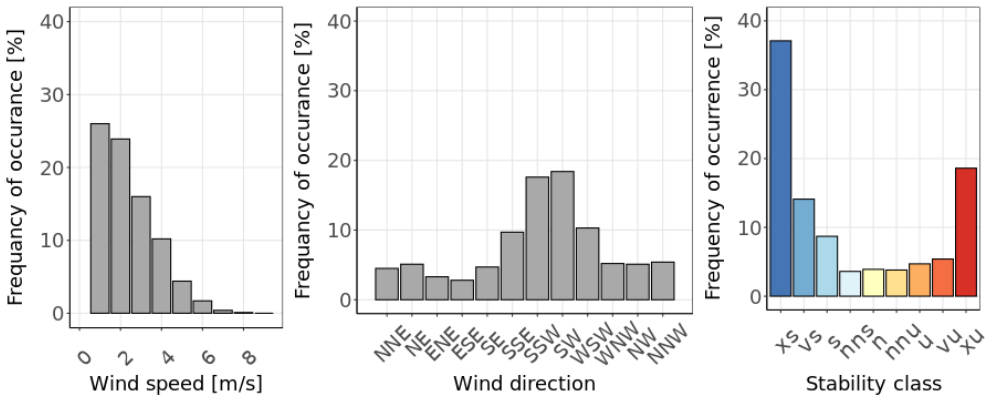
Month	**Month correction factor (f3)**
January	0.87
February	0.94
March	1.01
April	1.05
May	1.03
June	1.05
July	1.00
August	1.08
September	1.04
October	1.03
November	0.95
December	0.95

4.1.2 Meteorology

Multiple data sources were used to obtain the required meteorological input data for the models. Wind speed and wind direction at 9 m height were measured with a mobile station (L800) at the campaign location. The Flanders Environment Agency (VMM) provided RIVM with this data. For precipitation, observations from AWS Uccle were used. This Autonomous Weather Station is located at 25 km distance from the campaign location. ERA5 hourly meteorological data for the campaign location (50.889N, 04.135E) from the Copernicus Climate Change Service [18], [19] was used for solar radiation, dewpoint temperature, snow fall/depth, atmospheric pressure, sensible heat flux and friction velocity, and for missing wind direction and wind speed measurements by the local station (L800).

The average wind velocity that was measured on-site (2.2 m/s) was much lower than the average wind speed at AWS Uccle (3.3 m/s) and in the ERA5 dataset (3.8 m/s). The reasons for this difference are unknown but not related to the various measuring heights. The dominant wind direction during the measurement campaign was south-south-west (SSW). During the campaign, the dominant stability class was extremely stable, followed by hours with extremely unstable weather conditions (Figure 4.2).

Figure 4.2 Summary of meteorological variables from the Affligem motorway campaign. The data on wind speed and wind direction are from a local measurement station. Data on the stability classes come from the output from OPS-ST.



4.1.3 *Land use and surface roughness*

The vegetation within the measurement domain appeared to be mostly grass. Therefore, it was suggested to assume grass as vegetation in deposition calculations.

VITO provided RIVM with a map with surface roughness estimates at 250×250 m^2 resolution. According to this data, the average surface roughness length in a 1 km square around the measurement site was 0.28 m.

4.1.4 *Concentration measurements*

Fixed and mobile measurement equipment was placed at 6 to 150 m distance from the road (see Figure 4.1). The measured substances include nitrogen dioxide, ammonia, volatile organic components (VOC), and particulate matter (PM_{10} and $PM_{2.5}$). However, observations for ammonia, volatile organic components (VOC), and particulate matter were not used in this validation exercise, because the associated background concentrations were unavailable. In addition, ammonia measurements were apparently affected by emissions from a local horse stable.

Figure 4.3 Measurement locations for the Life+ ATMOSYS campaign. Image copied from [16].

Measurements of NO_2 were performed during 36 consecutive weeks with IVL passive samplers in triplicate at six separate locations and at a sampling height of 2 m. Measurements started on Friday 20 April 2012 and ended on Friday 28 December 2012 (total duration 36 weeks). It is assumed that samplers were replaced each Friday at noon, local time. Locations and their distance to the nearest lane of the highway are reported in Table 4.4.

Table 4.4 Locations of NO_2 measurements in the Affligem campaign

Location	x-coord.	y-coord.	Sampling height (m)	Distance to highway (nearest lane)
AF02	133532.5	175392.9	2	6 m
AF03	133540.5	175413.5	2	28 m
AF04	133551.0	175438.8	2	55 m
AF05	133564.4	175477.2	2	96 m
AF06	133585.1	175523.0	2	146 m
AF07	133520.7	175346.5	2	15 m

The triplicate measurements were in relatively good agreement: the relative standard deviation of pairs of three corresponding outcomes (same location, same period) ranges from 0.06% to 9%, with an average value of 2.7%. Model outcomes are compared with the mean value of the three triplicate measurements. No NO_2 measurement data was available for station AF03 regarding period 12-007.

For 5 out of the 36 periods, the measured NO_2 concentration at station AF06 was larger than the measured concentration at station AF05 (see Table 4.5). This could indicate the presence of a local source near AF06, as concentrations are expected to reduce with distance to the motorway. VMM classified the AF06 measurement for one of these periods (12-027) as being an outlier. The other measurements were not labelled outliers. For the comparison with model outcomes, RIVM will follow VMM; in other words, use all periodic NO_2 measurements except those for AF06 in period 12-027.

Table 4.5 NO_2 measurements (average of three samples) for station AF05 and AF06 when the latter is larger than the first

Period	Measurement for AF05	Measurement for AF06	Ratio of AF06 and AF05
12-002	27.0 µg/m^3	30.2 µg/m^3	1.1
12-006	19.4 µg/m^3	38.9 µg/m^3	2.0
12-007	27.6 µg/m^3	45.4 µg/m^3	1.6
12-008	21.4 µg/m^3	31.4 µg/m^3	1.9
12-027	32.7 µg/m^3	66.2 µg/m^3	2.0

The six measurement locations of the Affligem campaign are quite close to the motorway. Five extra receptors were used to provide more information on model behaviour at farther distances (see Chapter 7). The additional receptors were located at 200, 300, 500, 700, and 1000 m from the motorway in perpendicular direction towards the north-east.

Table 4.6 Additional receptors for the Affligem study

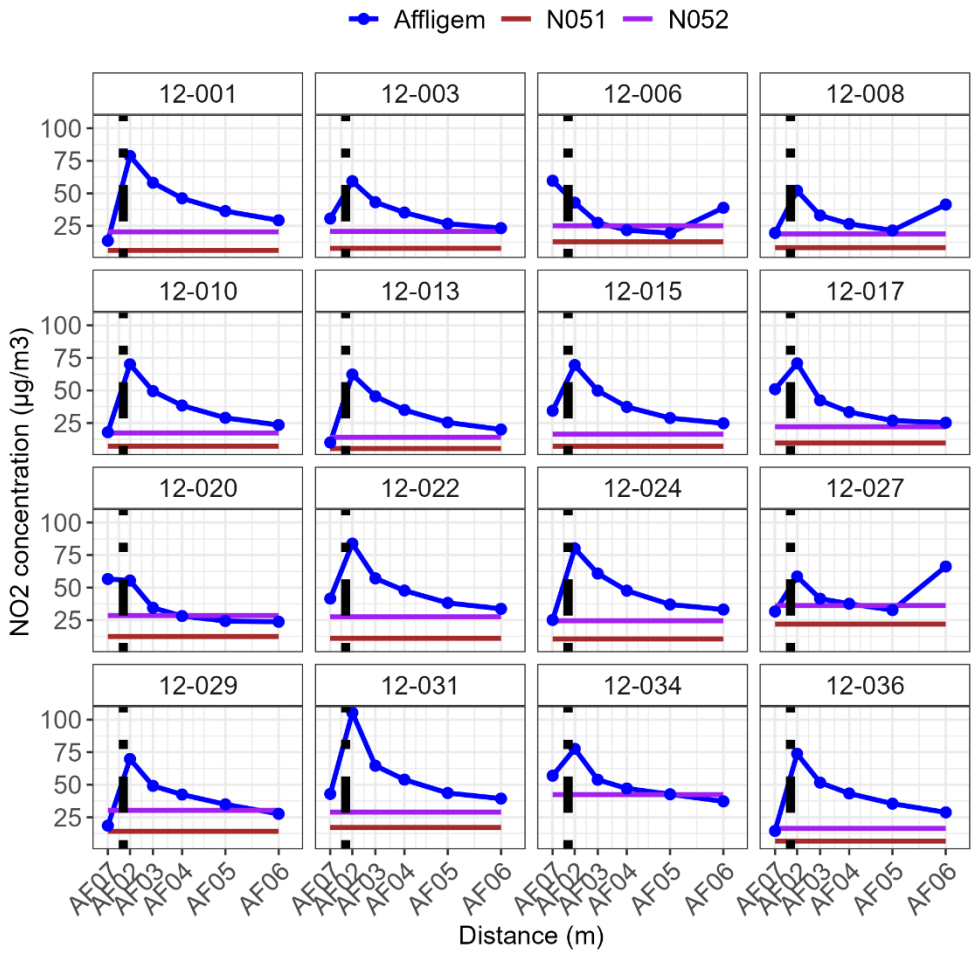
Label	X (m)	Y (m)	Height above ground (m)
AFX1	133607.6	175578.7	2
AFX2	133645.1	175671.6	2
AFX3	133720.2	175857.3	2
AFX4	133795.3	176043.1	2
AFX5	133908.0	176321.8	2

4.1.5 *Background concentrations*

Measurements of ozone (O_3) and nitrogen oxides (NO, NO_2, NO_x) at two background stations (N051 and N052) were provided to RIVM by VMM. N051 is located 17.5 km southwest from the campaign location, while N052 is about 60 km towards the west. During the period of NO_2 measurements at the Affligem site, the time-average NO_2 concentration at station N051 10.7 µg/m^3. The corresponding value for station N052 was 24.9 µg/m^3. Relative trends in time are similar between the two stations.

Figure 4.4 shows how the measured NO_2 concentrations at stations N051 and N052 compare with the measured NO_2 concentrations south (AF07) and north (AF02:AF06) of the Affligem motorway.[27]

Figure 4.4 Average measured NO_2 concentrations north and south of the Affligem motorway and 'background NO_2 concentrations' at stations N051 and N052. The dashed vertical line represents the location of the motorway (x=0).



The following can be observed in Figure 4.4:

- Measured NO_2 concentrations at station N052 (purple) are higher than those at station N051 (brown). The first (N052) are as high as the lowest concentrations around the Affligem motorway (blue) or, in some cases, even higher. The latter (N051) are lower than the measured concentrations around the motorway. Therefore, N051 appears to be a better approximation for background concentrations than N052.
- The measured concentration at location AF07 (just south of the motorway) is significantly lower than the measured concentrations just north of the motorway (AF02). This is consistent with the dominant wind direction during the campaign (the south-south-west direction).

[27] In order to restrict space, only 16 out of the 36 periods are shown in this graph.

- North of the motorway (from AF02 to AF06), concentrations normally reduce with distance to the motorway. This is, however, not true in some periods (see discussion above and Table 4.5). An increasing concentration between AF05 (96 m north of the motorway) and AF06 (146 m north of the motorway) could indicate that a local source was present near AF06, but there is no mentioning of such a source in the campaign description[16].

In the absence of further information, observed concentrations from station N051 were used as model input for background concentrations. Valid observations were, however, missing for 15% of the time. Missing observations were replaced by the average of the last prior and first posterior observation if the duration of missing measurements was no longer than 2 hours. The remaining missing observations were substituted with the average value of all valid observations. The resulting period-average NO_x and O_3 concentrations are listed in Table 4.7.

Table 4.7 Period-average NO_x and O_3 concentrations at station N051

Component	**Average concentration over the whole period**	
	Without replacing missing values	After replacing missing values
NO	2.4 µg/m^3	2.5 µg/m^3
NO_2	10.7 µg/m^3	10.7 µg/m^3
NO_x	13.5 µg/m^3	13.6 µg/m^3
O_3	45.9 µg/m^3	46.6 µg/m^3

4.2 Model validation results

Seven models provided outcomes for this case: ADMS, AERMOD, IFDM, OPS-LT, OPS-ST, SRM2, and STACKS-D. SRM2 is used in the Netherlands for permitting purposes relating to motorways. Calculations were carried out by the partners in 2024 and the discussion in this section relates to these calculations and outcomes. In 2025, OPS-LT was updated with new parametrisations to improve near-road modelling, derived from simulations with SRM2. The modifications and corresponding outcomes for the Affligem campaign are discussed in section 4.2.2.

4.2.1 General results

The model performance is analysed using five performance indicators (see section 2.2). These indicators do not account for measurement uncertainty, and therefore, they only quantify model performance if the measurements are sufficiently accurate.

The current section discusses the performance of models regarding average concentrations over the full period of 36 weeks (see section 2.3). Model performance results for weekly concentration measurements are presented in Appendix 3, for models that provided hourly output.

The measured concentrations include background from other sources. The background concentration level was estimated from station N051 at 17.5 km distance. The accuracy of the assumed background is unknown

but affects model outcomes for source plus background and model performance outcomes.

Figure 4.5 Calculated period-average concentrations (coloured lines), period-average observations (grey dashed lines) and assumed background concentrations (grey dotted lines) for the Affligem campaign

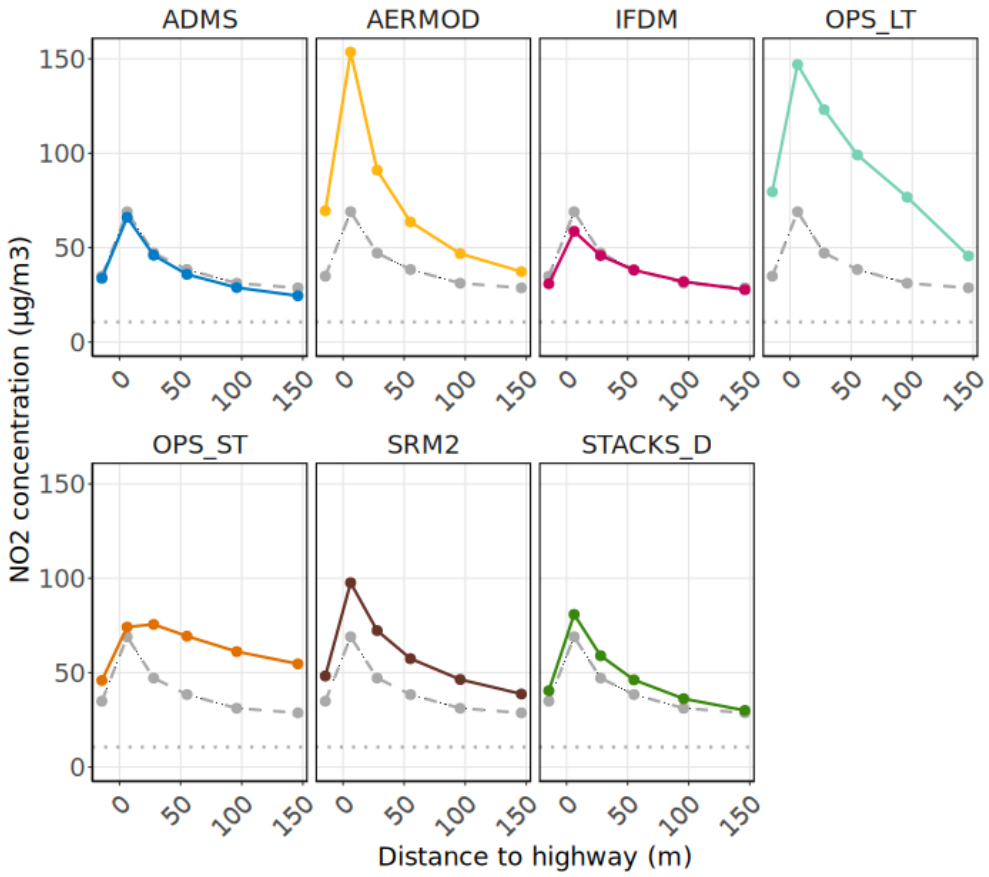


Figure 4.5 shows the measured (grey lines) and modelled (coloured lines) concentrations as a function of the distance from the motorway. The following can be observed in this graph:

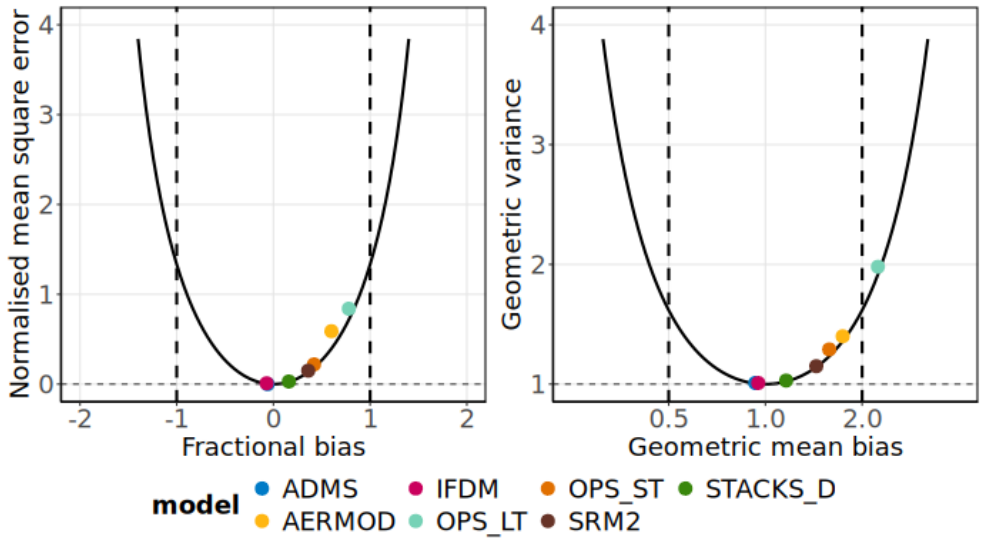
- ADMS, IFDM, and STACKS-D agree well with measurements, in terms of absolute values as well as trends with distance.
- AERMOD and SRM2 overpredict the concentration close to the road but show a logarithmic decrease of concentration with distance that is similar to the measurements.
- OPS-LT also overpredicts the concentration close to the motorway and consequently has a steeper decline with distance than in measurements.
- OPS-ST compares well with measurements close to the motorway but has a slower decline of concentrations with distance than the measurements indicate.

The outcomes for the key performance indicators are reported in Table 4.8 and visualised in Figure 4.6. Further details about receptor contributions are provided in Appendix 3.

Table 4.8 Model performance for period-average concentrations during the Affligem campaign

Model	FB	NMSE	MG	VG	FAC2
ADMS	-0.06	0.00	0.93	1.01	6/6
AERMOD	0.60	0.59	1.74	1.40	5/6
IFDM	-0.07	0.01	0.95	1.01	6/6
OPS-LT	0.78	0.84	2.24	1.98	1/6
OPS-ST	0.42	0.22	1.58	1.29	6/6
SRM2	0.36	0.15	1.44	1.15	6/6
STACKS-D	0.16	0.03	1.16	1.03	6/6

Figure 4.6 Deviations and biases of full period average concentrations for the Affligem campaign. Left pane: FB and NMSE, right pane: MG and VG. Outcomes for ADMS and IFDM overlap in both panes.



The following can be concluded for the calculated model performance:

- ADMS, IFDM, and STACKS-D have the best performance outcomes, followed by SRM2 and OPS-ST. OPS-LT and AERMOD have larger deviations with model outcomes than the above models.
- The models with the best outcomes (ADMS, IFDM, and STACKS-D) have no significant bias. The other models have a positive bias, which means (with our definition of bias) that modelled concentrations are larger than measured concentrations.
- Five models (ADMS, IFDM, SRM2, OPS-ST, and STACKS-D) score all six calculated period-average concentrations within a factor of two from measured concentrations. AERMOD scores five out of six. OPS-LT has the lowest correspondence with measurements and only scores one out of the six outcomes within a factor of two of the measurements.

The calculated performance indicators of Table 4.8 are compared with the performance criteria from Hanna & Chang (2012) for rural terrain [39]. As was noted in section 2.2.1, these criteria are intended as guidance; some failures to meet conditions are tolerated. SRM2, OPS-ST, AERMOD, and OPS-LT have too large biases, both in terms of

Fractional Bias (FB) and in terms of Geometric Mean Bias (MG). AERMOD and OPS-LT also have too large Geometric Variance (VG), and OPS-LT has too few modelled concentrations within a factor of two of the measured concentration.

Table 4.9 Comparison of calculated performance indicators with criteria from [39]

Test	**Models that fail the test (ordered from lesser to larger violations of the criterion)**
|FB| < 0.3	SRM2, OPS-ST, AERMOD, OPS-LT
0.7 ≤ MG ≤ 1.3	SRM2, OPS-ST, AERMOD, OPS-LT
NMSE < 3	
VG < 1.35	AERMOD, OPS-LT
FAC2 ≥ 0.5	OPS-LT

4.2.2 Results for a new version of OPS-LT

In recent years, two new options for modelling concentrations near motorways were added to OPS-LT. The first option involves the NO/NO_2 ratio in the vicinity of a motorway. Previously, a constant NO/NO_2 ratio was used, which was equal to the typical NO/NO_2 ratio at regional scale. The new version allows for using the NO/NO_2 ratio of SRM2 [51], [52]. The second option relates to the enhanced turbulence created by moving traffic. This turbulence increases the mixing of concentrations near the source. SRM2 already accounted for the influence of enhanced turbulence [51], [52]. A new parametrisation was recently added to OPS-LT to also allow for including this effect in OPS-LT calculations. The new parametrisation was derived from calibrations to SRM2 [53] outcomes. Both improvements only affect OPS-LT outcomes up to a 5 km distance from the relevant emission source along the road.

AERIUS Calculator is used in the Netherlands to calculate the NH_x and NO_y depositions relating to emissions from individual sources. Outcomes for AERIUS Calculator are used for permitting processes. Regarding emissions from road traffic, concentrations within the first 5 km from the source (a location along the motorway) are calculated with SRM2. Corresponding deposition fluxes and depletion corrections are calculated with OPS-LT. Since 2025, both new options of OPS-LT (the NO/NO_2 ratio and the enhanced turbulence) have been used to calculate these deposition velocities and depletion factors up to a 5 km distance. For distances beyond 5 km, concentrations, deposition fluxes, and depletion corrections are all calculated with OPS-LT. The new OPS-LT options have no effect at these larger distances.

Figure 4.7 shows how outcomes for the two OPS-LT versions compare with measured NO_2 concentrations in the Affligem campaign. The use of the new options (magenta points) results in a much better agreement with measured NO_2 concentrations than the previous version (turquoise points). Relatedly, outcomes for the five performance indicators (Table 4.9 and Figure 4.8) are much closer to their ideal values. The new results for FB, NMSE, MG and VG also fulfil the acceptance criteria from Hanna & Chang.

Figure 4.7 Calculated period-average concentrations for OPS-LT, either using both new options (magenta line) or not (turquoise line). Grey dashed line: period-average observations, grey dotted line: assumed period-average background concentration.

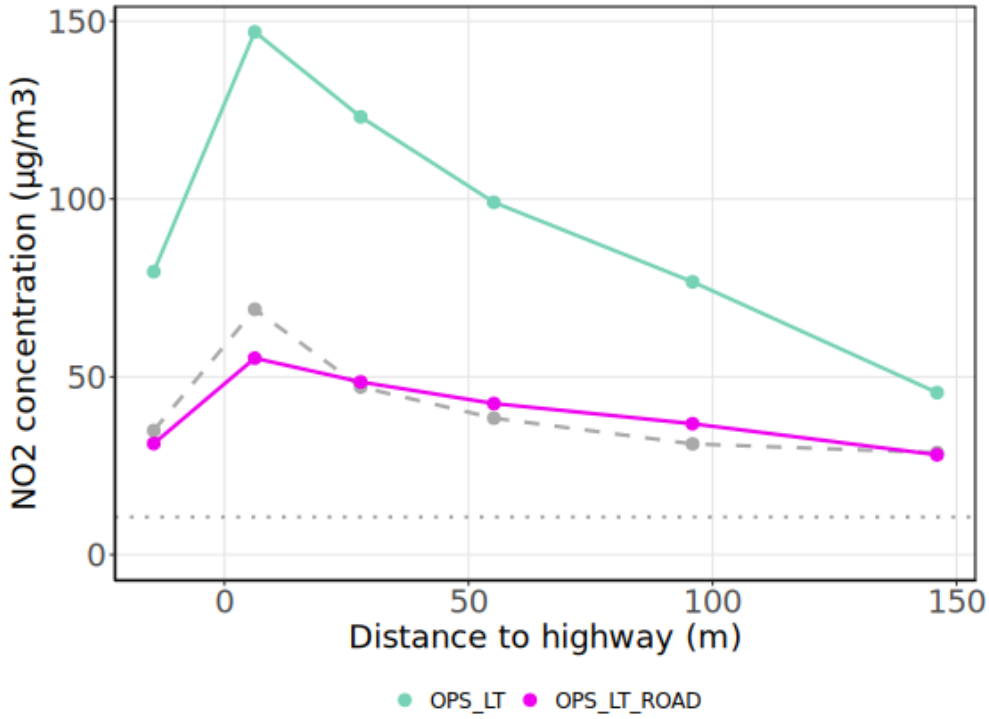
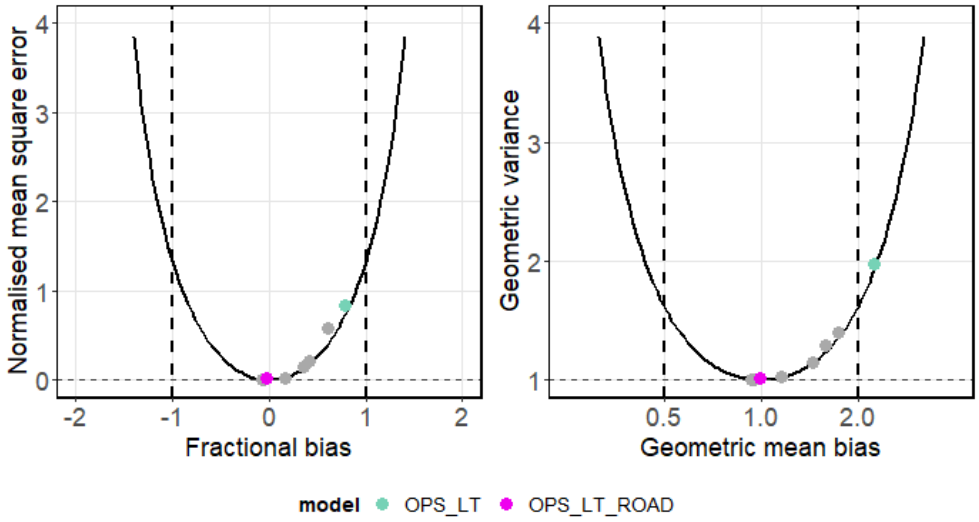


Table 4.10 Model performance for period-average concentrations during the Affligem campaign: update of results for OPS-LT

Model	FB	NMSE	MG	VG	FAC2
Old version	0.78	0.84	2.24	1.98	1/6
New version	-0.03	0.03	0.99	1.02	6/6

Figure 4.8 Deviations and biases of full period-average concentrations for the Affligem campaign for OPS_LT and OPS_LT_ROAD. Left-hand pane: NMSE and FB, right-hand pane: MG and VG. Magenta points: OPS-LT results using both new options, turquoise points: previous OPS-LT results, grey points: results for other individual models



5 Balko measurement campaign

5.1 Description of the measurement campaign

NO_x concentrations were measured around a compressor station in Balko, Oklahoma (USA), in order to identify hourly variance in NO_x concentrations for this type of source. The campaign was part of a larger experimental programme by the US Pipeline Research Council International (PRCI). The Balko campaign had a total duration of about 13 months, from 1 December 2015 until 31 December 2016. The measured data was later used to assess the performance of a new chemical conversion scheme (GRSM) in AERMOD.

The measurement campaign for the Balko compressor station is described in Panek et al. [20]. The use of the data for validating the GRSM code in AERMOD is described in [24] and [25]. Input data (and output data) of AERMET and AERMOD runs for Balko can be downloaded from the EPA website [17]. The files in this download were used to generate the input for the model runs in the current benchmark study.

Figure 5.1 shows an aerial site view. The site had four NO_x emissions sources (red dots) and four measurement locations (blue dots).

Figure 5.1 Top view of the release locations (red circles) and concentration monitors (blue circles) at the Balko compressor station (image copied from [25]).

5.1.1 *Sources and emissions*

According to [20], four sources of NO_x emissions were present at the site: two operational natural-gas fired combustion engines: C9 (model Clark TCV-12) and C10 (model Cooper Bessemer 12W330), an emergency power generator ('EGEN'), and a boiler.

Figure 5.2 Sideview of the compressor units at the Balko compressor station (image copied from [20]). The boiler and the emergency generator were located farther to the left.

These emission sources operated intermittently, depending on natural gas demand and operational requirements. Hourly emissions for each source were estimated. The estimated source strength, release temperature, and release velocity for each hour and source are available in the EPA download file.[28] Table 5.1 provides a summary of this data. The combustion engine C9 had the largest average emission over the whole period. The source term for NO_x in g/s is assumed to be in NO_2-equivalents.

Table 5.1 Summary of emissions (from the raw data in the EPA download [17])

	C9	**C10**	**EGEN**	**Boiler**
Percentage of time operational	20%	19%	0.9%	54%
Average source strength (total period, g/s)	2.5	0.34	0.0026	0.033
Source strength when operational (g/s)	13±3	1.7±0.3	0.3±0.3	0.06±0.02
Emission temperature when operational (K)	589	550	811	670
Emission velocity when operational (m/s)	17 ± 4	20	13	0.001
NO_2/NO_x emissions ratio	0.08	0.3	0.1	0.1

[28] File: /aermod/Inputs/HOUREMIS/Emission_File_C9_velocity.hrl

Table 5.2 Locations of the emission sources (from the raw data in the EPA download [17])

Source	Description	X (m)	Y (m)	Height (m)
C9	Clark TCV-12	331455	4047151	10.50
C10	Cooper Bessemer 12W330	331447	4047182	20.7
EGEN	Emergency generator	331431	4047231	8.4
Boiler		331439	4047235	6.7

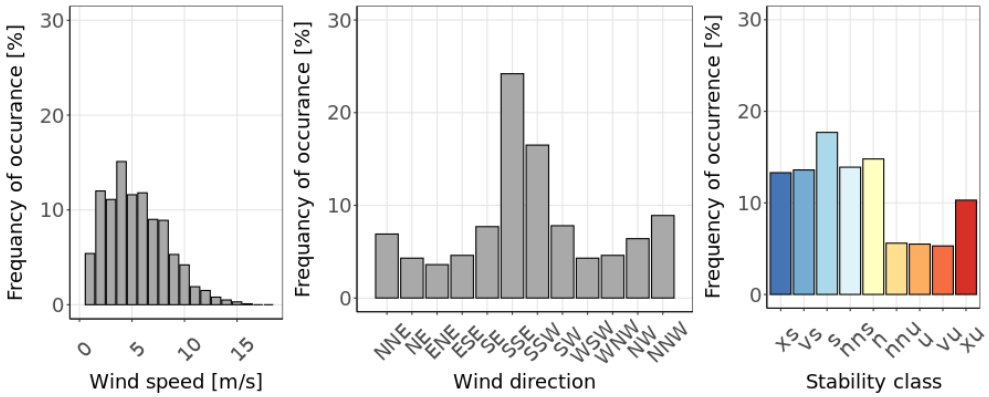
5.1.2 *Meteorological parameters*

The EPA dataset [17] contains raw observational data for meteorological parameters from three separate sources.

- Data from a meteorology tower located at the site. This file contains air temperature at 2 m and 10 m height, wind direction and wind speed at 10 m and 30 m height, hourly variance of wind speed at 10 m and 30 m height, relative humidity, and solar insolation.
- Data from the Automated Surface Observing System (ASOS) at Guymon Oklahoma. This file contains a large number of meteorological parameters, including dewpoint temperature, visibility, and precipitation.
- A file with observations from air soundings from Dodge City, Kansas (USA).

Data from the various data sources was combined and aggregated to hourly values as part of the AERMET pre-processing calculations. Missing values were replaced by estimates using AERMET conventions. Over the campaign period, the average wind speed was 5.4 m/s, with dominant winds from the SSE direction. Weather conditions during the campaign were quite evenly distributed across several stability classes, with slightly more hours with stable and neutral conditions compared with unstable conditions (Figure 5.3).

Figure 5.3 Summary of meteorological variables from the Balko compression station campaign. The data on wind speed and wind direction comes *from a local measurement station. Data on the stability classes comes from the output from OPS-ST.*



5.1.3 Terrain and environment

5.1.3.1 On-site buildings

Several buildings were located on the site of the compressor station. They were spread around the site (see Figure 5.1). The coordinates of the corners of the buildings and average building heights can be found in the EPA dataset[29] and are reproduced in Table 5.3.

Table 5.3 Characteristics of buildings at the Balko site

Name	Building centre x-coordinate (m)	Building centre y-coordinate (m)	Building height (m)
C2	331468	4047142	11.3
C3	331466	4047182	13.5
OFFICE	331387	4047305	5.3
UTILITY	331439	4047236	5.8
WH	331372	4047158	5.4
COOLER1	331431	4047184	4.5
COOLER2	331445	4047142	4.5

5.1.3.2 Surrounding environment

The case documentation does not provide details about the nature of the surroundings of the compressor station (surface roughness and vegetation).

The AERMET control files contain input values for the albedo, the Bowen ratio, and the surface roughness of the study area. The same values are used for each distance. These vary with the seasons. The parameters for Albedo and Bowen ratio in the AERMET control file correspond to default values for water in the AERMET user guide. The value for surface roughness corresponds to the default for cultivated land in the AERMET user guide. The year-average surface roughness from the control file is 6 cm.

Using online maps, we assume the surroundings to be mostly cultivated land (crops).

5.1.4 Concentration measurements

NO, NO_2, and ozone (O_3) were measured at four locations around the compressor station (see Figure 5.1). Hourly measurements are available for the period from 1 December 2015 until 31 December 2016 (13 months in total). Table 5.4 provides the coordinates of the measurement locations. Table 5.5 gives a summary of the numbers of hours with invalid (missing) outcomes and the number of measurements below the assumed background concentration (see section 5.1.5).

Table 5.4 Monitoring locations

Description	X (m)	Y (m)	Height above surface (m)	Distance to C9 (m)
Field	331458	4047575	2.5	425
North fence	331453	4047291	2.5	140
East fence	331548	4047189	2.5	100
Tower	331284	4047074	2.5	190

[29] See the file Bpipprm/Balko_Bpipprm.inp in the EPA download [17].

Table 5.5 Summary of invalid and zero measurements in the Balko campaign

Station	Measurement	Hours	invalid	below bg
East fence	NO_2	9559	6.3%	0.1%
East fence	NO_x	9559	6.3%	1%
North fence	NO_2	9559	4.8%	7.4%
North fence	NO_x	9559	4.8%	21%
Field	NO_2	9559	27%	14%
Field	NO_x	9559	27%	29%
Tower	NO_2	9559	4.9%	0.8%
Tower	NO_x	9559	4.9%	27%

The four monitoring locations were quite close to the emission sources. Twelve extra receptors were used to provide more information on model behaviour at farther distances (see Chapter 7). The additional receptors were located at 200, 500 and 1000 m from source C9 (the strongest source) in towards the north, east, south and west.

Table 5.6 Additional receptors for the Balko study

Label	X (m)	Y (m)	Height above ground (m)
rcp_200m_North	331455	4047351	2.5
rcp_200m_East	331655	4047151	2.5
rcp_200m_South	331455	4046951	2.5
rcp_200m_West	331255	4047151	2.5
rcp_500m_North	331455	4047651	2.5
rcp_500m_East	331955	4047151	2.5
rcp_500m_South	331455	4046651	2.5
rcp_500m_West	330955	4047151	2.5
rcp_1000m_North	331455	4048151	2.5
rcp_1000m_East	332455	4047151	2.5
rcp_1000m_South	331455	4046151	2.5
rcp_1000m_West	330455	4047151	2.5

5.1.5 *Background concentrations*

There were no other monitors apart from the concentration monitors described in the previous subsection. Therefore, the data from these monitors was used to define the background concentrations. CERC used observations from upwind monitors as reference for the background concentrations [25]:
'Meteorological data were recorded at the "Tower" monitoring station, located in the southwest corner of the site. NO_X, NO_2 and O_3 background concentration data were derived from upwind concentrations, with the Field monitor used as background for wind directions 305°-156° and the Tower monitor used as background for wind directions 156°-305°. Background values were low compared to peak concentrations, with background NOX concentrations on average approximately 5 µg/m^3 and over 99.5% of hourly values less than 20 µg/m^3.'

The corresponding hourly estimates are available in the EPA dataset [17].[30] The background concentration for NO_2 is reported in ppb, while NO_x is reported in µg/m^3 (assumed to be NO_2-equivalent mass). Ozone

[30] See folder /aermod/Inputs/Background/

values are reported in ppm. The average background concentrations and their standard deviation are provided in Table 5.7.

Table 5.7 Summary of background concentrations as used for AERMOD

Component	Average background concentration	Standard deviation
NO_2	2.3 (ppb)	1.7 (ppb)
NO_x	4.7 ($\mu g/m^3$)	3.5 ($\mu g/m^3$)
O_3	0.033 (ppm)	0.012 (ppm)

A significant number of hourly measured concentrations is lower than the calculated hourly background concentration (see Table 5.5). The underlying reasons were not investigated but could be related to measurement inaccuracy and concentrations often being close to the background (if the wind is not blowing towards the measurement location).

5.2 Model validation results

During the Balko campaign, hourly NO_2 and NO_x concentrations were measured at four locations near the compressor station, during a period of 13 months. Five models: ADMS, AERMOD, OML-Multi, OPS-ST and STACKS-D, provided both NO_2 and NO_x concentrations for the Balko campaign. OPS-LT could only provide NO_x concentrations (NO_2 concentrations are not exported to output files). IFDM only provided NO_2 concentrations, due to miscommunication between RIVM and VITO.

The current section discusses the performance of models regarding average concentrations over the full period of 13 months (see section 2.3). Hours with invalid (missing) observations were ignored when calculating total period average concentrations. Model performance results for hourly concentration measurements are provided in Appendix 4, for the models that provided hourly output.

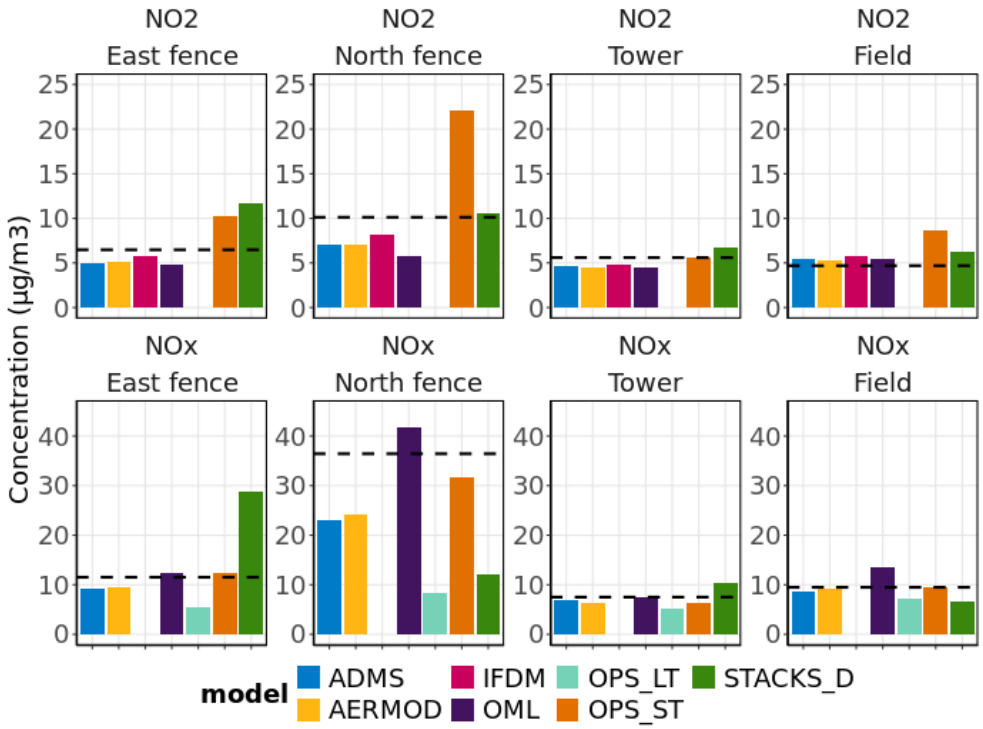
5.2.1 Model performance for period-average concentrations

The model performance is analysed using five performance indicators (see section 2.2). These indicators do not account for measurement uncertainty, and therefore, they only quantify model performance if the measurements are sufficiently accurate.

The measured concentrations include background from other sources. The height of this background was taken from an input file for AERMOD runs that was downloaded from the EPA website [17]. The accuracy of the assumed background is unknown, but affects model outcomes for source plus background and model performance outcomes.

Figure 5.4 shows the calculated period-average concentrations (NO_2 and NO_x) against measured values. The East Fence monitor is located at roughly 100 m distance from the main source (C9), the North Fence monitor at 140 m distance, the Tower monitor at 190 m distance, and the Field monitor at 425 m distance.

Figure 5.4 Calculated versus measured period-average NO_2 (top) and NO_x (bottom) concentrations for the Balko compressor station (dashed lines: measured concentrations, coloured bars: calculated concentrations). Note that vertical axis differs between NO_x and NO_2.



The following information can be deduced from Figure 5.4:

- NO_x measurements show more variation between locations than NO_2 measurements. NO_x concentrations reduce with distance due to mixing. NO_2 concentrations partly reduce with distance due to mixing but partly increase with distance due to conversion of NO into NO_2. As a result, the reduction with distance is less significant for NO_2 than for NO_x.
- Most model outcomes are reasonably close to measurements. The largest discrepancies between model outcomes and measured outcomes occur for the NO_x measurements at North Fence, which is the location closest to the dominant source and surrounding buildings.
- Regarding NO_x, the variation of measured concentrations between locations is captured quite well by OPS-ST, OML-Multi, AERMOD, and ADMS. STACKS-D calculates concentrations that are too low at North Fence and too high at East Fence. OPS-LT calculates too low concentrations for both locations (North Fence and East Fence).
- Regarding NO_2, ADMS, AERMOD, IFDM, and OML outcomes are close to measured concentrations; in more detail, they are slightly lower. OPS-ST calculates higher concentrations for North Fence than measured, and, to a lesser extent, also for East Fence and Field. STACKS-D calculates higher concentrations for East Fence, but is in good agreement with measurements for the three other locations.

The outcomes for the key performance indicators are reported in Table 5.8 (NO_2) and Table 5.9 (NO_x), and are visualised in Figure 4.6 (NO_2 and NO_x). Further details about receptor contributions are provided in Appendix 4.

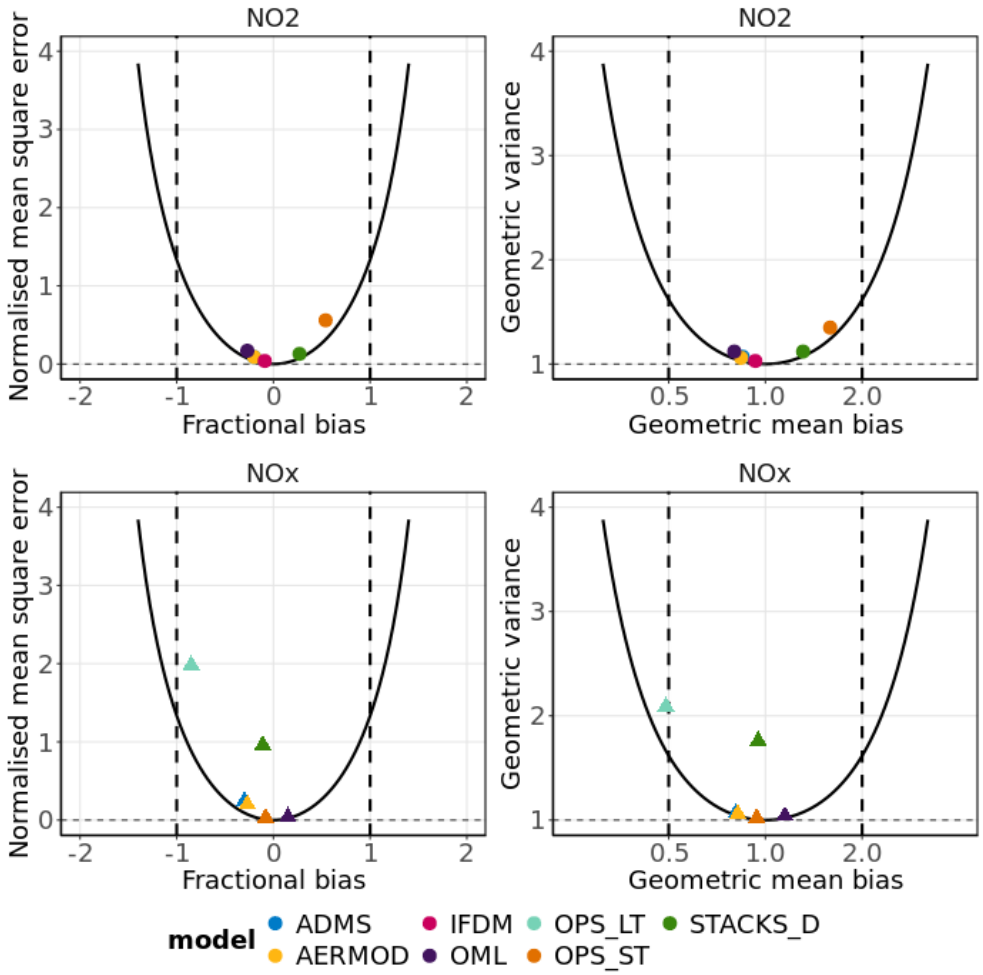
Table 5.8 Model performance for total period average NO_2 concentrations during the Balko campaign

Model	**FB**	**NMSE**	**MG**	**VG**	**FAC2**
ADMS	-0.20	0.09	0.85	1.1	4/4
AERMOD	-0.20	0.09	0.84	1.1	4/4
IFDM	-0.09	0.04	0.93	1.0	4/4
OML-Multi	-0.27	0.17	0.80	1.1	4/4
OPS-ST	0.54	0.56	1.60	1.4	3/4
STACKS-D	0.27	0.13	1.30	1.1	4/4

Table 5.9 Model performance for total period average NO_x concentrations during the Balko campaign

Model	**FB**	**NMSE**	**MG**	**VG**	**FAC2**
ADMS	-0.30	0.24	0.81	1.1	4/4
AERMOD	-0.27	0.20	0.82	1.1	4/4
OML-Multi	0.15	0.04	1.15	1.0	4/4
OPS-LT	-0.85	2.00	0.49	2.1	2/4
OPS-ST	-0.08	0.03	0.94	1.0	4/4
STACKS-D	-0.11	0.96	0.95	1.8	2/4

Figure 5.5 Deviations and biases of full period-average NO_2 (top) and NO_x (bottom) concentrations for the Balko campaign. Left-hand pane: FB and NMSE, right-hand pane: MG and VG.



The following can be concluded from these figures and tables:

- OPS-ST and OML-Multi have the best scores for NO_x measurements, while OPS-LT has the poorest scores for these measurements. Regarding NO_2 measurements, IFDM has the best model performance results, and OPS-ST has the worst.
- The fraction of calculated outcomes within a factor of two of the measured outcomes is smallest for OPS-LT (two out of four). STACKS-D has six out of eight outcomes within a factor of two, and OPS-ST has seven. All other models have all outcomes within a factor of two measured concentrations (note: IFDM only provided outcomes for NO_2).
- Outcomes for North Fence have a great impact on calculated NMSE: the contribution from this location to total NMSE is between 61% and 82% for all models except STACKS-D (see Appendix 4). North Fence is also most important for VG outcomes, but not as prominent as for NMSE.

The calculated performance indicators of Table 5.8 (NO_2 concentrations) and Table 5.9 (NO_x concentrations) are compared with the performance

criteria from Hanna & Chang (2012) for rural terrain [39] in Table 5.10 and Table 5.11. As was noted in section 2.2.1, these criteria are intended as guidance; some failures to meet conditions are tolerated. Regarding NO_2 concentrations, OPS-ST is the only model that fails some tests. Regarding NO_x concentrations, OPS-LT fails the two tests for bias (FB and MG), while STACKS-D and OPS-LT fail the test for Geometric Variance (VG).

Table 5.10 Comparison of calculated performance indicators for NO_2 measurements with criteria from [39]

Test	**Models that fail the test**
|FB| < 0.3	OPS-ST
0.7 ≤ MG ≤ 1.3	OPS-ST
NMSE < 3	
VG < 1.35	OPS-ST just fails (VG=1.35)
FAC2 ≥ 0.5	

Table 5.11 Comparison of calculated performance indicators for NO_x measurements with criteria from [39]

Test	**Models that fail the test (ordered from lesser to larger violations of the criterion)**
|FB| < 0.3	OPS-LT
0.7 ≤ MG ≤ 1.3	OPS-LT
NMSE < 3	
VG < 1.35	STACKS-D, OPS-LT
FAC2 ≥ 0.5	

5.2.2 *Comparing calculated biases for NO_2 and NO_x*

For emissions of nitrogen oxides (NO_x), model outcomes for nitrogen dioxide (NO_2) depend on assumptions for the NO_2/NO_x ratio in the emission and the modelling of dispersion and NO_X chemistry. Biases of model outcomes to measurements should, therefore, be similar for NO_2 and NO_x if the assumed NO_2/NO_x ratio and the modelling of NO_X chemistry are both reasonably accurate. The comparison of biases for NO_2 and NO_x is only useful if the measurements have similar spatial and temporal coverage. This is true for the Balko measurements.

Figure 5.6 shows the calculated Fractional Biases (FB) for NO_2 and NO_x, for the models that provided both outcomes, and Figure 5.7 shows the Geometric Variances (VG) of these outcomes. The two graphs present similar patterns. Calculated biases for ADMS and AERMOD are very consistent between the two components, with only slightly higher biases for NO_x, compared with NO_2. Calculated biases for OML, OPS-ST, and STACKS-D are much less consistent, showing either average underprediction for NO_2 and average overprediction of NO_x (OML), or vice versa (OPS-ST and STACKS_D). The former could indicate that the assumed NO_2/NO_x ratio for the emission is too small, that the conversion of NO into NO_2 is too slow, or both. Similarly, the latter could indicate that the assumed ratio is too large, that the chemical conversion is too fast, or both.

Figure 5.6 Comparison of the Fractional Biases (FB) of the models regarding NO_2 and NO_x concentration measurements for the Balko campaign

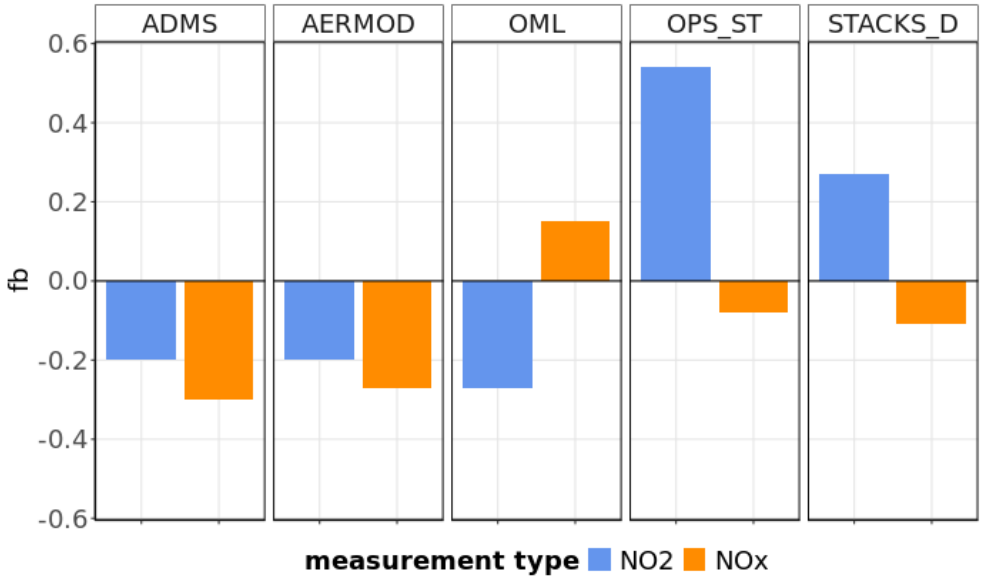
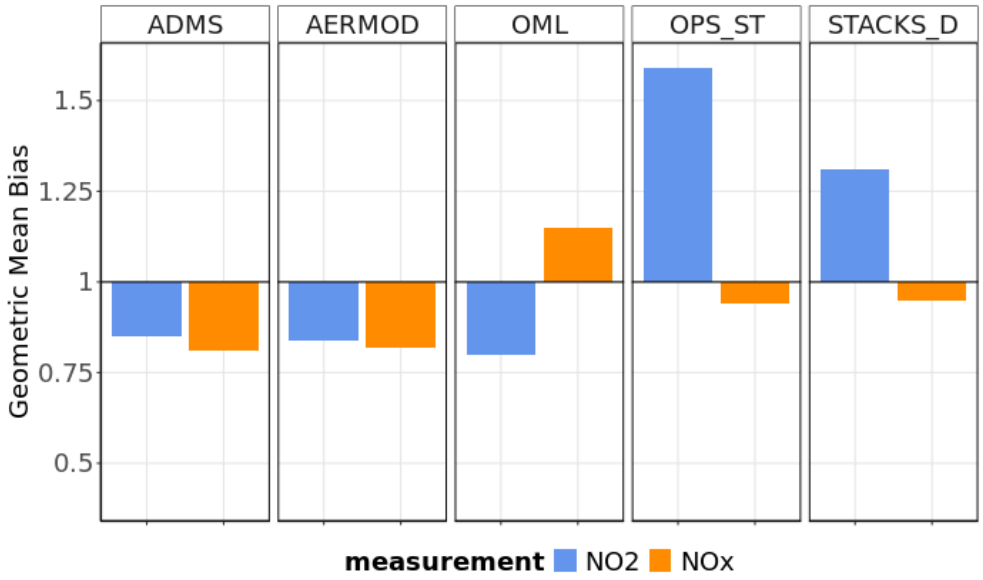


Figure 5.7 Comparison of the Geometric Mean Biases (MG) of the models regarding NO_2 and NO_x concentration measurements for the Balko campaign



6 Model ensemble outcomes

In this chapter, the performance of model ensembles is investigated and compared with the performance of individual models. The same measurements as before are used to calculate and compare performances, while the same performance indicators are also used. Two research questions will be addressed:

- Will the ensemble outcome show better agreement with measurement outcomes than individual model outcomes?
- Can the uncertainty of ensemble outcomes be derived from the variance in individual model outcomes?

Each ensemble consists of a subset of the models included in this study. Multiple ensembles (subsets of models) are investigated for each measurement campaign. The ensemble outcome is derived from the outcomes for the individual models. That ensemble outcome is compared with measurements, similar to the comparison of individual model outcomes with measurements above.

6.1 Definition of ensembles and ensemble outcomes

For each considered campaign (Ringsted, Affligem and Balko), we will define three separate ensembles. This approach allows us to identify how the size of the ensemble affects the performance of the ensemble.

Group 1: an ensemble of three hourly models;
Group 2: the ensemble of all hourly models (ADMS, AERMOD, IFDM, OML-Multi, and OPS-ST) that provided output for a case;
Group 3: the ensemble of all models that provided output.

For the smallest model set, we will use three models that calculate concentrations and deposition for each our (hourly models). A priori, we would expect that such models would perform better than models that use classified meteorology (OPS-LT and SRM2) or only use a selection of hours (STACKS-D). The default choice for group 1 is to select ADMS, IFDM, and OML-Multi.[31] AERMOD replaces OML in the Affligem campaign (because OML had no outcomes for this campaign) and replaces IFDM regarding Balko NO_x concentrations (because IFDM did not provide that output). The models in group 1 were developed by different research organisations and are thus reasonably independent.

In group 2, the remaining hourly models are added, and in group 3 the remaining non-hourly models. OPS-ST and OPS-LT in group 3 have the same origin and cannot be regarded as independent. In practice, OPS-ST and OPS-LT hardly show better agreement between the two than the other models.[32] Regarding deposition, IFDM outcomes are not independent of OPS-LT, because OPS-LT output was used to calculate the IFDM deposition velocity. Despite this dependence, IFDM deposition outcomes for the Ringsted campaign showed better agreement with

[31] Between the four models with hourly varying emissions, AERMOD input and output was more difficult to generate and process than ADMS, IFDM, and OML input and output. Therefore, it was decided to only include AERMOD if needed; in other words if either ADMS, IFDM, or OML output was missing.
[32] This claim was verified by studying covariance parameters in a correlations matrix for each campaign.

AERMOD and OPS-ST outcomes than with OPS-LT outcomes. Thus, the overall effect of the dependency appears to be limited.

For the joint outcome of the models in an ensemble, also referred to as the 'ensemble outcome', we will use the Geometric Mean (GM) of the individual model outcomes (see section 2.4). The Geometric Mean gives equal weight to models that underpredict by a factor of A and models that overpredict by the same factor of A. The Arithmetic Mean (AM) outcome is believed to be less suited, as was substantiated in the model intercomparison report [2] (see, for example, the Summary and Appendix 3 of that report). Ensemble performance relating to the use of the AM as mean ensemble outcome was calculated, and did not differ much from the performance based on the GM. For the purpose of conciseness, results for AM are not included in this report.

For the comparison of the ensemble outcome with measurements, we will use the same performance indicators that were previously used for the individual models (see section 2.2). In the visualisation and discussion, we will focus on the Geometric Mean Bias (MG) and the Geometric Variance (VG). It is believed that these indicators provide a better image of model performance than Fractional Bias (FB) and Normalised Mean Square Error (NMSE), because concentrations and deposition fluxes in this study show a significant reduction with distance from the source. The geometric indicators (VG and MG), then, give a more balanced outcome for the full dataset (see section 2.2.1).

The ensemble outcome is designed to be in the middle of the group of individual model outcomes. Performance indicators for the ensemble will only be substantially better than the indicators of individual models if some models in the ensemble underpredict, while others overpredict. If all individual models show a bias in the same direction, the performance of the ensemble will normally be somewhere in the range of individual models performances.

6.2 Outcomes for Ringsted concentrations

In this subsection, we will discuss the performance of three sets of models (ensembles), regarding concentration and deposition measurements in the Ringsted campaign (see Table 6.1). The ensemble outcome is defined as the Geometric Mean of the underlying individual model outcomes (see section 6.1). The Ringsted measurement campaign itself was described in Chapter 3.

Table 6.1 Selection of ensembles for the Ringsted campaign

Ensemble	**Included models**
Group 1 (n=3)	ADMS, IFDM, and OML-Multi
Group 2 (n=5)	The above + AERMOD and OPS-ST
Group 3 (n=7)	The above + OPS-LT and STACKS-D

The performance outcomes for the three ensembles are reported in Table 6.2 and visualised in Figure 6.1 (MG versus VG only). Together, the table and figure show:

- The calculated performance indicators are quite similar between the three ensembles: the Geometric Mean Bias (MG) ranges

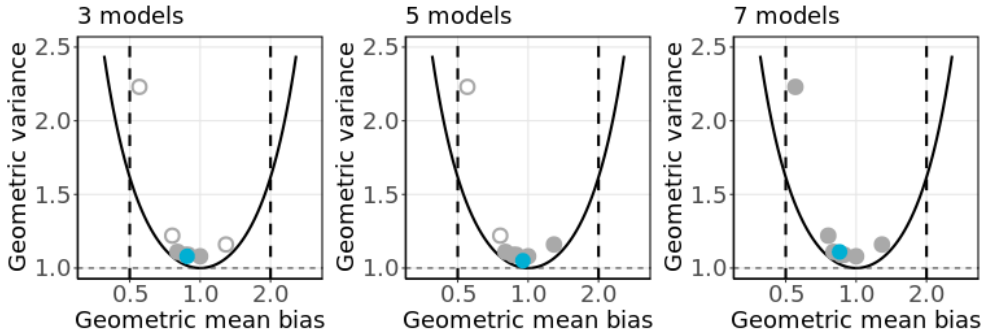
between 0.85 and 0.95, and the Geometric Variance (VG) ranges between 1.05 and 1.11. Between the three ensembles, the set of five models performs slightly better than the two other sets.

- Overall, the ensembles are believed to perform as well as the hourly models, but not better.
- The (minor) differences between the three ensembles can be explained in terms of the outcomes for the underlying individual models. For example, the smallest ensemble uses three models that all have a negative bias (MG<1), and that also has a negative bias itself.[33] The set of five models contains one model (OPS-ST) with positive bias, which here reduces the mean deviation (underpredictions are now balanced by an overprediction) and moves the ensemble mean bias towards neutrality. The largest ensemble contains one model (STACKS-D) that ended up with worse performance statistics than the other models. Adding this model to the ensemble has a negative effect on the ensemble performance statistics.

Table 6.2 Performance of the model ensembles for Ringsted concentration measurements. Outcomes for individual models were added in italic for the purpose of completeness.

Model selection	**FB**	**NMSE**	**MG**	**VG**	**FAC2**
ADMS	*-0.35*	*0.38*	*0.80*	*1.1*	*0.96*
IFDM	*-0.15*	*0.21*	*1.0*	*1.1*	*1.0*
OML-Multi	*-0.29*	*0.34*	*0.86*	*1.1*	*0.96*
3 models	**-0.27**	**0.30**	**0.88**	**1.08**	**1**
AERMOD	*-0.28*	*0.34*	*0.89*	*1.1*	*0.96*
OPS-ST	*0.39*	*0.58*	*1.3*	*1.2*	*0.89*
5 models	**-0.15**	**0.14**	**0.95**	**1.05**	**1**
OPS-LT	*-0.47*	*0.81*	*0.76*	*1.2*	*0.81*
STACKS-D	*-0.85*	*2.8*	*0.55*	*2.2*	*0.74*
7 models	**-0.32**	**0.43**	**0.85**	**1.11**	**0.96**

Figure 6.1 Performance of the model ensembles (blue points), relative to that of individual models included in the ensemble (grey points) or not included in the ensemble (grey open circles), for Ringsted concentration measurements



[33] Ensembles can have a close to neutral bias if overpredictions by some models are balanced by underpredictions from other models. In contrast, if all underlying either underpredict or overpredict, the ensemble outcome is likely (though not necessarily) to have similar positive or negative bias.

6.3 Outcomes for Ringsted deposition

The performance outcomes for the three ensembles relating to the Ringsted deposition measurements are reported in Table 6.3 and visualised in Figure 6.2 (MG versus VG only). Two receptors with negative measured deposition flux were not used for calculating the Geometric Mean Bias (MG) and the Geometric Variance (VG), because these indicators do not allow negative values. Similarly, two outcomes equal to 0 from STACKS-D also had to be excluded from the ensemble of all models.

The following can be concluded from Table 6.3 and Figure 6.2:

- The performance of the ensembles containing five or seven models is slightly better than the performance of the smallest ensemble.
- The three models in the smallest ensemble all have positive bias (MG>1). The score for this ensemble is in between the individual scores of the three models.
- Adding AERMOD and OPS-ST gives better performance statistics because AERMOD has a negative bias, which to some extent compensates the positive biases of the other models.
- Furthermore, adding OPS-LT and STACKS-D improves the ensemble scores slightly more, mostly due to the (positive) influence of OPS-LT.
- Overall, the scores for group 2 and group 3 are considered to be as good as[34] those for OML-Multi and OPS-LT, and better than the results for the other individual models.

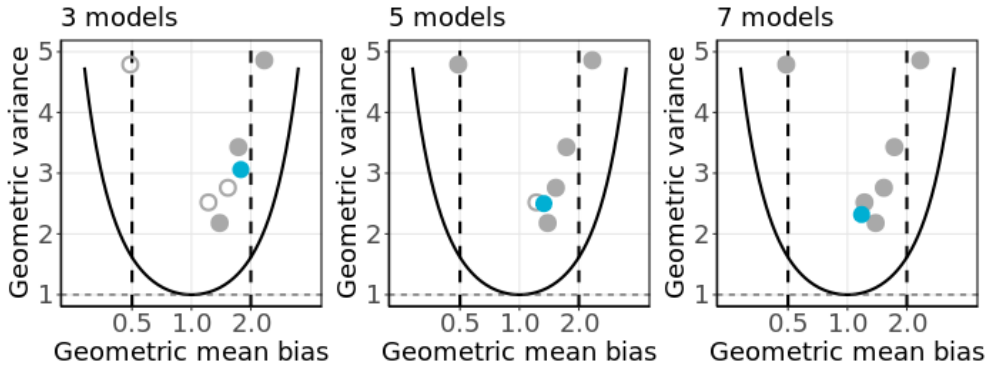
Table 6.3 Performance of the model ensembles for Ringsted deposition measurements. Outcomes for individual models were added in italic for the purpose of completeness.

Model selection	**FB**	**NMSE**	**MG**	**VG**	**FAC2**
ADMS	*0.73*	*1.03*	*2.3*	*4.9*	*0.44*
IFDM	*0.37*	*0.75*	*1.7*	*3.4*	*0.56*
OML-Multi	*0.13*	*0.48*	*1.4*	*2.2*	*0.56*
Group 1 (n=3)	**0.4**	**0.52**	**1.78**	**3.1**	**0.56**
AERMOD	*-0.81*	*3.28*	*0.49*	*4.8*	*0.25*
OPS-ST	*0.24*	*0.59*	*1.5*	*2.8*	*0.50*
Group 2 (n=5)	**0.11**	**0.60**	**1.33**	**2.5**	**0.44**
OPS-LT	*0.01*	*0.87*	*1.2*	*2.5*	*0.38*
STACKS-D	*-0.58*	*3.5*	*0.57 **	*38 **	*0.31*
Group 3 (n=7)	**-0.04**	**0.92**	**1.18**	**2.3**	**0.38**

* Two receptors with outcome 0 had to be excluded when calculating VG and MG for STACKS-D.

[34] As good as: neither significantly better nor significantly worse.

Figure 6.2 Performance of the model ensembles (blue points), relative to that of individual models included in the ensemble (grey points) or not included in the ensemble (grey open circles), for Ringsted deposition measurements



6.4 Outcomes for Affligem concentrations

The measurements in the Affligem campaign were described in Chapter 4. Similar to the evaluation of individual models, we will use full period-average concentrations to calculate ensemble performance.

The sets of ensembles used for Affligem are slightly different from the ones used for Ringsted, because OML-Multi did not provide output for this campaign. Therefore, OML was replaced by AERMOD in the smallest set of models (Table 6.4). OPS-ST is added to the second set of models, which only contains four models (not five). The three annual models are only included in the full set of (seven) models. OPS-LT outcomes relate to the 2024 version of OPS-LT.

Table 6.4 Selection of ensembles for the Affligem campaign

Ensemble	**Included models**
Group 1 (n=3)	ADMS, IFDM, and AERMOD
Group 2 (n=4)	The above + OPS-ST
Group 3 (n=7)	The above + OPS-LT, SRM2, and STACKS-D

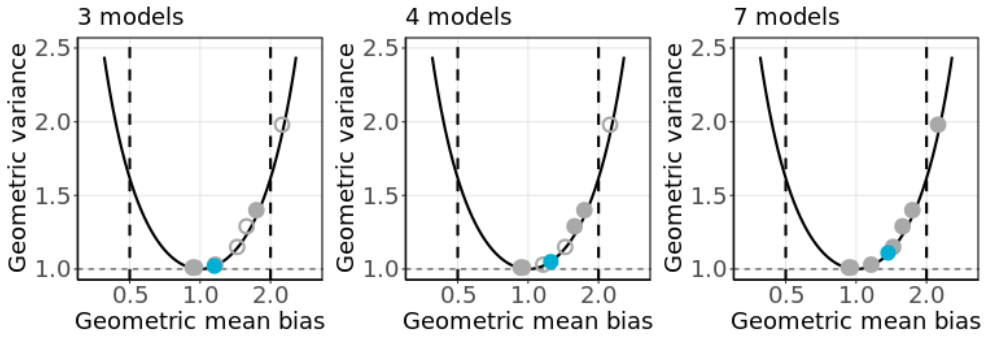
The performance outcomes for the three ensembles are reported in Table 6.5 and visualised in Figure 6.3 (MG versus VG only). Together, the table and figure show:

- The calculated performance indicators for the three ensembles are very similar. Out of the three, the smallest ensemble performs just slightly better than the other two.
- Two out of the three models in the smallest set have a bias (GM) that is very close to 1. Average overpredictions by the third model are, therefore, not countered by underpredictions from the other models. As a result, the score for the smallest ensemble is in between the scores of the three individual models.
- Adding more models to the ensemble makes the overall performance slightly worse. The average positive bias of these extra models now moves the ensemble outcome into the wrong direction.
- Overall, the ensembles are believed to perform better than most models, but not as well as the best hourly models.

Table 6.5 Performance of the model ensembles for Affligem concentration measurements. Outcomes for individual models were added in italic for the purpose of completeness.

Model selection	FB	NMSE	MG	VG	FAC2
ADMS	*-0.06*	*0.00*	*0.93*	*1.01*	*6/6*
AERMOD	*0.60*	*0.59*	*1.74*	*1.40*	*5/6*
IFDM	*-0.07*	*0.01*	*0.95*	*1.01*	*6/6*
Group 1 (n=3)	**0.16**	**0.04**	**1.15**	**1.02**	**6/6**
OPS-ST	*0.42*	*0.22*	*1.58*	*1.29*	*6/6*
Group 2 (n=4)	**0.22**	**0.05**	**1.25**	**1.05**	**6/6**
OPS-LT	*0.78*	*0.84*	*2.24*	*1.98*	*1/6*
SRM2	*0.36*	*0.15*	*1.44*	*1.15*	*6/6*
STACKS-D	*0.16*	*0.03*	*1.16*	*1.03*	*6/6*
Group 2 (n=7)	**0.32**	**0.11**	**1.37**	**1.11**	**6/6**

Figure 6.3 Performance of the model ensembles (blue points), relative to that of individual models included in the ensemble (grey points) or not included in the ensemble (grey open circles), for Affligem concentration measurements



6.5 Outcomes for Balko concentrations

The measurements in the Balko campaign were described in Chapter 5. Similar to the evaluation of individual models, we will use full-period average concentrations to calculate ensemble performance.

6.5.1 *Balko NO_x concentrations*

The sets of ensembles used for evaluating NO_x concentrations in the Balko campaign are reported in Table 6.6. IFDM was replaced by AERMOD in the smallest set because IFDM did not provide output for NO_x concentrations. In absence of IFDM results, the second group has four models (not five) and the third group has six models (not seven).

Table 6.6 Selection of ensembles for Balko NO_x concentrations

Ensemble	Included models
Group 1 (n=3)	ADMS, AERMOD, and OML-Multi
Group 2 (n=4)	The above + OPS-ST
Group 3 (n=6)	The above + OPS-LT and STACKS-D

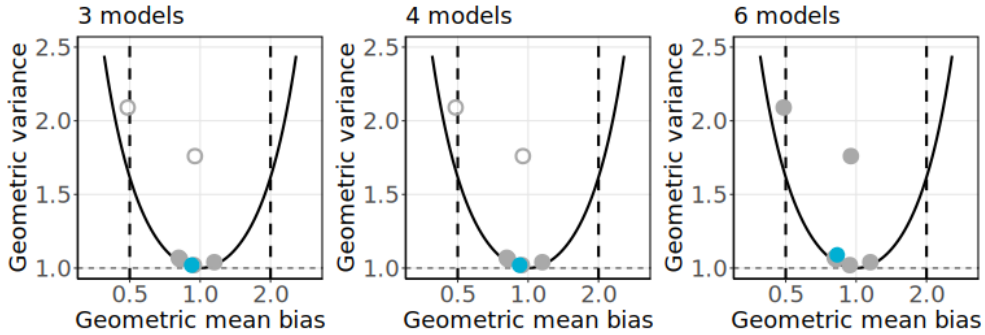
The performance outcomes for the three ensembles are reported in Table 6.7 and visualised in Figure 6.4 (MG versus VG only). Together, the table and figure show:

- The calculated performance indicators for the three ensembles are (again) very similar. The two smallest ensembles perform slightly better than the largest ensemble (all six models).
- The outcomes for the four hourly models are all close together and are all reasonably close to the perfect outcome (MG=VG=1). This explains why the group of three models and the group of four models perform almost equally well. Adding OPS-LT and STACKS-D to the ensemble, however, has a negative influence on performance outcomes, as these two models have larger VG.
- Overall, the ensemble scores are better than those of most individual models, provided the two annual models are not included in the ensemble.

Table 6.7 Performance of the model ensembles for Balko NO_x concentration measurements. Outcomes for individual models were added in italic for the purpose of completeness.

Model selection	FB	NMSE	MG	VG	FAC2
ADMS	*-0.30*	*0.24*	*0.81*	*1.1*	*4/4*
AERMOD	*-0.27*	*0.20*	*0.82*	*1.1*	*4/4*
OML-Multi	*0.15*	*0.04*	*1.20*	*1.0*	*4/4*
Group 1 (n=3)	**-0.15**	**0.07**	**0.92**	**1.02**	**4/4**
OPS-ST	*-0.08*	*0.03*	*0.94*	*1.0*	*4/4*
Group 2 (n=4)	**-0.13**	**0.06**	**0.92**	**1.02**	**4/4**
OPS-LT	*-0.85*	*2.00*	*0.49*	*2.1*	*2/4*
STACKS-D	*-0.11*	*0.96*	*0.95*	*1.8*	*2/4*
Group 3 (n=6)	**-0.31**	**0.33**	**0.83**	**1.09**	**4/4**

Figure 6.4 Performance of the model ensembles (blue points), relative to that of individual models included in the ensemble (grey points) or not included in the ensemble (grey open circles), for Balko NO_x concentration measurements



6.5.2 *Balko NO_2 concentrations*

The sets of ensembles used for evaluating NO_2 concentrations in the Balko campaign are reported in Table 6.8. OPS-LT did not provide outcomes for these NO_2 measurements. Therefore, the third ensemble only contains six models.

Table 6.8 Selection of ensembles for Balko NO_2 concentrations

Ensemble	**Included models**
Group 1 (n=3)	ADMS, IFDM, and OML-Multi
Group 2 (n=5)	The above + AERMOD and OPS-ST
Group 3 (n=6)	The above + STACKS-D

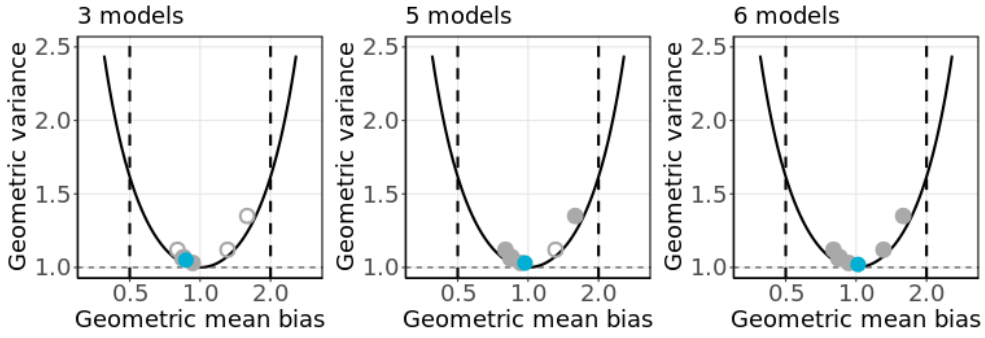
The performance outcomes for the three ensembles are reported in Table 6.9 and visualised in Figure 6.5 (MG versus VG only). Together, the table and figure show:

- The calculated performance indicators for the three ensembles are reasonably close. Out of the three ensembles, the largest ensemble (six models) has the best score, and the smallest ensemble scores worst.
- The smallest ensemble only contains models that have a (small) negative bias. The two other ensembles also contain models with positive biases. The resulting bias is, then, closer to neutrality, and the resulting mean deviation closer to zero (see also footnote 33).
- Overall, the two largest ensembles get better scores than the individual models. The smallest ensemble does not score as well, and is average compared with the scores of the individual models.

Table 6.9 Performance of the model ensembles for Balko NO_2 concentration measurements. Outcomes for individual models were added in italic for the purpose of completeness.

Model selection	**FB**	**NMSE**	**MG**	**VG**	**FAC2**
ADMS	*-0.20*	*0.09*	*0.85*	*1.07*	*4/4*
IFDM	*-0.09*	*0.04*	*0.93*	*1.03*	*4/4*
OML-Multi	*-0.27*	*0.17*	*0.80*	*1.12*	*4/4*
Group 1 (n=3)	**-0.19**	**0.09**	**0.86**	**1.07**	**4/4**
AERMOD	*-0.20*	*0.09*	*0.84*	*1.06*	*4/4*
OPS-ST	*0.54*	*0.56*	*1.59*	*1.35*	*3/4*
Group 2 (n=5)	**-0.05**	**0.03**	**0.97**	**1.03**	**4/4**
STACKS-D	*0.27*	*0.13*	*1.31*	*1.12*	*4/4*
Group 3 (n=6)	**0.00**	**0.02**	**1.02**	**1.02**	**4/4**

Figure 6.5 Performance of the model ensembles (blue points), relative to that of individual models included in the ensemble (grey points) or not included in the ensemble (grey open circles), for Balko NO_2 concentration measurements



6.6 Uncertainty in ensemble outcomes

Benefits of ensemble modelling were expected to be two-fold:
(i) ensemble outcomes were expected to be more accurate than individual model outcomes and (ii) ensemble modelling was expected to provide relevant information on outcome uncertainty (the uncertainty in the ensemble outcome). The latter topic is addressed in this subsection.

Model uncertainty is a measure for the (limited) accuracy of model outcomes. Model uncertainty can be derived from measurements if those measurements are sufficiently accurate. Similarly, the uncertainty in the ensemble outcome of models can be derived by comparing ensemble outcomes with observations. As single model outcomes (and also single ensemble outcomes) can be either close to or far away from measurements by sheer coincidence, an accurate estimate of model uncertainty requires a significant number of measurements.

Model spread refers to the differences between outcomes for individual models in an ensemble. The assumption is that this spread is an adequate indicator for the uncertainty in the ensemble outcomes.

Both model spread and deviations from measurements can be expressed in different ways. In order to limit the number of analyses, we will do two tests.

1. The individual model outcomes are assumed to represent a normal (Gaussian) probability density distribution for the parameter of concern, which is defined by the Arithmetic Mean (AM) and the Standard Deviation (SD) of the individual model outcomes. If the measurements are sufficiently accurate and large in number, 68% of the measured values should range between AM-SD and AM+SD, and 95% should range between AM-2·SD and AM+2·SD (see section 2.4).
2. The individual model outcomes are assumed to represent a lognormal probability density distribution for the parameter of concern, which is defined by the Geometric Mean (GM) and the Geometric Standard Deviation (GSD) of the individual model outcomes. If the measurements are sufficiently accurate and large in number, 68% of the measurements should have an outcome between GM/GSD and GM·GSD, and 95% should have an outcome between GM/GSD^2 and $GM \cdot GSD^2$ (see section 2.4).

Figure 6.6 shows how individual model outcomes in this study were distributed. For this analysis, the model outcomes for the five additional receptors in the Affligem study (Table 4.6) and for the eight additional receptors in the Balko study (Table 5.6) were also included. The left-hand pane shows to what extent the distribution of outcomes resembles the normal distribution, while the right-hand pane shows the resemblance with the lognormal distribution. For the normal comparison, deviations of individual model outcomes relative to their Arithmetic Mean value, were normalised by their Standard Deviation. For the lognormal comparison, all concentrations were transposed to their logarithmic value: the distribution of concentrations is lognormal if the distribution of logarithmic concentrations is normal. Overall, the

distribution of individual model outcomes resembles the lognormal distribution (right) better than the normal distribution (left). [35]

Figure 6.6 Distribution of individual model outcomes in the current study. Left-hand pane: distribution of (outcome-AM)/SD, right-hand pane: distribution of $log_{GSD}((outcome/GM))$.

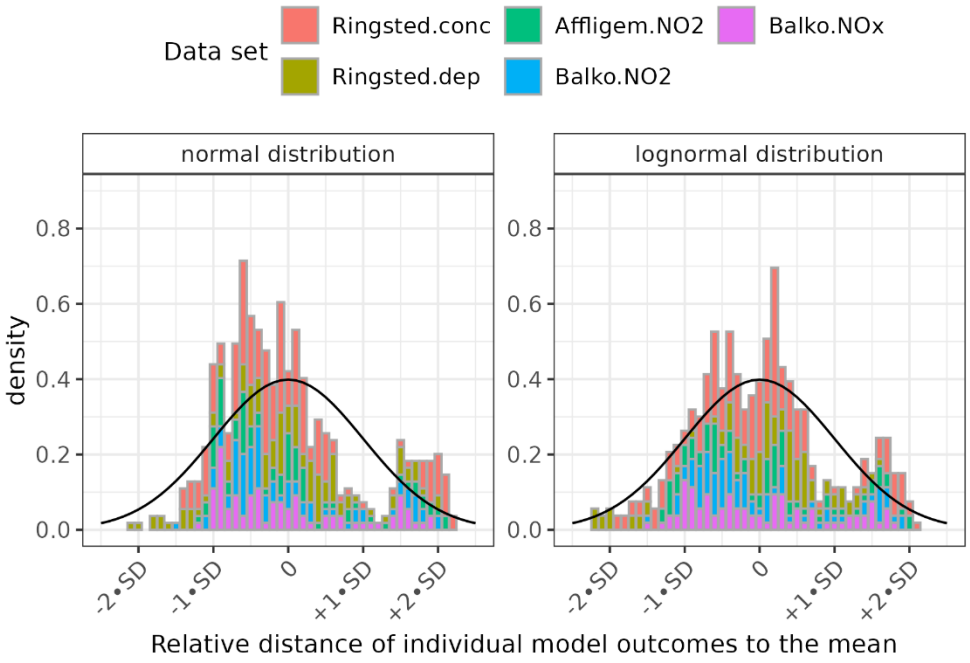
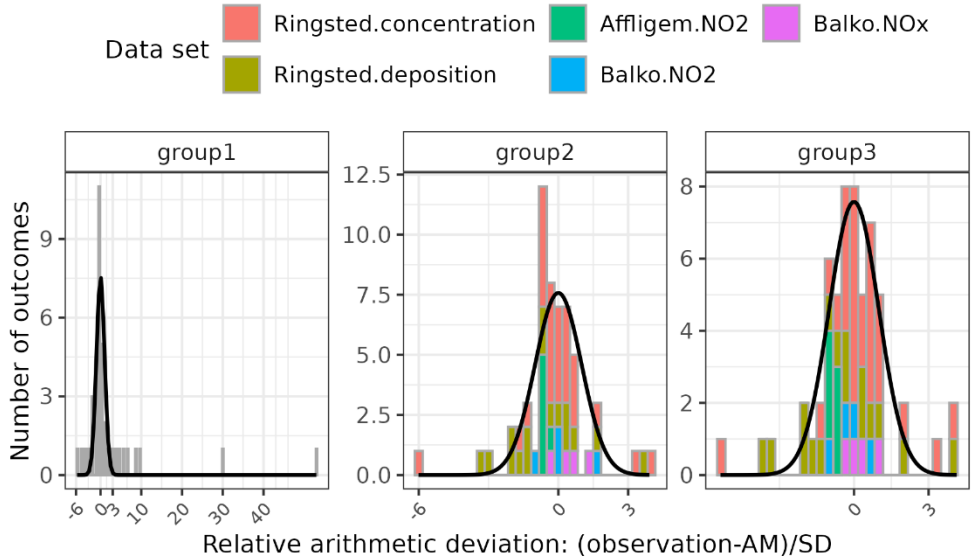


Figure 6.7 shows how measured values are distributed across the Arithmetic Mean ensemble outcome (test 1). Differences between measured values and AM are normalised by the Standard Deviation: $Relative\ arithmetic\ deviation\ = \frac{observation-AM}{SD}$. Therefore, it shows how many SDs the measured value is away from the AM. The plot distinguishes between the three separate ensembles (small, medium and large), but aggregates over the different measurements investigated in this study. The normal distribution itself is indicated by a solid line and can be used to see how well the data fits the normal distribution.

[35] This was tested with the Cramér-von Mises test for goodness of fit. Q^2 was smallest for the lognormal distribution (0.54 versus 2.2) and the p-value largest for the lognormal distribution (0.03 versus 0.000006). See https://doi.org/10.32614/CRAN.package.goftest for more information about this test.

Figure 6.7 Histogram of relative arithmetic deviations: (observation-AM)/SD

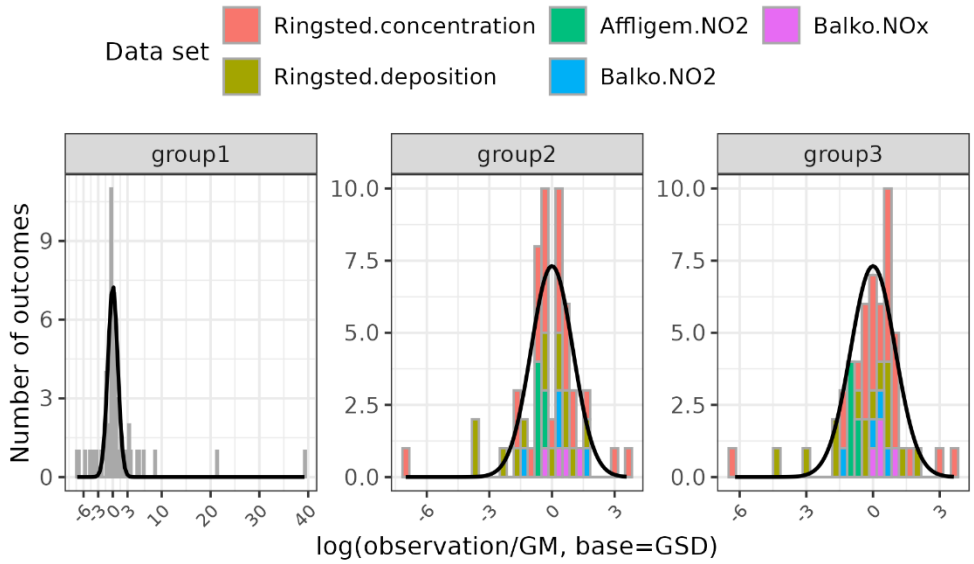


When looking across the different measurements, outcomes for the two larger ensembles (groups 2 and 3) appear to be in reasonable accordance with the normal distribution. As such, the normal distribution around AM with standard deviation SD appears to be a reasonably good approximation of the uncertainty in the ensemble outcome AM. The same conclusions do not apply to the smallest ensemble (group 1), which has too many deviations smaller than AM-2·SD or larger than AM+2·SD. Outliers for group 1 include data pairs for Ringsted concentration measurements, Ringsted deposition measurements and Balko NO_2 measurements.

In more detail, the Balko data for groups 2 and 3 shows the best agreement with the normal distribution. Group 2 and 3 outcomes for Affligem appear to be slightly closer to zero, relative to what should be expected for a normal distribution, while Ringsted concentration and deposition measurements are slightly farther away relative to what should be expected. The number of data points is, however, small and a hypothesis that uncertainty is either smaller (Affligem) or larger (Ringsted) than SD would probably not pass any sound statistical test.

Figure 6.8 shows to what extent differences between measured values and the Geometric Mean (GM) ensemble outcome fit the lognormal distribution. For this 'lognormal evaluation', relative geometric deviation is defined as the logarithm of the ratio between the observed and predicted (GM) outcome, and using GSD as a base for the logarithm: $Relative\ geometric\ deviation\ = log_{GSD}\left(\frac{observation}{GM}\right)$. A value of -2 is the equivalent of $observation = GM \cdot GSD^{-2}$. Similar to Figure 6.7, the plot distinguishes between three separate ensembles and aggregates over the different measurements investigated in this study. The normal distribution is indicated by a solid line and can be used to see whether the distribution of $log\left(\frac{observation}{GM}\right)$ is normal. The latter would imply that the distribution of $\frac{observation}{GM}$ itself is lognormal.

Figure 6.8 Histogram of relative geometric deviations (log(obs/GM, base=GSD)



Overall, the resemblance between Figure 6.8 and Figure 6.7 is considerable. If it was formerly stated that uncertainty can, to a reasonable extent, be expressed in terms of a normal distribution with standard deviation SD, now, it turns out that it can also be expressed in terms of a lognormal distribution with GSD as a (geometric) standard deviation. The data do not favour one of the two options, both appear to be equally valid.

Although the data cannot identify whether uncertainty is better represented by a normal distribution and SD or by a lognormal distribution and GSD, the latter is preferred from a conceptual point of view. Because concentrations are bounded by 0, the asymmetric lognormal density function is expected to represent reality better than the symmetric normal density function.

6.7 Summary

6.7.1 Ensemble outcomes for concentration

The ensemble performances for the four sets of concentration measurements are summarised in Figure 6.4. In general, differences between the three ensembles are small. In more detail, the two ensembles of hourly models (groups 1 and 2) perform slightly better than the ensemble of all models (group 3). Only for Balko NO_2 concentrations, group 3 had the better scores, although group 2 comes very close.

Between the two ensembles of hourly models (groups 1 and 2), group 1 achieves the better outcomes for Affligem NO_2 concentrations, while group 2 achieves the better outcomes for Ringsted NH_3 concentrations and for Balko NO_2 and NO_x concentrations. For the latter three sets of measurements, group 1 only contained models with the same sign for bias, while group 2 also had at least one model with different bias. The slightly worse results for group 1 could, then, be a consequence of this smallest group not being diverse enough.

When looking across all four sets of concentration measurements, the performance results for the ensemble of all hourly models (group 2) were similar to those of IFDM, and better than the results for the other individual models. Outcomes for the smallest ensemble (group 1) were, on average, better than all individual models except IFDM, and the performance of group 3 was similar to that of ADMS and OML.

Figure 6.9 Summary of the performance of ensembles against the performances of individual models, for the four sets of concentration measurements

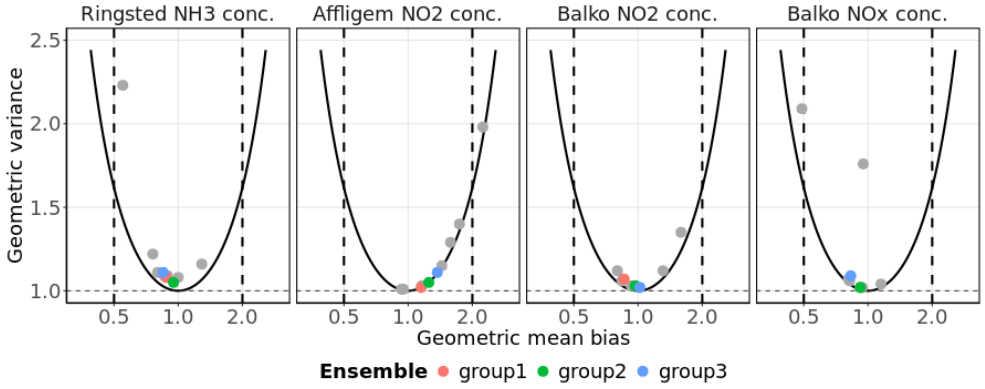


Table 6.10 Comparison of calculated performance indicators with criteria from [33]: outcomes that fail the test

Test	Three hourly models	All hourly models	All models
|FB| < 0.3			Test fails for Ringsted NH_3, Affligem NO_2, and Balko NO_x
0.7 ≤ MG ≤ 1.3			Test fails for Affligem NO_2 concentrations
NMSE < 3			
VG < 1.35			
FAC2 ≥ 0.5			

6.7.2 *Ensemble outcomes for deposition*

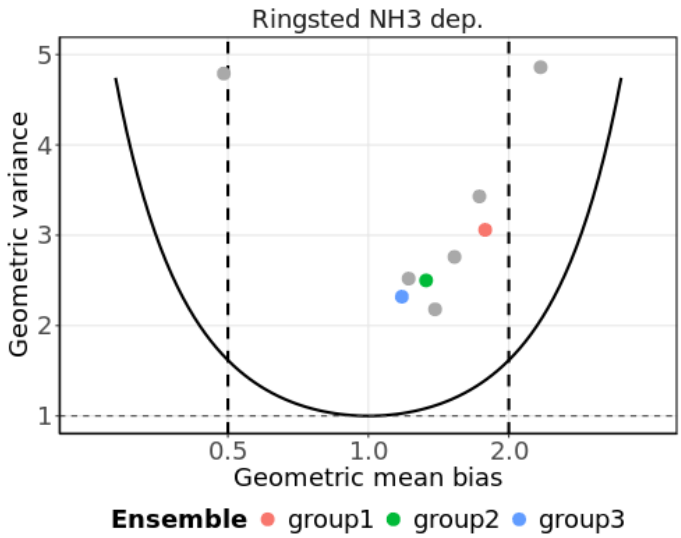
Regarding the deposition measurements in the Ringsted campaign, model performance was not as good as observed for the concentration measurements in Ringsted, Affligem, and Balko. This observation applies to both individual models and ensembles. The worse results in terms of performance statistics are expected to relate to three causes:

- These deposition measurements appear not to have been as accurate as the four sets of concentration measurements.
- The deposition measurements were, on average, carried out closer to the source than the concentration measurements.
- Deposition modelling is inherently more complex because it is a mix of concentration modelling and deposition flux calculations.

Among the ensembles, the ensembles of all hourly models (group 2) and all models (group 3) had better outcomes than the group of three hourly models (group 1). The results for groups 2 and 3 are about as good as

those for OML-Multi and OPS-LT (the best performing individual models for this set of measurements), and perhaps more robust.

Figure 6.10 Summary of the performance of ensembles against the performances of individual models, for the Ringsted deposition measurements (STACKS-D not visible)



Ensemble outcomes depend on the underlying individual model outcomes. It has not been investigated whether some models are more sophisticated than others. An ensemble ideally only contains models that are expected to perform reasonably well.

7 Comparison with the model intercomparison study

Eight models were used in this study to compare model outcomes with measurements. The same models have previously been used to identify differences among models for hypothetical cases [2]. Outcomes of the current study are used to verify whether the observations made in that study also apply to the current cases. The spread of outcomes for different models is discussed in the current chapter. The relative order of models is discussed in Appendix 7.

7.1 Introduction

The spread of concentration and deposition outcomes within the group of individual models was analysed for hypothetical cases and different types of land cover in the model intercomparison study [2]. Several options to define 'spread' were explored in that report. Among the alternatives, the Geometric Standard Deviation (GSD) was believed to be the most useful metric to define the spread of model outcomes. The definition of GSD is provided in section 2.4.

In the intercomparison study, GSDs were calculated for different cases and types of land cover. The most relevant outcomes are repeated in Figure 7.1 (concentration outcomes) and Figure 7.2 (deposition outcomes). In general, GSDs for concentration mostly ranged between 1.5 and 2. Higher values were found at very short distances from an industrial source, very close to an animal farm building, and, for sources of ammonia, also at distances beyond 2 km. GSDs for deposition were slightly higher than for concentration; they usually ranged between 2 and 3, but were higher at very short distance from the industrial stack and the farm building.

Figure 7.1 GSD for calculated annual-average concentrations in the model intercomparison study, using outcomes for heterogeneous land cover

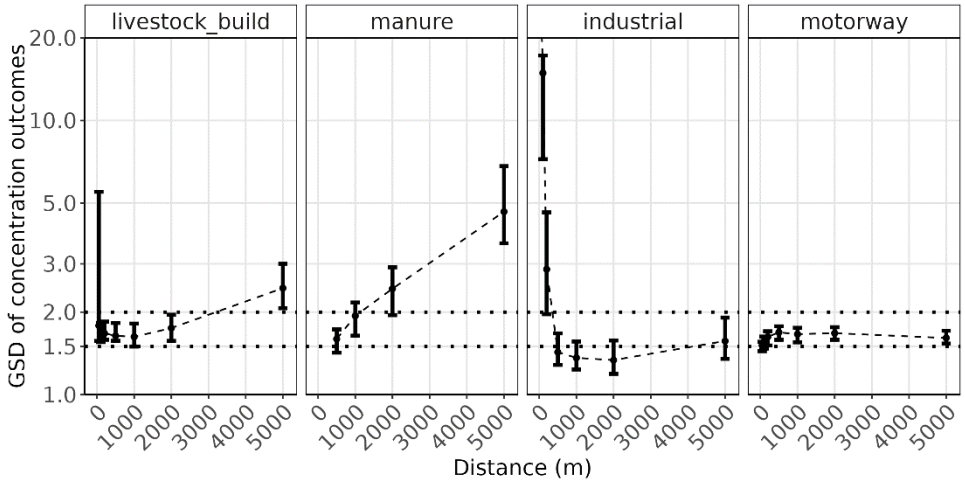
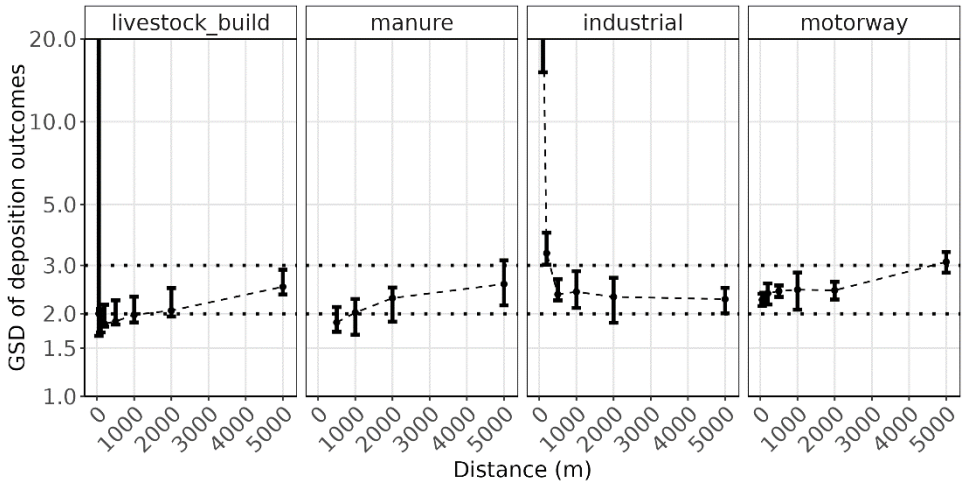


Figure 7.2 GSD for calculated annual-average deposition fluxes in the model intercomparison study, using outcomes for heterogeneous land cover



The individual model outcomes in the current study can be used to verify whether the above values for hypothetical cases also apply to the current cases. In line with the model intercomparison study, we now compare concentrations without background across models, and for deposition only the contribution by the source.[36]

7.2 Ringsted poultry farm

Outcomes for the Ringsted campaign are compared with the outcomes for the livestock farm with building case (and heterogeneous terrain) in the model intercomparison study [2]. The results of the comparison are shown in Figure 7.3.

[36] Model outcomes are more similar when background is included and less similar when background is excluded. In order to compare with the model intercomparison study, we chose to exclude background.

Figure 7.3 Comparison of the GSD's in the intercomparison study (black) and the Ringsted case (blue) for concentration (left) and deposition outcomes (right). The error bar shows the range of GSDs from the 5th to the 95th quantile in the intercomparison study.

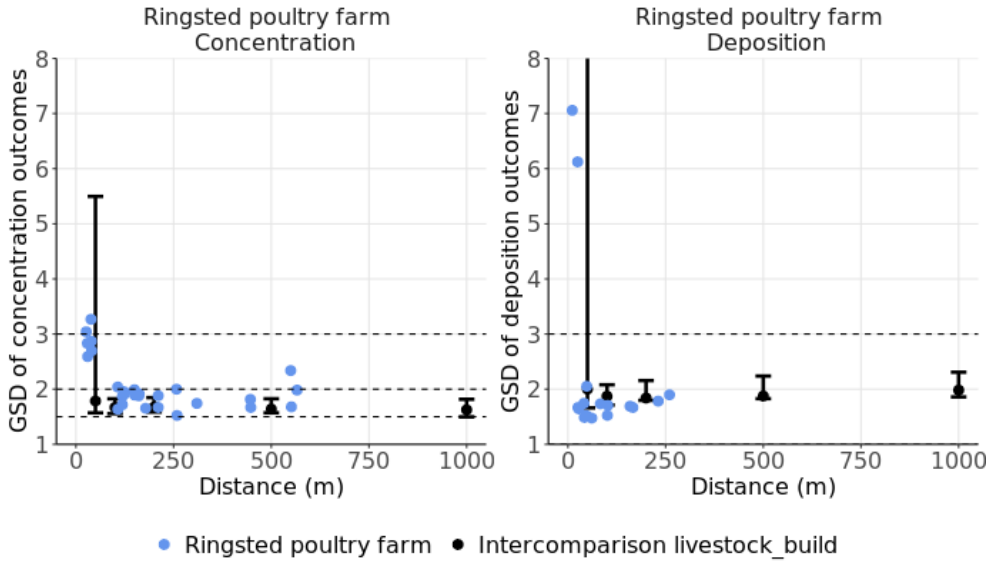


Figure 7.3 shows that the spread of models regarding outcomes for Ringsted are very similar to the spread that was observed in the model intercomparison study, especially farther away from the source. GSDs for concentration mostly range between 1.5 and 2. GSDs for deposition are below 2 for Ringsted, except for two points very close to the buildings of the Ringsted farm. Higher GSDs very close to buildings had also been observed in the intercomparison study.

7.3 Affligem motorway

Outcomes for the Affligem campaign are compared with the outcomes for the motorway case (without a noise barrier) in the model intercomparison study [2]. Outcomes for Affligem involve the six measurement locations from the campaign plus five additional locations that were defined by RIVM for extra model intercomparison. The results of the comparison are shown in Figure 7.4. GSD outcomes for the intercomparison study apply to NO_x concentrations, while GSDs for Affligem involve NO_2 concentrations. In both datasets, land cover was assumed to be homogeneous grass.

Figure 7.4 Comparison of the GSD's in the intercomparison study (black) and the Affligem case (blue) for concentration outcomes. The error bar shows the range of GSD's from the 5th to the 95th quantile in the intercomparison study.

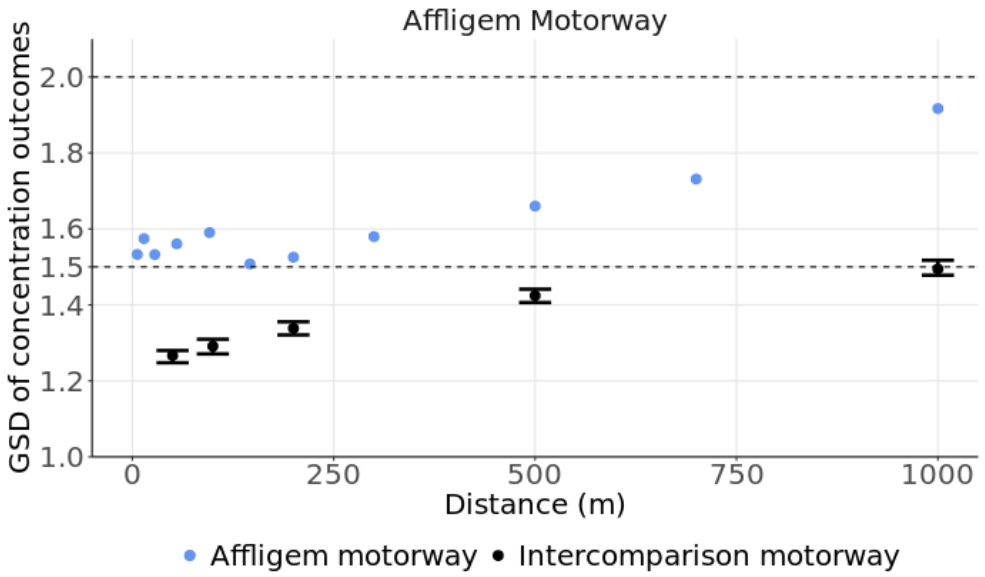


Figure 7.4 shows that the spread across model outcomes in the Affligem study is slightly higher than the spread in the motorway case of the model intercomparison study. Similar to the intercomparison study, GSDs also show some increase with distance.

The reason for the slightly higher GSDs for the Affligem study may be related to modelling NO_x concentrations versus NO_2 concentrations: the first does not require the modelling of chemistry while the latter does. Another cause could be the larger fraction of hours with extremely stable weather conditions for Affligem. The model intercomparison study already highlighted that these hours are important for outcomes and differences between models. Figure 4.2 shows that the time fraction of hours with extremely stable weather conditions was close to 40% in the Affligem campaign[37], compared with 20% in the model intercomparison study.

7.4 Balko compressor station

Outcomes for the Balko campaign are compared with the outcomes for the industrial case (and heterogeneous terrain) in the model intercomparison study [2]. Outcomes for Balko involve the four measurement locations from the campaign plus twelve additional locations that were defined by RIVM for extra model intercomparison. The results of the comparison are shown in Figure 7.5. GSD outcomes apply to modelled NO_x concentrations. IFDM did provide NO_x output for the intercomparison study, but not for Balko.

[37] This number is an estimate, using the Monin-Obukhov length calculated with OPS-ST for each hour.

Figure 7.5 Comparison of the GSD's in the intercomparison study (black) and the Balko case (blue) for concentration outcomes. The error bar shows the range of GSD's from the 5th to the 95th quantile in the intercomparison study. The y-axis is on a logarithmic scale.

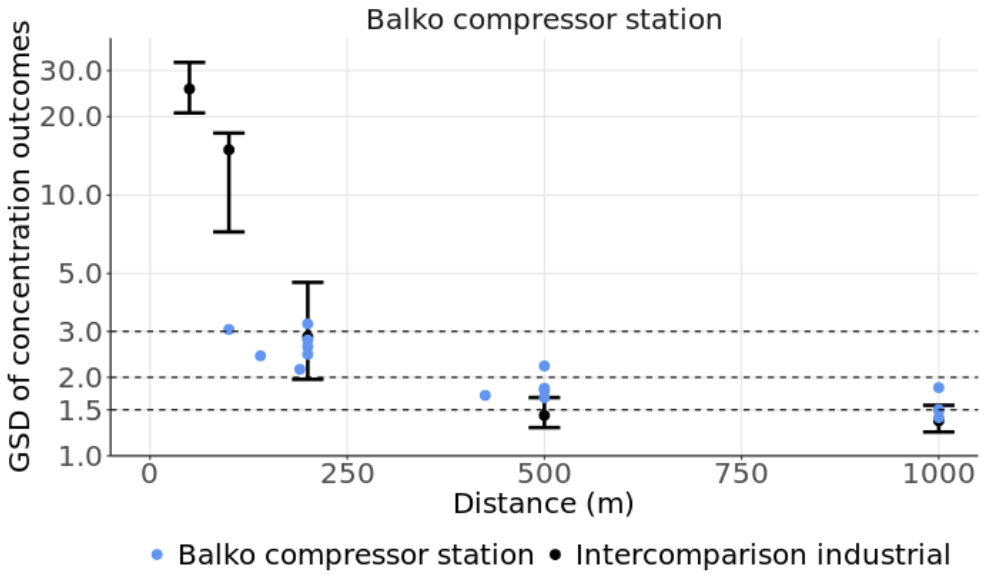


Figure 7.5 shows that the spread among model outcomes in the Balko study is similar to the values calculated for the industrial case in the model intercomparison study at distances beyond 500 m, and (much) smaller at shorter distances. The smaller GSDs at short distances for Balko may be related to the lower emission height of the source, resulting in less uncertainty about where the plume impacts the ground. Emissions, temperatures and velocities were, however, higher in the Balko study.

8 Discussion

In this chapter, we discuss the scope of the findings in this study and the limitations on the conclusions we can draw from these findings.

The first objective of this study was to make a statement about the accuracy of individual models regarding calculated concentrations and deposition fluxes close to the source, and on the basis of a comparison of model results with measurements. The second objective was to determine whether the use of ensemble modelling has added value for model calculations at local scale.

8.1 Using performance indicators for measuring the accuracy of models

The accuracy of individual models was determined by comparing model outcomes with measurements from three measurement campaigns. The degree of agreement with measurements was derived from five commonly used performance indicators: Fractional Bias (FB), Normalised Mean Square Error (NMSE), Geometric Mean Bias (MG), Geometric Variance (VG), and the fraction of model outcomes within a factor of two of measured outcomes (FAC2).[38] These indicators show whether a model has an average bias compared with measurements and what the average deviation is between model outcomes and measurements [26].

Indicators are based on differences between model results and measurements and depend on the accuracy of the model, the accuracy of the inputs for the model calculation, and the accuracy of the measurements. The main input parameters for model calculations are the source strength, the emission characteristics, the local weather conditions, the surface roughness of the environment, and the background concentration.

Outcomes for performance indicators relate to the measurements that were used: the characteristics of the source, the measurements locations, the duration of the measurements, and the weather conditions during that period. Table 8.1 provides an overview of the measurement locations that were used in this study.

Regarding the calculation and use of the selected performance indicators, the following limitations were identified in this study:

- Fractional Bias (FB) and Normalised Mean Square Error (NMSE) depend on mean calculated outcomes, mean measured values and mean differences between the two. In this study, measured and modelled outcomes often reduced significantly with increasing distance to the source. Then, locations close to the source have a much larger influence on mean values and corresponding results for FB and NMSE than locations farther

[38] The signs of FB and MG were reversed, compared with Hanna & Chang. In this study, FB>0 and MG>1 both indicate average overprediction (not underprediction) by a model.

away (see sections 13.1.2, 13.2.2, 14.2, and 15.2 for details). In other words, outcomes for FB and NMSE mostly reflect the model performance at the shortest distances. Several papers in literature [26], [30], [37], highlighted that the geometric indicators Geometric Mean Bias (MG) and Geometric Variance (VG) may provide a more balanced outcome for all measurements if outcomes vary significantly in either space or time. Consequently, FB and NMSE would be more useful when outcomes fluctuate around a mean value, as is often observed in time series.

- However, the geometric indicators, MG and VG, can be strongly affected by large relative deviations for extremely low outcomes (concentrations or deposition fluxes), for example, near or below instrument thresholds or model cut-offs ([26], [30]). This is particularly relevant for the Ringsted deposition measurements, where assumed background contribution had been subtracted from the total measured deposition. Model results for source contribution will deviate significantly from the 'measured' contribution, if the assumed background contribution in measurements is inaccurate. Therefore, Chang & Hanna [26], Borrego et al. [30] and Schatzmann et al. [38] recommend using a threshold for calculating MG and VG: a lower bound for measured and modelled outcomes. For the Ringsted deposition measurements, it was decided not to use such a threshold. The motivation was that any threshold value would have been arbitrary, and would have had a large impact on results. Only negative measurements and model outcomes were excluded when calculating MG and VG, because the underlying formulae do not allow negative values.
- FAC2 is defined as the fraction of calculated outcomes within a factor of two from measured outcomes. In this study, we had few outcomes to intercompare between models and measurements: six for the Affligem study and four for the Balko campaign. Outcomes for FAC2 are, then, particularly sensitive to minor differences in model or measurement outcomes, for example, it can range from 0.5 to 0.75 and vice versa.

On the basis of the above considerations, the indicators MG and VG were believed to provide a better picture of model performance than FB, NMSE, and FAC2. However, results for all five indicators were generally reasonably consistent. It was, therefore, possible to get an understanding of model performance and relative differences between models and model ensembles by considering all five indicators together.

The accuracy of model results derived from validation studies is not identical to the accuracy of outcomes for standard model applications. In validation studies, most input parameters are known with relative high accuracy. For example, extra efforts are made to accurately determine the source strength and to measure local weather conditions during the measurement period. For standard model applications, input values for source strength and meteorology involve estimates, with associated uncertainties. The increased uncertainty in input data translates into larger uncertainties in output parameters. The contribution of input uncertainties to overall output uncertainty can be derived from

sensitivity analyses and subsequently be compared with the 'intrinsic' model inaccuracy.

8.2 Acceptance criteria for model performance indicators

It is generally accepted that model outcomes may deviate from measurements. This is because no model can be expected to be perfect, but also because the accuracy of both the input data and the measurements is known to be finite [26]. For concentration calculations, this has resulted in acceptance criteria that define which margins between model outputs and measurements are generally considered acceptable. Many of the publications that were examined in this study use criteria proposed by Chang and Hanna, either their first set of criteria from 2004 [26], or the later criteria for rural and urban application from 2012 [39].

In this study, the Hanna & Chang criteria for measurements in rural environments [39] were used to test the acceptability of performance indicator outcomes. These criteria were believed to be better suited than the less stringent criteria for measurements in urban environments. Due to differences between the measurements used in the current study and measurements used by Hanna & Chang, the criteria should be used with prudence: results for the current study do not indicate whether or not a model is suitable for regular model applications.

- Hanna & Chang's criteria were intended as a starting point for further discussion and have no official status. 'The suggestions of acceptance criteria in this paper are intended to represent a start to the discussion and it is hoped that further studies can help develop acceptance criteria for other applications and types of available data.' [39] 'The proposed rural and urban model acceptance criteria themselves are somewhat arbitrary, and a more valid and widely recognized set is expected to result from further testing of these acceptance criteria with more models and field datasets by a wide array of research groups.' [39]
- Hanna & Chang's criteria are based on experiences with field experiments and apply to arc-wise maximum concentrations. The field experiments mainly consisted of short-duration runs with well-known emission characteristics of a single source. 'The acceptance criteria suggested here are for evaluations with research-grade field experiment data, with the sources well known, with on-site meteorological observations, and with an extensive array of high-quality concentration samplers.' [39] The current study involves a 'point-by-point' comparison of time-averaged concentrations at specific locations, for more complex dynamic sources in a realistic environment with obstacles (such as buildings near the source).
- The dataset that Hanna & Chang used to derive their acceptance criteria includes more tests and more measurement locations than the current study. In addition, distances to the source appear to be on average larger than in the current study.
- Hanna & Chang propose to measure the performance of models with FB, NMSE, FAC2, and NAD. In the present study, we use FB, NMSE, MG, VG, and FAC2. For the measurement data in this study, the results for FB and NMSE depend to a large extent on

the measured and modelled concentrations at the measurement location closest to the source (see section 8.1). Therefore, FB and NMSE do not provide a balanced picture for the accuracy of the models across the entire range of measurement locations. MG and VG are more representative of the entire set of measurement locations. Acceptance criteria for MG and VG were not reported in [39], but could be derived from the accepted factor random scatter factor reported in [39]. NAD was not used in this study; first, because its outcomes depend on subjective threshold concentrations, and second, because measurement locations were not placed in concentric circles.

No specific acceptance criteria for deposition modelling were found in literature. The criteria for concentration modelling cannot be applied to deposition modelling because (i) deposition modelling is inherently more complex, and (ii) deposition measurements are less accurate.

8.3 Availability of measurements for validating modelled deposition

An important limitation of this study is that only one campaign was found that fulfilled the requirements for validating deposition outcomes. We believe that validating deposition modelling at local scale requires prolonged deposition measurements close to a dedicated source (see section 2.1.1). Such measurements are believed to be scarce because (i) it is not the main focus of research programmes, and (ii) measuring reactive nitrogen deposition near a source is difficult (see section 2.1.1 and Appendix 1, section 12.3.1).

When useful deposition measurements near a source are lacking, the question arises whether the validity of deposition modelling at local scale cannot be tested or explored in a different way. Then, the most plausible route is to validate concentrations in the vicinity of the source with local concentration measurements and to validate deposition schemes separately with regional deposition measurements. This would be a solution if concentrations near a source were similar in magnitude to regional average concentrations, but in general, they are not; that is, local-scale concentrations are much higher than regional average concentrations.

Yet another option is to compare measured spatial variation in air concentrations with modelled variation. The assumption is that emissions vary locally. Then, local patterns in air concentration are only correctly predicted by models if the models use reasonably accurate deposition velocities. Such an analysis requires that models can account for multiple sources and that the emissions of local sources are known with sufficient detail. In addition, the spatial resolution of such concentration measurements must be high, in order to identify local variation with precision. A comparison with measurements at regional scale was carried out in [3], using different models.

8.4 Accuracy of the Ringsted deposition measurements

Deposition in the Ringsted measurement campaign was analysed at 25 separate locations, using three ryegrass pots for each location [6]. The deposition rate was calculated for each pot, using the ^{15}N fraction in

that pot relative to the average ^{15}N fraction at a 'control location' 320 m west of the farm. Because the initial ^{15}N fraction in the pots from the feed was larger than the ^{15}N fraction in air, reduced ^{15}N fraction corresponds to increased deposition. Six locations had larger average ^{15}N fractions than the average ^{15}N fraction at the control location, resulting in negative calculated deposition rates. The choice of the reference locations was motived by the claim that 'there were few occasions when the wind was blowing from the farm towards this location, that is, from the 60° direction'. However, Figure 1 in Sommer's paper shows that the frequency of winds from this direction was certainly not negligible (see also Figure 3.4 in this report). It is, therefore, unclear whether the ^{15}N fraction relating to NH_3 background concentrations was properly chosen.

Subsequently, the mean value of three calculated deposition rates per location (one outcome for each pot) was used as the best estimate for deposition. In addition, a standard error was calculated for each location, equal to the standard deviation of the three outcomes divided by √3. Averaged over all 25 locations, the standard error was 137% of the assumed deposition rate per location.

Five measurements in southerly direction were discarded by Sommer et al. because of possible interference with an ammonia source south of the farm (not further specified, but visible in Figure 3.8). Four more measurements were discarded because the measurement outcome was negative and considered to be unrealistically low for the corresponding locations. According to Sommer, 'the loss could have been caused either by plant disease or by damage from wild animals eating the plants'. Two more measurements with negative outcomes were not discarded, because the deposition was expected to be very low at these locations. It is unclear why the deposition at these two locations is lower than the calculated deposition rate at the control location (set to zero), when the latter represents the deposition relating to NH_3 background concentrations, and the first two represent source contribution plus background.

The resulting outcomes for measured deposition (Figure 3.8) did not demonstrate the same smooth decline with distance, that measured concentrations (Figure 3.7), modelled concentrations (Figure 3.9), and modelled deposition (Figure 3.11) demonstrate. At least five out of the sixteen remaining outcomes are close to the assumed background value. These five outcomes are particularly sensitive to inaccuracies in the estimation of the background.

Overall, the Ringsted deposition measurements do not appear to be as accurate as had been hoped for, and are less accurate than the concentration measurements used in this study.

8.5 Individual model performance

Five sets of measurement data from three different measurement campaigns were used for validation (see Table 8.1). For Affligem and Balko, models were validated against averaged concentrations for the full measurement period, even if weekly (Affligem) and hourly (Balko)

measurement outcomes were available. The period-averages were used, however, because deposition is normally expressed in terms of annual average values. Therefore, this study focused on the accuracy of model outcomes for annual average concentrations and deposition fluxes (see section 2.3). The majority of measurement locations ranged between 5 and 200 m distance from the source. In the Ringsted campaign, the largest distance for concentration measurements was 570 m, and the largest distance for deposition measurements was 260 m. In the Affligem campaign, the furthest measurement location was 150 m from the motorway. For the Balko campaign, the furthest location was located at 425 m distance from the dominant source.

Table 8.1 Summary of selected measurements and their distances to the nearest source

Measurement	**Data selection**	**Number of measurements at specified distances**		
		<50m	**50-200m**	**>200m**
Ringsted NH_3 concentrations	27 measurements with durations of 10 to 17 days at 15 different locations	6	11	10
Ringsted N-deposition	16 valid measurements, duration 54 days	8	6	2
Affligem NO_2 concentrations	6 locations, concentrations averaged out over full period (36 weeks)	3	3	
Balko NO_2 concentrations	4 locations, concentrations averaged out over full period (13 months)		3	1
Balko NO_x concentrations	4 locations, concentrations averaged out over full period (13 months)		3	1
Total		17	26	14

Unless otherwise stated, outcomes for OPS-LT in this section refer to its 2024 version. The current (2025) version includes two new model improvements for emissions from road traffic. When using these two improvements, OPS-LT shows a much better agreement with calculated NO_2 concentrations in the Affligem campaign (see section 4.2.2).

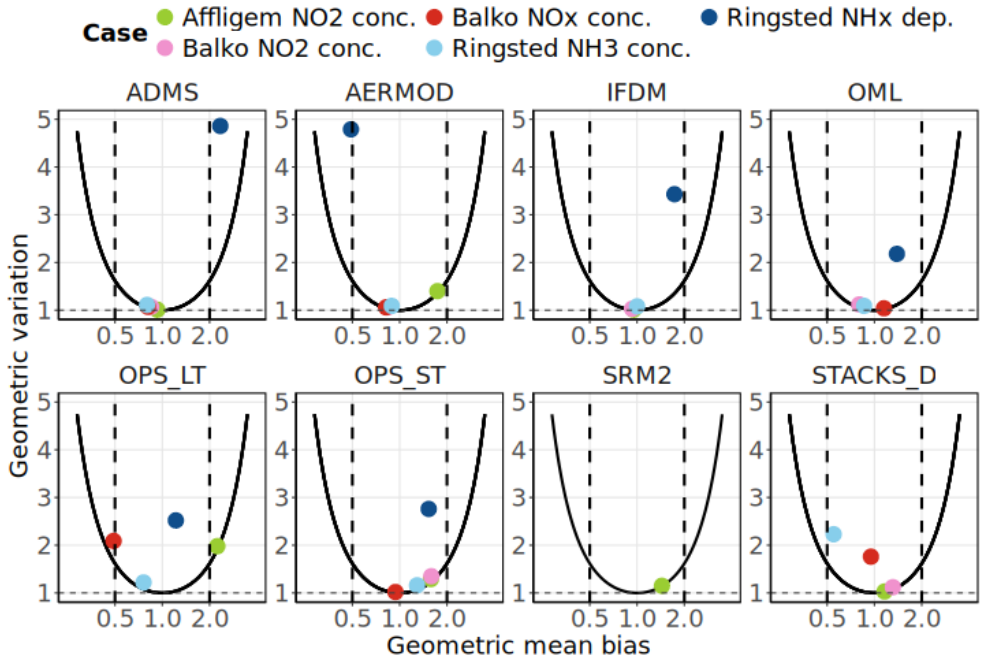
Figure 8.1 shows the Geometric Variance (VG) and Geometric Mean Bias (MG) for all cases and models combined. The following can be observed in the graph:

- For all models that calculated deposition in the Ringsted campaign, VG is the largest for that dataset. Discrepancies between measured and modelled deposition in this campaign are

thus considerably larger than what is normally observed for concentrations. It is expected that these larger deviations are the result of two factors: (i) deposition modelling is inherently more complex than concentration modelling, and (ii) deposition measurements in this campaign are expected to be (much) less accurate than the concentration measurements of this study (see section 8.4).

- In contrast to VG, the Geometric Mean Biases (MG) for Ringsted deposition measurements (ranging between 0.5 and 2.3) are similar to the values found for concentration measurements.
- Hourly models generally perform well for concentrations. Out of these models, IFDM, OML-Multi, and ADMS yield better performance results than AERMOD and OPS-ST. Regarding these concentration measurements, OPS-LT and STACKS-D do not perform as well as the hourly models.
- For deposition, OML-Multi and OPS (both LT and ST) seem to perform better than the other models. The limited accuracy of these deposition measurements does not allow drawing strong conclusions.

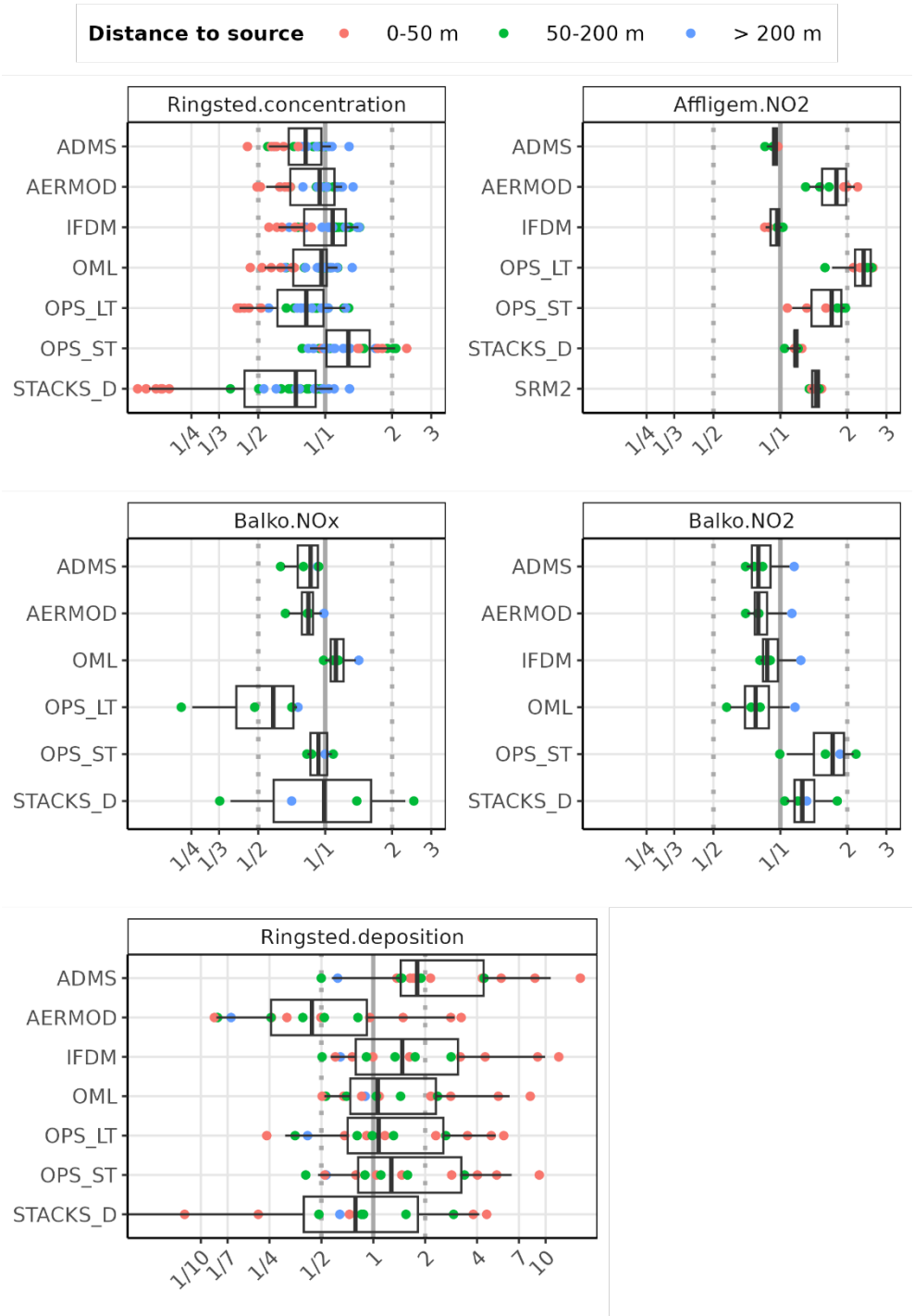
Figure 8.1 Geometric Variance (VG) and Geometric Mean Bias (MG) for the different campaigns per model. The STACKS-D outcome for the Ringsted NH_3 deposition is not visible (off the scale).



An alternative way of visualising differences between calculated and measured data, is to show the range of ratios between calculated and measured outcomes (Figure 8.2). This plot shows the median ratio as well as the variance of the ratios around their median value. When using colours to discriminate between various distances, we can see, for example, that the largest underpredictions of concentrations in the Ringsted campaign, occur relatively close to the source. The total performance score of models is much affected by these outcomes. The

largest overpredictions of measured deposition in the Ringsted campaign also correspond to locations close to the source.

Figure 8.2 Ratios of calculated and measured outcomes for different models and campaigns. Open circles: individual measurements, diamond: median ratio. box: range from 25th to 75% percentile, whiskers: range to 5th and 95th quantile. Note: y-axis scale differs between concentration and deposition measurements.



For concentration measurements, it was furthermore investigated whether performance criteria from Hanna & Chang regarding concentration measurements in a rural environment [39] were fulfilled. According to Hanna & Chang, models cannot be expected to fulfil all criteria for each campaign. Instead, the performance criteria 'should be met over half the time, on average, at all field experiments tested' or 'at least half of the performance criteria are met for at least half of the field experiments considered' [39].

Table 8.2 summarises the results of the comparison of calculated indicators with the performance criteria for concentration measurements in rural terrain [39]. ADMS, IFDM and OML-Multi had particularly good scores, with all (IFDM and OML), or nearly all (ADMS) performance criteria being fulfilled. AERMOD had substandard results for the Affligem campaign (three criteria not met), but good results for the Ringsted and Balko measurements. STACKS-D had substandard results for the Ringsted campaign, and mostly good results for Affligem and Balko measurements. OPS-ST were substandard for Balko NO_2 measurements and, to slightly lesser extent, also for Affligem NO_2 measurements. OPS-LT outcomes (2024 version) were poor for Affligem NO_2 measurements (four criteria not met) and substandard for Balko NO_x measurements.[39] SRM2 was only validated against Affligem data, with below-average results (two out of the five criteria are not met).

[39] Note: the latest 2025 version of OPS-LT had a much better agreement with Affligem NO_2 measurements. When outcomes of this version were used, all five performance criteria were fulfilled. See Appendix 8 for more information.

Table 8.2 Fulfilment of performance criteria for rural campaigns from Hanna & Chang [39] for individual models

Model	Ringsted	Affligem	Balko
ADMS	Criterion for FB not fulfilled	All five criteria fulfilled	All five criteria fulfilled, for NO_2 and NO_x
AERMOD	All five criteria fulfilled	Criteria for FB, MG and VG not fulfilled	All five criteria fulfilled, for NO_2 and NO_x
IFDM	All five criteria fulfilled	All five criteria fulfilled	All five criteria fulfilled for NO_2. No output for NO_x
OML-Multi	All five criteria fulfilled	Not run	All five criteria fulfilled, for NO_2 and NO_x
OPS-LT 2024 version	Criterion for FB not fulfilled	Criteria for FB, MG, VG and FAC2 not fulfilled	Criteria for FB, MG and VG not fulfilled for NO_x. No output for NO_2.
OPS-LT 2025 version	Not run	All five criteria fulfilled	Not run
OPS-ST	Criterion for FB not fulfilled	Criteria for FB and MG not fulfilled	All five criteria fulfilled, for NO_x. Criteria for FB, MG and VG not fulfilled for NO_2.
SRM2	Not run	Criteria for FB and MG not fulfilled	Not run
STACKS-D	Criteria for FB, MG and VG not fulfilled	All five criteria fulfilled	Criterion for MG not fulfilled for NO_2. Criterion for VG not fulfilled for NO_x

Between the models, the hourly models (ADMS, AERMOD, IFDM, OML-Multi and OPS-ST) mostly get better results than the annual models (OPS-LT, SRM2 and STACKS-D). Underlying reasons why some models perform better than others were not further investigated. Such an analysis either requires a thorough review of the various model formulations, which was not feasible within the provided budget, and/or comparison of more detailed model output data,[40] which was not available. One possible explanation of the general better performance of hourly models is the higher temporal resolution of these models relating to the use of hourly meteorology. In addition, ADMS, AERMOD, IFDM, and OML-Multi model hourly variation in emissions in high detail. It is expected that higher temporal resolution improves model performance, in particular at distances relatively close to the source. The Affligem

[40] Most prominently plume centreline height, plume depth (σ_z), effective deposition velocity, and underlying parameters.

case is slightly different because this source (a motorway) is a line-source. In this case, the concentration at a receptor is the sum of contributions from different motorway sections at different distances. Thus, perpendicular distance has limited meaning. Indeed, results for STACKS-D and the new version of OPS-LT for Affligem were as good as the results for hourly models.

Deposition outcomes were not compared with performance criteria because no such criteria have been proposed yet. For all models, the performance results for the Ringsted deposition measurements were worse than the performance results for any of the concentration measurements. This is expected to be a result of (i) deposition modelling being inherently more complex than concentration modelling, and (ii) deposition measurements being less accurate than concentration measurements.

8.6 Understanding causes of differences between models

The current study highlighted differences between models and corresponding differences in the agreement of model outcomes with measurements. Underlying causes of differences between models could not be investigated, because only calculated concentrations and deposition fluxes were provided and model formulations could not be compared due to limitations in the amount of time available for this study. For future projects, understanding the causes of differences is believed to be as important as quantifying differences. Models can only be improved if causes of deviations are sufficiently understood.

8.7 Benefits of ensemble modelling

The main goal of this study was to identify whether ensemble modelling provides benefits for the evaluation of concentration and deposition at local scale. In the SAGEN programme plan [1], the following assumed benefits are listed:

1. The mean/median outcome of the ensemble usually performs better in comparison with measurements than a single model (assuming all models are of good/comparable quality).
2. Ensembles provide insight into the uncertainty range due to different modelling approaches.
3. The mean/median outcome reduces the risk of using the most optimistic or pessimistic model.
4. The use of an operational ensemble promotes the exchange of knowledge, and leads to faster model improvements.

These four assumed benefits will be discussed in the following subsections.

Investigating possible disadvantages of ensemble modelling was outside of the scope of this study. Two disadvantages of ensemble modelling are very briefly discussed in section 8.7.5.

8.7.1 Performance of ensembles

The performance of ensembles was investigated in Chapter 6, using three separate ensembles: a subset of three hourly models (group 1), the set of all hourly models (group 2), and the set of all models

(group 3). The performance of these ensembles was measured using the same indicators as those used for individual models. Results in terms of MG and VG are repeated in Figure 8.3.

When looking across the three campaigns with concentration measurements, the performance results for the ensemble of all hourly models (group 2) were similar to those of IFDM, and better than the results for the other individual models. On average, outcomes for the smallest ensemble (group 1) were better than all individual models except IFDM, and the performance of group 3 was similar to that of ADMS and OML.

For deposition measurements in the Ringsted campaign, the ensembles of all hourly models (group 2) and all models (group 3) had better outcomes than the group of three hourly models (group 1). The results for groups 2 and 3 are about as good as those for OML-Multi and OPS-LT (the best performing individual models for this set of measurements), and perhaps more robust.

Overall, the two largest ensembles achieved consistently good performance results in this study, while the smallest ensemble (group 1) had good results for concentration measurements but poorer results for deposition measurements. On average, therefore, using ensemble modelling has a positive effect on the accuracy of outcomes. At the same time, the increased accuracy is perhaps not as large as may have been anticipated prior to this study, in particular for concentration measurements (where many hourly models also had good results).

Figure 8.3 Comparison of the performances of ensembles and individual models, using outcomes for Geometric Mean Bias (MG) and Geometric Variance (VG). Regarding Ringsted deposition: outcomes for STACKS-D are not visible (off the scale).

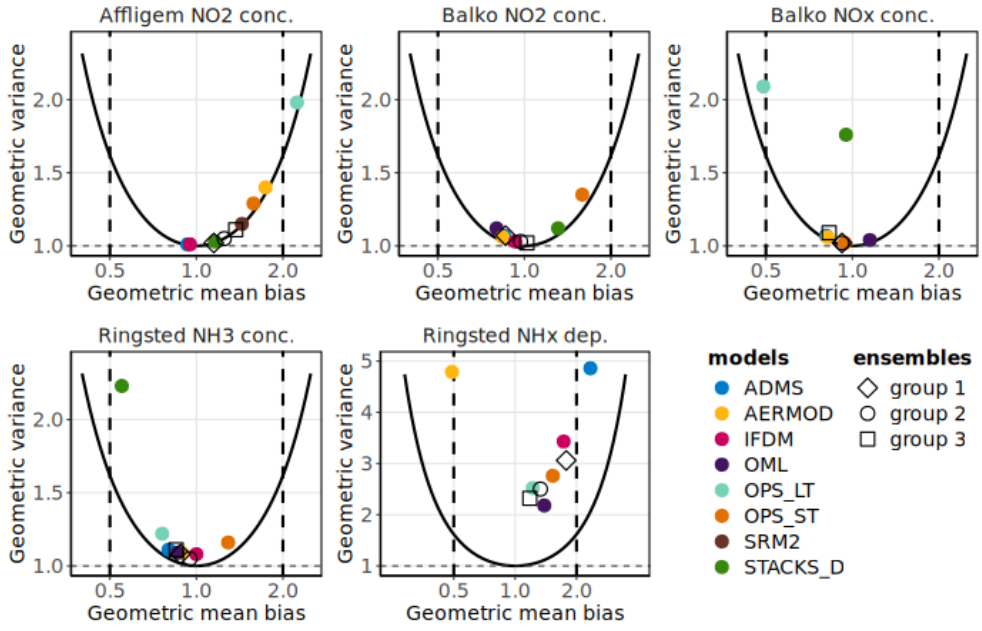
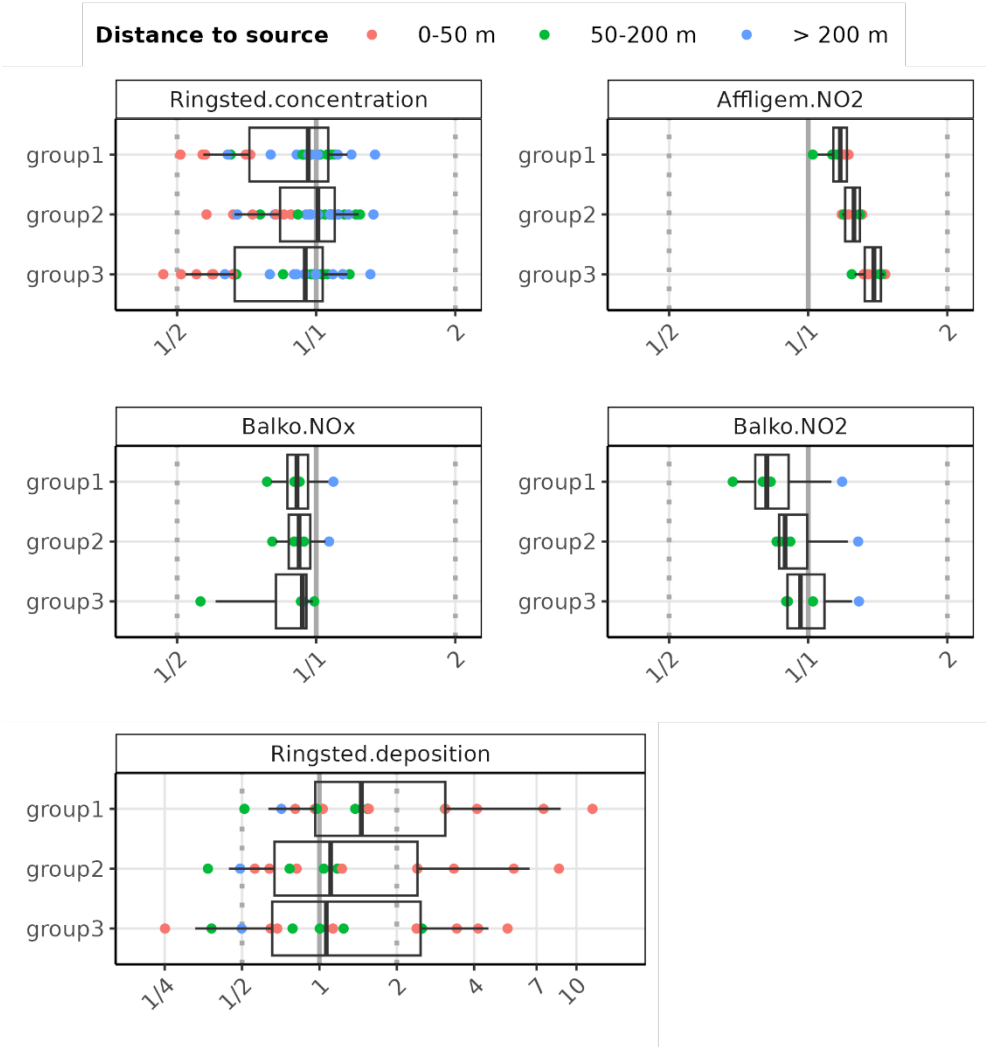


Figure 8.4 shows ratios of calculated and measured concentrations for the three ensembles. Similar to Figure 8.3, it shows which ensembles perform better or worse, and to what extent, for which campaigns. Partly due to the different scale, differences between the three ensembles are now clearer.

Figure 8.4 Ratios of calculated and measured outcomes for different ensembles and campaigns. Coloured points: individual measurements. Box: range from 25th to 75% percentile, whiskers to 5% and 95%. Note: y-axis scale differs between concentration and deposition measurements.



Compared with the acceptance criteria proposed by Hanna & Chang for rural environments [39], groups 1 and 2 fulfil all five criteria for all considered concentration measurements (Table 8.3). Group 3 fulfilled most criteria for all these measurements.

Table 8.3 Fulfilment of performance criteria for rural terrain from Hanna & Chang [39] for model ensembles

Ensemble	Ringsted	Affligem	Balko
Group 1	All five criteria fulfilled	All five criteria fulfilled	All five criteria fulfilled, for NO_2 and NO_x
Group 2	All five criteria fulfilled	All five criteria fulfilled	All five criteria fulfilled, for NO_2 and NO_x
Group 3	Criterion for FB not fulfilled.	Criteria for FB and NMSE not fulfilled.	Criterion for FB not fulfilled for NO_2. All five criteria fulfilled for NOx

8.7.2 Uncertainties derived from ensembles

A second assumed benefit of ensemble modelling was that ensembles may be capable of providing accurate estimates of uncertainties in the ensemble outcome. This topic was investigated in section 6.6. The results of this study revealed that the Standard Deviation (SD) and the Geometric Standard Deviation (GSD) of individual model outcomes are both good measures for differences between measured values and the ensemble outcomes. This conclusion, however, only applied to the two larger ensembles (groups 2 and 3). For the smallest ensemble, deviations with measurements could be much larger than should be expected from the SD and GSD. The most likely explanation is that this ensemble is too small; outcomes for the three models in this ensemble can be very similar, resulting in an underestimation of uncertainty.

From a conceptual point of view, GSD is expected to be a better indicator of uncertainty than SD, because outcomes (concentrations or deposition fluxes) are bounded by a minimum (zero). Indeed, the distribution of individual model outcomes around their joint mean, resembled the lognormal distribution better than the normal distribution (see section 6.6). Chapter 7 provides a comparison of all GSDs calculated in this study with GSDs calculated in the model associated model intercomparison study [2]. With some exceptions, GSDs in this study were very similar to the values that were found in the intercomparison study (see Chapter 7).

8.7.3 Reducing the risk of using the most optimistic or pessimistic model

The third assumed benefit of using ensemble modelling, expressed in the SAGEN programme plan [1], was that it reduces the risk of using the most optimistic or pessimistic model. This topic has not been investigated in this study. Indeed, ensemble outcomes are by construction in the middle of underlying individual model outcomes. Ensemble modelling also has a price: larger calculation times and higher financial costs. Thus, using proper procedures for selecting an adequate neutral model, may be more efficient and cost-effective.

8.7.4 Promoting the exchange of knowledge and further model development

The fourth and final assumed benefit is that ensemble modelling is supposed to promote the exchange of knowledge, and result in faster

model improvements. The experience from the current project is that exchanging knowledge is mostly a matter of available time. It requires causes of differences to be investigated and addressed. Causes of differences can only be found by comparing model formulations and by studying outcomes, including intermediate outcomes, at a detailed level. The recent improvements in OPS-LT regarding emissions from motorway traffic (section 4.2.2) are an example of successful model improvement by intercomparing model formulations and intermediate outcomes.

The most efficient way appears to be to start with specific cases, such as a benchmark exercise, and under the condition that sufficient time is provided to explore differences between models. An automatic interface for running models in parallel as such is not sufficient to identify and understand differences between models. Ensemble modelling could, however, be a way to formalise collaborations between institutes and to obtain structural funding for the exchange of knowledge and model development.

8.7.5 Disadvantages of ensemble modelling

Investigating possible disadvantages of ensemble modelling fell outside of the scope of this study. Ensemble modelling, however, also has a price. First, financial costs are higher: model input and output must be aligned and a more complex model interface must be developed and kept up-to-date. These costs can be significant. Second, ensemble modelling requires larger computing capacity and increases the duration of calculations. Required computer power for individual model runs can be either small or large, depending on the number of sources and the temporal and spatial resolution.

9 Conclusions

The main goals of this study were to validate concentration and deposition outcomes of operational atmospheric transport models against measurement at local scale, and to investigate possible benefits of ensemble modelling relating to the use of model outcomes at local scale. The work is part of Work Package 2.2 of the SAGEN project [1]. The overall aim of that project is to reduce uncertainties in determining the deposition of reactive nitrogen by using satellite measurements and ensemble modelling. Work package 2.2 investigates possible benefits of ensemble modelling regarding calculations of nitrogen deposition at local scale, where local scale is defined as ranging 'from a hundred metres to a few kilometres' from a dedicated source. This scale is important for permit application.

Mere differences between eight separate atmospheric transport models regarding calculated concentrations and deposition fluxes close to a source, have recently been reported in a model intercomparison study at local scale [2]. The same eight atmospheric transport models were used for the current study. Differences between operational models and regional scale concentration and deposition measurements have been reported in [3], even if a different set of models was used.

The SAGEN programme plan [1] suggested multiple benefits of ensemble modelling. Three ensembles of different sizes were defined to test whether model ensembles provide more accurate outcomes and whether the uncertainty in the outcomes can be derived from the spread of underlying individual model outcomes. The smallest ensemble (group 1) consisted of three models that also allow for calculating concentrations and deposition fluxes for each individual hour, the second group contained all these hourly models and the third group contained all models. This approach allowed for identifying differences between smaller and larger ensembles.

The performance of individual models and ensembles was tested by comparing calculated outcomes with measurements. The available budget allowed for validation of the selected models against measurement data from three campaigns. Significant efforts were undertaken to find suited campaigns for validating deposition outcomes at local scale and with sufficient long duration. Such campaigns proved to be scarce: deposition is mostly measured in nature areas and away from strong local sources. They can be used for validating calculated deposition at regional scale (Work Package 2.1 of the SAGEN project), but not at local scale. Reasons why few deposition measurements are carried out close to a source with sufficient duration can be:

- Understanding the effects of deposition on nature areas and vulnerable ecosystems is regarded as more important.
- Measuring deposition close to a source is more difficult. Dry deposition is important near a source and should be measured with corresponding accuracy. Plant-specific influences on dry deposition should be accounted for. Gradient and eddy covariance techniques appear to be fit for that purpose, but at

local scale suffer from their inability to identify which part of a vertical flux is related to mere mixing (transport of substances away from the plume centreline height), and which part is the result of deposition to vegetation (or re-emission from vegetation). Biomonitors measure effective deposition to vegetation directly, but the accuracy of these measurements still appears to be limited.
- Measuring deposition for a prolonged time is expensive.

Eventually three campaigns were selected to validate outcomes of individual models and ensembles: the Ringsted campaign with concentration and deposition measurements around a Danish chicken farm [6], the Affligem campaign with NO_2 concentration measurements around a Belgian motorway [16] and the Balko campaign with NO_2 and NO_x concentration measurements around a compressor station in the USA [20]. All measurements had been performed during normal operations of these sources. The campaign durations were all large enough to validate annual average output of models. Measurement locations were mostly close to the source; in majority ranging between 5 and 200 m distance from the source. The performance outcomes in this study, therefore, correspond to the performance of models and ensembles at these short distances.

Deposition measurements in the Ringsted campaign were not as accurate as the concentration measurement that were used in this study. Deposition was not measured directly, but calculated from the ^{15}N fractions in pots. The relative standard error of the resulting deposition rates was, on average, higher than 100%. The control location, representing deposition relating to background NH_3 concentration, was not the location with the lowest outcome for deposition. As a result, six pots had lower calculated deposition rates than the control location, possibly contradicting the concept of 'control location'. Four of these were discarded by Sommer, while two other locations with negative calculated deposition rates were not. Five additional locations were discarded by Sommer due to the presence of an additional nearby source. The remaining measurements did not show smooth trends with distance, in contrast to the concentration measurements in the same campaign and modelled deposition fluxes for this campaign.

A literature analysis revealed that the performance of atmospheric transport models is most often investigated using performance indicators proposed by Chang & Hanna [26] and the associated BOOT software for validating air dispersion models [27]. Relating to the decrease of concentrations and deposition fluxes with increasing distance to the source, the Geometric Mean Bias (MG) and Geometric Variance (VG) were believed to be more suited to this study than their arithmetic equivalents Fractional Bias (FB) and Normalised Mean Square Error (NMSE). Results for individual model outcomes were discussed in Chapter 3 (Ringsted), Chapter 4 (Affligem), and Chapter 5 (Balko). Results for ensembles were discussed in Chapter 6. Figure 8.3 (section 8.7.1) shows the calculated values for MG and VG for both individual models and ensembles, and can be regarded as a summary of all outcomes. The boxplots of ratios of calculated and measured outcomes

(Figure 8.2 for individual models and Figure 8.4 for model ensembles) provide an alternative way of summarising outcomes.

Most of the investigated validation studies for air quality models also use acceptance criteria defined by the same authors, either the initial criteria from 2004 [26], or the updated criteria from 2012 [39] that discriminate between concentration measurements in rural and urban environments. These proposals relate to arc-wise maximum concentrations measured in field experiments with well-known sources. The suitability of these criteria for the measurements used in the current study is unknown. The acceptance criteria outcomes should, therefore, be used with prudence, and do not indicate whether or not a model is suitable for regular model applications. Regarding deposition, no acceptance criteria were found in literature. Criteria for deposition modelling should be more tolerant than criteria for concentration modelling, because deposition modelling is inherently more complex and deposition measurements are less accurate. The FAIRMODE framework provides an interesting alternative to measuring model performance but could not be applied because the required parameters were still undefined for the measurement data that was used in this study.

Regarding the performance of individual models and model ensembles, the following can be concluded:

- IFDM, OML-Multi, and ADMS had consistent good scores for all considered concentration measurements.[41] These models, together with AERMOD, are also the models that calculate dispersion for each individual hour, and take hourly variations in emissions into account in great detail. Across the three campaigns with concentration measurements, the ensemble of all hourly models (group 2) performed as well as IFDM and better than all other individual models. The other ensembles considered (group 1 and group 3) performed slightly worse, but still better than most individual models. All the above models and ensembles fulfilled all, or nearly all, of the acceptance criteria proposed by Hanna & Chang for rural environments [39]. SRM2, OPS-ST, STACKS-D and OPS-LT had poorer results than the other individual models and had more failures to meet the Hanna & Chang performance criteria. These four models do not allow for using hourly specific emission rates, and, except for OPS-ST, do not calculate dispersion for each individual hour. Therefore, these limitations appear to result in substandard performance at close distances to the source. For line sources, such as motorways, ‘distance to the source’ is more ambiguous. Indeed for Affligem, both STACKS-D and an updated (2025) version of OPS-LT performed as well as the best hourly models and ensembles.
- Regarding the deposition measurements in the Ringsted campaign, discrepancies between measured and calculated outcomes were generally larger than previously observed for concentration measurement. This is expected to be the result of (i) deposition modelling being inherently more complex, and (ii) these deposition measurements being (much) less accurate than

[41] IFDM wasn’t compared with Balko NO_x concentration measurements, OML-Multi wasn’t compared with Affligem NO_2 concentration measurements, and OPS-LT wasn’t compared with Balko NO_2 concentration measurements.

the concentration measurements used in this study. Out of the individual models, OML-Multi and OPS-LT yielded the best performance results. Scores for the two larger ensembles (groups 2 and 3) were similar to those for OML-Multi and OPS-LT. All three ensembles had better results than most individual models.

As stated before, these conclusions apply to the sources and distances used for validating outcomes. The majority of distances was between 5 and 200 m from the source.

According to the SAGEN programme plan [1], ensemble modelling offers multiple benefits compared with the use of one single model: it is supposed to result in more accurate estimates, to provide more insight into uncertainties in outcomes, to prevent users from using the most optimistic or pessimistic model, and to promote the exchange of knowledge between developers resulting in faster model improvements.

Regarding the increase of accuracy, this study showed that ensembles mostly perform as well as the better individual models in specific campaigns, and better than the majority of individual models when looking across campaigns. At the same time, the increased accuracy is perhaps not as large as may have been anticipated prior to this study, in particular for concentration measurements (where many hourly models also had good results).

Regarding insight into uncertainty, this study showed that the spread between individual models is a good indicator of uncertainty in the ensemble outcome, if the ensemble is not too small. For the two larger ensembles, discrepancies between the ensemble outcome and measurements were in reasonable agreement with the Standard Deviation (SD) or Geometric Standard Deviation (GSD) of individual model outcomes. Between these two indicators of uncertainty, GSD is preferred from a conceptual point of view (see section 8.7.2). For the smallest ensemble, which contained only three models, deviations from measurements can be larger than what is expected from the spread in individual model outcomes.

The third assumed benefit was not investigated. Indeed, ensemble outcomes are by construction in the middle of underlying individual model outcomes. Using proper procedures could be an alternative solution for selecting an adequate neutral model.

Lastly, ensemble modelling was assumed to promote the exchange of knowledge and result in faster model improvements. The experience from the current project is that exchanging knowledge is mostly a matter of available time. It requires causes of differences to be investigated and addressed. This requires studying model formulations and intermediate outcomes at a detailed level (see, for example, section 4.2.2). An automatic interface for running models in parallel as such is not sufficient to identify and understand differences between models. Ensemble modelling could, however, be a way to formalise collaborations between institutes and to obtain structural funding for the exchange of knowledge and model development.

Possible disadvantages of ensemble modelling were not investigated. However, ensemble modelling also has a price: the clearest disadvantages are higher financial costs and longer calculation times.

In summary, this study provides new insight into the performance of models and model ensembles at short distances from the source. An important limitation is that only one suited campaign with deposition measurements was found for validating model outcomes, and with limited precision of the measured deposition fluxes. Measuring dry deposition of reactive nitrogen close to a source appears to be particularly difficult and may require the design of new measurement techniques or significant improvement of the accuracy of experiments with biomonitors. Indirect evidence that models use the right deposition velocities can be obtained by comparing model outcomes with concentration measurements at regional scale, however, only for those models that can be applied to regional scale.

Recommendations

This study showed that some models correspond better with certain measurements than other models. RIVM recommends that the causes of these differences be further analysed in a follow-up study. This can be done by looking at model formulations and intermediate model results in detail. Then, the models that deviate the most from the measurements can be improved using the knowledge gained. The measurement campaigns from the current study and the underlying input data for the model calculations are suited as a basis for such research.

The study also showed that there are hardly any reliable measurements available to validate calculated deposition fluxes in the vicinity of a source. This is because measurement methods, particularly for dry deposition, have very specific requirements. RIVM recommends investigating what is needed to set up a reliable measurement campaign with deposition measurements near an individual source.

10 Acknowledgements

This study received funding from the following organisations:

- The Dutch Ministry of Agriculture, Fisheries, Food Security, and Nature funded the work of RIVM and WUR and partly funded contributions by Aarhus University, UKCEH, CERC, and VITO.
- The UK Atmospheric Dispersion Modelling Liaison Committee (ADMLC) and the Northern Ireland Department of Agriculture, Environment and Rural Affairs providing additional funding to CERC and UKCEH.
- The Flemish Department of the Environment and Spatial Development provided additional funding for the contribution by VITO.

Furthermore, all project partners made use of internal resources to complete this work.

DGMR supported this work by permitting the use of STACKS-D, providing an executable for running STACKS-D, and describing how to use STACKS-D.

The Flanders Environment Agency (VMM) provided measurement data for the Life+ Affligem campaign and allowed RIVM and partners to use this data without any restrictions.

Aarhus University provided measurement data for the Ringsted campaign and allowed RIVM and partners to use this data without any restrictions.

11 References

[1] The use of satellite data and ensemble modelling in the National knowledge program nitrogen (NKS). SAGEN research proposal.

[2] Kooi ES, Thorkelsdottir G, Meijer PA, Stocker J, Lefebvre W, Vigier A, Lansø AS, Krol M, Jacobs CMJ, Van Pul WAJ & Wichink Kruit RJ. Differences in calculations of concentration and deposition of ammonia and nitrogen oxides at local scale - A comparison of eight atmospheric transport models. Report 2025 – 0047. RIVM. 2025. Available from https://www.rivm.nl/bibliotheek/rapporten/2025-0047.pdf (accessed 5-10-2025).

[3] Schaap M, Kranenburg R, Michaud van der Wal CH, Geers L. National scale modelling of nitrogen deposition in the Netherlands. TNO 2025 R11248. TNO. 2025. Available from https://publications.tno.nl/publication/34645402/68FvGQTS/TNO-2025-R11248.pdf (accessed 9-3-2026).

[4] R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2023. www.R-project.org (accessed 14-4-2025).

[5] Hoogerbrugge R, Braam M, Siteur K, Jacobs C, Hazelhorst S, Stefess G, Van der Swaluw E, Wichink Kruit R, Wesseling J & Van Pul A. Uncertainty in the determined nitrogen deposition in the Netherlands - status report 2023. Report 2022-0085. RIVM. 2022. Available from https://www.rivm.nl/bibliotheek/rapporten/2022-0085.pdf (accessed 20-11-2025).

[6] Sommer SG, Østergård HS, Løfstrøm P, Andersen HV & Jensen LS, Validation of model calculation of ammonia deposition in the neighbourhood of a poultry farm using measured NH3 concentrations and N deposition, *Atmospheric Environment*, **43:4**, 915-920 (2009). Available from https://doi.org/10.1016/j.atmosenv.2008.10.045 (accessed 29-1-2025).

[7] Sommer SG. A simple biomonitor for measuring ammonia deposition in rural areas. *Biology and Fertility of Soils* **6**, 61–64 (1988). Available from https://doi.org/10.1007/BF00257922 (accessed 29-1-2025).

[8] Sommer SG and Jensen ES. Foliar absorption of atmospheric ammonia by ryegrass in the field. *Journal of Environmental Quality* **20**, 153–156 (1991).

[9] Phillips SB, Pal Arya S & Aneja, VP, Ammonia flux and dry deposition velocity from near-surface concentration gradient measurements over a grass surface in North Carolina, *Atmospheric Environment* **38:21**, 3469-3480 (2004). Available from https://doi.org/10.1016/j.atmosenv.2004.02.054 (accessed 29-1-2025).

[10] Tietema A, Barmentlo H, Van Loon E, Bol R, Ebben B, Tulp T, Tromp M, Schwennen C, Maas L & Averkamp J. Nitrogen deposition measurements around dairy farms: spatial and temporal patterns. University of Amsterdam. 2023. Available from https://www.uva.nl/binaries/content/assets/faculteiten/faculteit-der-natuurwetenschappen-wiskunde-en-informatica/fnwi-nieuws/eindverslag_12092023_definitief.pdf (accessed 24-6-2025).
[11] Sutton MA, Nemitz E, Theobald MR, et al. Dynamics of ammonia exchange with cut grassland: strategy and implementation of the GRAMINAE Integrated Experiment, *Biogeosciences* **6**, 309–331 (2009). Available from https://doi.org/10.5194/bg-6-309-2009 (accessed 24-6-2025).
[12] Milford C, Theobald MR, Nemitz E, et al. Ammonia fluxes in relation to cutting and fertilization of an intensively managed grassland derived from an inter-comparison of gradient measurements, *Biogeosciences* **6**, 819–834 (2009). Available from https://doi.org/10.5194/bg-6-819-2009 (accessed 24-6-2025).
[13] Loubet B, Milford C, Hensen A, Daemmgen U, Erisman J.-W, Cellier P & Sutton MA. Advection of NH3 over a pasture field, its effect on gradient flux measurements, *Biogeosciences* **6**, 1295-1309 (2009). Available from https://bg.copernicus.org/articles/6/1295/2009/ (accessed 29-1-2025).
[14] Kirchner M, Jakobi G, Feicht E, Bernhardt M & Fischer A. Elevated NH3 and NO2 air concentrations and nitrogen deposition rates in the vicinity of a highway in Southern Bavaria, *Atmospheric Environment* **39:25**, 4531-4542 (2005). Available from https://doi.org/10.1016/j.atmosenv.2005.03.052 (accessed 29-1-2025).
[15] Bettez ND, Marino R, Howarth, RW & Davidson EA. Roads as nitrogen deposition hot spots, *Biogeochemistry* **114**, 149–163 (2013). https://doi.org/10.1007/s10533-013-9847-z (accessed 29-1-2025).
[16] Life+ ATMOSYS snelwegcampagne: Luchtkwaliteit nabij de E40-snelweg in Affligem (in Dutch). Vlaamse Milieumaatschappij (VMM). 2013. Available from https://vmm.vlaanderen.be/publicaties/http-www-vmm-be-pub-life-atmosys-snelwegcampagne-luchtkwaliteit-nabij-de-e40-snelweg-in-affligem-view/@@download/attachment (accessed 25-3-2025).
[17] Website: Air Quality Dispersion Modeling - Preferred and Recommended Models | US EPA (accessed 25 June 2024).
[18] Website: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5 (accessed 11 July 2024).
[19] Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, Nicolas J, Peubey C, Radu R, Rozum I, Schepers D, Simmons A, Soci C, Dee D & Thépaut J-N. (2023): ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.f17050d7 (Accessed on 09-07-2024)

[20] Panek JA, McCarthy JM, Huth AZ, Krol AJ & Nowak C. PRCI ambient NO2 AERMOD performance assessment and model improvement project: Modeled to observed comparison, *Journal of the Air & Waste Management Association*, **70:5**, 504-521 (2020). Available from https://doi.org/10.1080/10962247.2020.1743382 (accessed 25 June 2024).
[21] Jensen SS, Ketzel M, Im U, Løfstrøm P, Poulsen MB, Monies C & Ellermann T. Validering af luftkvalitetsmodellen OML-Highway. Science report nr. 240, Aarhus University, 2006. Available from: http://dce2.au.dk/pub/SR240.pdf (accessed 4-2-2025).
[22] Website: Modelling Tools (HARMO.org) (accessed 4-2-2025).
[23] Website: Datasets | ADMLC (accessed 4-2-2025).
[24] Environmental Protection Agency, 2023. Technical Support Document (TSD) for Adoption of the Generic Reaction Set Method (GRSM) as a Regulatory Non-Default Tier-3 NO2 Screening Option, Publication No. EPA-454/R-23-009. Office of Air Quality Planning & Standards, Research Triangle Park, NC. Available from https://gaftp.epa.gov/Air/aqmg/SCRAM/conferences/2023_13th_Conference_On_Air_Quality_Modeling/Review_Material/AERMOD_GRSM_TSD_20231012.pdf (accessed 25 June 2024).
[25] Evaluation of the Generic Reaction Set Method for NO2 conversion in AERMOD: The modification of AERMOD to include ADMS chemistry. Stocker J, Seaton M, Smith S, O'Neill J, Johnson K, Jackson R & Carruthers D. Cambridge Environmental Research Consultants (CERC) Technical Report. 2023. Available from: https://gaftp.epa.gov/Air/aqmg/SCRAM/conferences/2023_13th_Conference_On_Air_Quality_Modeling/Review_Material/20230808_GRSM_Evaluation_Report_CERC.pdf (accessed 25 June 2024).
[26] Chang JC & Hanna SR. Air quality model performance evaluation, *Meteorology and atmospheric physics* **87**, 167-196 (2004). Available from https://link.springer.com/article/10.1007/s00703-003-0070-7 (accessed 21-7-2025).
[27] Chang JC & Hanna SR. *Technical descriptions and user's guide for the BOOT statistical model evaluation software package, version 2.0*. 2005. Available from https://www.harmo.org/Kit/Download/BOOT_UG.pdf (accessed 30-12-2024).
[28] Dennis, R., Fox, T., Fuentes, M. *et al.* A framework for evaluating regional-scale numerical photochemical modeling systems. *Environ Fluid Mech* **10**, 471–489 (2010). https://doi.org/10.1007/s10652-009-9163-2
[29] Thunis P, Pederzoli A & Pernigotti D. Performance criteria to evaluate air quality modeling applications, *Atmospheric Environment* **59**, 476-482 (2012). Available from DOI: https://doi.org/10.1016/j.atmosenv.2012.05.043 (accessed 19-8-2025).
[30] Borrego C, Monteiro A, Ferreira J, Miranda AI, Costa AM, Carvalho AC & Lopes M, Procedures for estimation of modelling uncertainty in air quality assessment. *Environment International* **34**, 613-620 (2008). Available from https://doi.org/10.1016/j.envint.2007.12.005 (accessed 19-8-2025).

[31] Olesen HR & Chang JC. Consolidating tools for model evaluation, *International Journal of Environment and Pollution* ***40*** *(1-3)*, 175-183 (2010). Available from https://www.harmo.org/Docs/OlesenChang.pdf (accessed 18-12-2024).
[32] Monteiro A, Durka P, Flandorfer C *et al*. Strengths and weaknesses of the FAIRMODE benchmarking methodology for the evaluation of air quality models. *Air Qual Atmos Health* **11**, 373–383 (2018). https://doi.org/10.1007/s11869-018-0554-8 (accessed 18-12-2024)
[33] Janssen S & Thunis P. FAIRMODE Guidance Document on Modelling Quality Objectives and Benchmarking. Version 3.3. European Commission Joint Research Centre. 2023. doi:10.2760/41988. Available from: https://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/Guidance_MQO_Bench_vs3.3_20220519.pdf (accessed 18-12-2024).
[34] Website: FAIRMODE - Forum for Air quality Modeling (accessed 30-12-2024).
[35] Gryning SE, Batchvarova E, Brümmer B, Jørgensen H & Søren Larsen S. On the extension of the wind profile over homogeneous terrain beyond the surface boundary layer. *Boundary-Layer Meteorol* **124**, 251–268 (2007). Available from https://doi.org/10.1007/s10546-007-9166-9 (accessed 21-3-2025).
[36] Overview of tools and methods for meteorological and air pollution mesoscale model evaluation and user training. Joint Report of COST Action 728 (Enhancing Mesoscale Meteorological Modelling Capabilities for Air Pollution and Dispersion Applications) and GURME (GAW Urban Research Meteorology and Environment Project). Schlünzen KH & Sokhi RS (eds.). WMO/TD-No. 1457. 2008.
[37] COST ES1006 – Model Evaluation Case Studies. COST Action ES1006. April 2015.
[38] Schatzmann M, Olesen H & Franke J. COST 732 Model evaluation case studies: approach and results. 2010. ISBN: 3-00-018312-4.
[39] Hanna S & Chang J. Acceptance criteria for urban dispersion model evaluation. *Meteorology and Atmospheric Physics* **116**, 133–146 (2012). Available from https://doi.org/10.1007/s00703-011-0177-1 (accessed 12-3-2025).
[40] Hanna S & Chang J. Setting Acceptance Criteria for Air Quality Models. In: Steyn, D., Trini Castelli, S. (eds) *Air Pollution Modeling and its Application* **XXI**. NATO Science for Peace and Security Series C: Environmental Security. Available from https://doi.org/10.1007/978-94-007-1359-8_80 (accessed 23-12-2025).
[41] Theobald MR, Løfstrøm P, Walker J, Andersen HV, Pedersen P, Vallejo A & Sutton MA. An intercomparison of models used to simulate the short-range atmospheric dispersion of agricultural ammonia emissions, *Environmental Modelling & Software* **37**, 90-102 (2012). Available from https://doi.org/10.1016/j.envsoft.2012.03.005 (accessed 29-1-2025).

[42] Heist D, Isakov V, Perry S, Snyder M, Venkatram A, Hood C, Stocker J, Carruthers D, Arunachalam S & Owen RC. Estimating near-road pollutant dispersion: A model inter-comparison. *Transportation Research Part D: Transport and Environment* **25**, 93-105 (2013). Available from https://doi.org/10.1016/j.trd.2013.09.003 (accessed 22-7-2025).
[43] Walker JT, Robarge WP & Austin R. Modeling of ammonia dry deposition to a pocosin landscape downwind of a large poultry facility. *Agriculture, Ecosystems and Environment* **185**, 161-175 (2014). Available from https://doi.org/10.1016/j.agee.2013.10.029 (accessed 22-7-2025).
[44] Stocker J, Ellis A, Smith S, Carruthers D, Venkatram A, Dale W. & Attree M. A review of dispersion modelling of agricultural emissions with non-point sources. *International Journal of Environment and Pollution* **62**, 247-263 (2017). Available from 10.1504/IJEP.2017.089410 (accessed 22-7-2025).
[45] SCAIL Agriculture Update - Sniffer ER26 Final Report. Available from: https://www.scail.ceh.ac.uk/agriculture/Sniffer%20ER26_SCAIL-Agriculture%20Final%20report_Issue_11032014.pdf (accessed 4-2-2025).
[46] Website: SCAIL - Simple Calculation of Atmospheric Impact Limits (accessed 4-2-2025).
[47] Patiño WR & Duong VM. Intercomparison of Gaussian Plume Dispersion Models Applied to Sulfur Dioxide Emissions from a Stationary Source in the Suburban Area of Prague, Czech Republic. *Environmental Modeling and Assessment* **27**, 119-137 (2022). Available from https://doi.org/10.1007/s10666-021-09803-4 (accessed 22-7-2025).
[48] Mazzola T, Hanna S, Chang J, Bradley S, Meris R, Simpson S, Miner S, Gant S, Weil J, Harper M, Nikmo J, Kukkonen J, Lacome JM, Nibart M, Björnham O, Khajehnajafi S, Habib K, Armand P, Bauer T & Ek N. Results of comparisons of the predictions of 17 dense gas dispersion models with observations from the Jack Rabbit II chlorine field experiment. *Atmospheric Environment* 244, 117887 (2021). Available from https://doi.org/10.1016/j.atmosenv.2020.117887 (accessed 19-8-2025).
[49] Lorber, M., Eschenroeder, A., Robinson, R., Testing the USA EPA's ISCST-Version 3 model on dioxins: a comparison of predicted and observed air and soil concentrations, Atmospheric Environment 34:23, 3995-4010 (2000). Available from https://doi.org/10.1016/S1352-2310(00)00163-1 (accessed 4-2-2025).
[50] PRCI. Alternate East Fence NO2 monitor data adjustments technical memo. Catalog No. PR-312-15201-R01. February 8, 2018. Document not consulted in this study.
[51] Wesseling J en Van Velze K. Technische beschrijving van standaardrekenmethode 2 (SRM-2) voor luchtkwaliteitsberekeningen (in Dutch). RIVM Briefrapport 2014-0109. 2014. Available from https://www.rivm.nl/bibliotheek/rapporten/2014-0109.pdf (accessed 9-10-2025).

[52] Van den Hout KD en Baars HP. Ontwikkeling van twee modellen voor de verspreiding van luchtverontreiniging door verkeer: het TNO-Verkeersmodel en het CAR Model (in Dutch). TNO rapport R88/192. 1988.
[53] Sauter F, Siteur K, Sterk M, Van der Swaluw E, Wichink Kruit R and De Vries W. The OPS-model, Description of OPS 5.3.1.0. RIVM. 2025. Available from https://www.rivm.nl/sites/default/files/2025-05/ops_documentation_5.3.1.0.pdf (accessed 9-10-2025).
[54] Kooi E, Van Loon R, Sterk M and Jonkers S. NNM Gebouwmodule - Beschrijving en analyse van de NNM gebouwmodule ten behoeve van de implementatie in OPS-ST (in Dutch). Report 2024-0031. RIVM. 2024. Available from NNM Gebouwmodule - Beschrijving en analyse van de NNM gebouwmodule ten behoeve van de implementatie in OPS-ST | RIVM (accessed 7-11-2025).

12 Appendix 1 Measurement campaigns for model validation

In this appendix, we describe how we searched for measurement data to validate models and which measurement data we eventually selected. Finding useful deposition measurements was considered to be most important, but this turned out to be difficult.

12.1 Introduction

The aim of this work was to find datasets that could be used for validating deposition outcomes at local scale. Local scale is loosely interpreted as 'observed concentrations and deposition rates predominantly stemming from a known source at a short distance (typically less than 1 km away)'.

A combined internet and literature search was used to find measurement data that could be used for model validation. In addition, partners in the project were asked which datasets they would recommend. The measurements thus found were compared with sets of essential features and strongly desired features for model validation. Since the search for suited deposition measurement campaign only resulted in the selection of one campaign (section 12.3), additional efforts were made to find useful concentration measurement campaigns (section 12.4).

Regional measurements, at locations where air concentrations do not relate to a single dominant source, were not considered for the following reasons:

- the goal of this work was to determine the accuracy of outcomes for a single source, as used, for example, for permit application;
- the spatial scope of this research was defined in the SAGEN programme plan [1] as ranging 'from a hundred metres to a few kilometres' from the (dominant) source.

12.2 Literature search

A search in Scopus was carried out to find scientific articles about deposition measurements near sources. The search queries used are shown in Table 12.1. Combined, these queries yielded 153 references (the three queries have remarkable overlaps in terms of results). For each reference, the title and abstract were used to identify whether useful information could be found in the article. The more promising papers were downloaded and scanned for relevant information.

Table 12.1 Scopus search queries for identifying validation studies for deposition

Query	# ref.
TITLE-ABS-KEY ((nitrogen OR ammoni*) AND (deposition) AND (measure* OR validat* OR observ*) AND (model*) AND (local*)) AND (LIMIT-TO (SUBJAREA , "ENVI") OR LIMIT-TO (SUBJAREA , "AGRI") OR LIMIT-TO (SUBJAREA , "MULT")) AND (LIMIT-TO (EXACTKEYWORD , "Nitrogen") OR LIMIT-TO (EXACTKEYWORD , "Ammonia")) AND (LIMIT-TO (	150

Query	# ref.
LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "German"))	
TITLE-ABS-KEY ((air OR atmospher*) AND (quality OR pollution) AND model* AND observ* AND (evaluat* OR valida*) AND (performance) AND (gauss* OR plume) AND dispersion)	87
TITLE-ABS-KEY (((air OR atmospher*) AND (quality OR pollution) AND model* AND observ* AND (evaluat* OR valida*) AND (perform*) AND (gauss* OR plume) AND dispersion))	125

12.3 Campaigns with deposition measurements

A general observation from the analysis of literature is that most deposition measurement campaigns are carried out in nature areas, and not particularly close to a single dedicated source. Reasons why deposition is rarely measured in the vicinity of a dominant emission source could be:

- Deposition fluxes depend on concentrations in ambient air, micrometeorological transport conditions, and interactions with plants. The exact source (origin) of concentrations is not relevant for deposition.
- Many campaigns try to measure fluxes to and from plants in order to identify daily and seasonal patterns for a specific plant species and correlations between deposition rates and ambient concentrations. In these studies, the emission sources are not described, or only in little detail.
- Deposition flux measurements (using the gradient method or the eddy covariance method, see section 12.3.1) are most accurate if concentrations vary little in horizontal direction and in time. Close to a source, horizontal variation of concentrations can cause advection errors that make observations less accurate. For this reason, deposition fluxes are normally measured away from strong sources.

12.3.1 Deposition measurement techniques

Multiple techniques exist for measuring the deposition of reactive nitrogen. The analysis of literature highlighted some limitations of different techniques, either for measuring reactive nitrogen deposition in general, or specifically for measuring it close to a source. These limitations are listed below. The list is not meant to be exhaustive.

1. Bulk samplers: Bulk samplers are open samplers that collect total deposition: wet deposition during precipitation events and dry deposition during the remainder of the time. For vegetation, it is known that specific plant characteristics are important because of the way they absorb, adsorb, and re-emit reactive nitrogen components. Bulk samplers cannot mimic the specific behaviour of vegetation and the measured deposition is not representative of vegetation.
2. Wet-only samplers: Wet-only samplers are samplers that open during precipitation events and then collect the precipitation and the corresponding wet deposition. They do not measure dry

deposition, which, therefore, needs to be measured with a different technique.

3. Throughfall methods: Throughfall methods collect rain water that falls through the canopy of a forest and measure the mass of different components (e.g. reactive nitrogen) in that water. By comparing it with rain water collected in an open field or above the canopy, the mass of components that is 'washed' from the leaves in the canopy can be calculated. It does not measure the mass absorbed by the plant including mass used for plant growth, nor does it measure any mass re-emitted to air during dry periods. Last, the method can only be used for vegetation from a specific minimum height. Overall, the relevance of these measurements is very limited.
4. Concentration gradient measurements: These measurements measure air concentrations at two or more heights and then calculate the vertical concentration gradient. Advantages are that the temporal resolution of such measurements is normally very high and that these measurements can be carried out above different types of land cover (e.g. grassland, crops, moors, heather, or forest). An important limitation is that this method cannot distinguish between different causes for vertical concentration gradients. Dry deposition is one cause. Close to sources, mere mixing of air concentrations can be another cause: as long as the concentration at plume centreline height exceeds concentrations above and below that height, mixing will result in a downward flux below the centreline and an upward flux above the centreline. Close to a source, concentrations can also fluctuate significantly in time due to changing weather conditions. This also limits the reliability of the measurement, if the concentration measurements at different heights are not carried out simultaneously.
5. Eddy covariance techniques: eddy covariance techniques measure mass fluxes at one specific height. Similar to the concentration gradient measurements (see before), it cannot discriminate between different contributions to the total flux. Close to a source, vertical fluxes due to mixing (the upward and downward transportation of mass away from centreline height) are non-negligible.
6. Flux chambers: flux chambers are cylinders that are placed directly on soil or vegetation. The top of the cylinder can be opened or closed in alternating sequences. When the lid is closed, changes in concentrations reflect either emissions by vegetation (increasing concentrations) or absorption and adsorption (decreasing concentrations). A limitation of this technique is that turbulence inside the closed compartments is much reduced compared with turbulence in an open field. Therefore, the method is not reliable if turbulence is important for transporting mass towards the plant (deposition) or away from the plant (re-emission). The latter is true if exchange rates with plants are high, as is expected for reactive nitrogen components (particularly ammonia).
7. Biomonitors: biomonitors, also referred to as bio-indicators, are monitored plants that grow while absorbing nutrients. The effective uptake of reactive nitrogen can be derived by comparing

the amount of nitrogen in the plant at the start and at the end of the measurement period. Plants can be put in pots to obstruct the exchange of reactive nitrogen with soil and ground water. The measured deposition is combined wet and dry deposition over the full period. For practical purposes, plants usually start as seedlings. It is, then, uncertain whether the measured deposition is also representative of more mature plants. In addition, measuring nitrogen content of plant and soil appears to come with measurement errors. Last, animals eating from the plants can disturb the measurement. For the two biomonitoring experiments that were explored in this study ([6], [10]), the accuracy was not as good as had been hoped for.

In some studies, the deposition rate is not measured, but derived from observed concentrations at one specific height (also referred to as 'inferential methods') and calculated or predefined deposition velocities. These studies have not been further investigated, because they did not measure deposition itself.

12.3.2 Criteria for selecting datasets

Criteria were defined to differentiate between more useful and less useful datasets. An important criterium was that we wanted to be able to validate OPS-LT (the more important tool for deposition modelling in the Netherlands from a legislative perspective). OPS-LT uses a statistical approach to calculating annual concentrations and deposition values. The minimum time duration to get reasonably accurate outcomes in all wind directions is assumed to be somewhere around one month.

Essential features:

- The main features of the campaign have been published in a report or scientific paper. That report or paper should also address which measurement techniques were used and provide information on the quality/accuracy/reliability of the observations.
- Observations are carried out close to a known source. Observed concentrations and deposition rates are largely linked to that source. The source should be real or representative of real scale. Emissions rates from that source have been quantified or estimated and a description of the source rate estimation is available.
- The measured component is a reactive nitrogen compound (NH_x or NO_y)
- Either total deposition was measured or dry and wet deposition were measured separately. Dry deposition should be related to plants (vegetation) and should be measured with sufficient accuracy, because close to a source, dry deposition is believed to be more important than wet deposition.
- The duration of the campaign should be four weeks or more.
- Measurement data is publicly available or can be made publicly available.

Strongly desired features:

- Spatial resolution: it is preferred that deposition is measured at multiple locations, to such an extent that gradients in horizontal

(radial) profiles can be observed. The most interesting distance range is 50 to 500 m.
- Concentration is measured along with the deposition.
- Ideally, the error in observations has been quantified or estimated and is small for most observations. Next best but also acceptable: the error has been quantified or estimated but is significant.
- Availability of meteorological observations: Ideally, meteorological conditions were measured at the test location. If not, data from a nearby station representative of the test site, should be available.
- Observations have already been used for validation of some air transport model(s): it has been demonstrated that the data can indeed be used for validating models.
- Background concentrations are known.

Weakly desired features:
- Temporal resolution: The temporal resolution (sampling frequency) is less important. The research programme is aimed at deposition on nature areas. Deposition rates do not need to be specified per hour, week, or even month. Please note that some models do not (and cannot) report hourly values.

The criteria were applied to deposition measurements near livestock farms and fields with manure application (section 12.3.3), roads (section 12.3.4) and industrial sources (section 12.3.5).

12.3.3 Campaigns relating to emissions from livestock farms and fields with manure application

More than a dozen campaigns were found where deposition was measured near a livestock farm or in or around a field with manure application. Quite a few deposition measurements were carried out in order to obtain more accurate estimates of emissions from farms or fields. Most campaigns either had a very short time span (hours or days) and/or a suboptimal spatial coverage (inside a field or only very close to the source (e.g. [13]). Other campaigns focused on deposition measurement techniques and did not measure or estimate source characteristics (e.g. [7], [8]). Deposition measurements by the University of Amsterdam around a dairy farm in the Netherlands in 2022-2023 [10] were not used because the outcomes had not yet been reported in literature and were not yet available in the public domain.

The most relevant deposition measurement campaigns near sources are discussed in more detail in the following subsections. Out of all these campaigns, only one campaign fulfilled the essential criteria of section 12.3.2: the measurement campaign in 2005 around a poultry farm in Ringsted, Denmark. The details of this campaign were reported in Sommer et al. [6]. Deposition was measured with biomonitors. This campaign also fulfilled many strongly or weakly desired features for campaign selection (see section 12.3.2). Both measurement data and input data for modelling was available at Aarhus University.

12.3.3.1 Ringsted chicken farm campaign (2005)

The Ringsted campaign involves concentration and deposition measurements around a chicken farm in Ringsted, Denmark. The measurements were carried out in 2005 with the aim of validating the dispersion and deposition modelling in OML. Details of the campaign and the comparison with model outcomes have been described in a paper by Sommer et al. [6].

Emissions from two buildings with chickens were measured with sensors in the inlets and outlets for mechanical ventilation of the buildings. Ambient concentrations around the farm were measured with passive samplers, while deposition around the farm was measured with biomonitors, using rye grass in triplets of pots at 25 locations around the source. The triple measurements per location allowed for assessing the precision of the deposition measurements. The total duration of the measurements was about eight weeks. As stated above, the measurement data had already been used to validate OML [6]. Aarhus University was able to share the experimental data and assist with the interpretation of the data.

The measurement campaign fulfilled all six requirements for selecting campaigns. The accuracy of the deposition measurements was, however, lower than had been hoped for (see also section 3.1.4.2):

- The relative standard error in measured deposition fluxes was significant.
- While measured concentrations mostly decreased smoothly with distance, measured deposition fluxes did not.
- Several outcomes for deposition were smaller than zero.
- Measurements south of the farm were affected by a non-specified local source.

More details about the campaign, and the accuracy of the measurements, are provided in Chapter 3.

12.3.3.2 Denmark dairy farm deposition measurements (1988)

This campaign involves concentration and deposition measurements around a dairy farm in Denmark. The campaign was carried out in 1986 and details have been reported in a short paper by Sommer [7]. The campaign appears to have been a proof of principle that deposition fluxes around a dairy farm can be measured with biomonitors.

Deposition was measured with plants of barley in groups of three pots at 13 separate locations around the dairy farm. The duration of the deposition measurements was approximately one month. The paper by Sommer [7] shows how deposition rates vary with distance.

The campaign was not selected for the validation study for the following reasons:

- The campaign was carried out long time ago (in 1986) and was described in little detail (4 pages). It is likely that at least some required input data for model calculations is missing. It is also unknown whether all measurement data is still available.
- Emissions for the animal houses certainly have not been measured. For this reason, Sommer decided to organise a new

measurement campaign (Ringsted), when asked to validate OML [6].

12.3.3.3 Denmark dairy farm campaign (1990)

This campaign also involves concentration and deposition measurements around a dairy farm in Denmark. The campaign was most probably carried out in 1990. Details have been reported in a short paper by Sommer & Jensen [8]. The aim of the campaign was to 'simultaneously measure atmospheric NH_3 concentrations and foliar absorption of NH_3 by plants in the field'.

Deposition was measured with plants of rye grass in groups of four or five pots at nine separate locations around the dairy farm. In addition, bulk deposition was measured with bulk samplers. The duration of the deposition measurements was approximately six weeks.

The reasons for not selecting this campaign are identical to the reasons for not selecting the related 1988 campaign (section 12.3.3.2):

- The campaign was carried out a long time ago (most likely in 1990) and was described in little detail (4 pages). It is likely that at least some required input data for model calculations is missing. It is also unknown whether all measurement data is still available.
- Emissions for the animal houses have certainly not been measured. For this reason, Sommer decided to organise a new measurement campaign (Ringsted), when asked to validate OML [6].

12.3.3.4 North Carolina pig farm campaign (2001-2002)

This campaign involves concentration and deposition measurements (although indirectly) around a research pig farm in North Carolina, USA. The campaign was carried out in 2001 and 2002 and has been described in a paper by Philips et al. [9]. The paper focuses on seasonal patterns in concentration and deposition.

The site consists of seven barns housing swines and a lagoon containing 'swine waste' (possibly manure). Deposition fluxes to the ground were derived from concentration measurements at 2 heights (2 m and 6 m) and at a distance of approximately 800 m from the farm. The measurements were carried out for four 4-week periods, spread over four seasons.

The campaign was not selected for the validation study for the following reasons:

- There is no reference to source characteristics or source strength in the paper by Philips et al. There is no indication that emissions from the pig houses and the lagoon have been measured.
- Deposition flux was only measured at one distance, and quite far from the source (800 m).
- Deposition fluxes were not measured directly but calculated from measured gradients and calculated micrometeorological functions. At this distance from the source, it is unknown whether the calculation is accurate. According to Philips et al.,

the relative error in the resulting fluxes (1 standard deviation) was roughly 100% (see Table 1b in [9]).

12.3.3.5 Netherlands dairy farms campaign (2020-2022)

This campaign involves concentration and deposition measurements around two dairy farms in the Netherlands and was carried out from 2020 to 2022 [10]. The main goal of the campaign was to analyse spatiotemporal patterns of reactive nitrogen deposition around dairy stables. Some measurement data has been compared with model outcomes for OPS-LT. Details on the campaign are provided in a report by Tietema et al. [10].

Deposition around the farms was measured in various ways; using bulk samplers, flux chambers, and two types of biomonitors. Isotope measurements were carried out to determine whether nitrogen in local biomass originated from natural sources or human activities (traffic and industry). Both the bulk samplers and the biomonitors were placed at 24 separate locations, with a maximum distance to the source of roughly 500 m. The flux chamber was placed at one single location.

Measurements have been carried out around two farms, but neither farm is described in detail. Most of the discussion appears to focus on one farm (the text is not always coherent). Emissions of ammonia originate from a partly open dairy stable and from production grassland that had been fertilised with manure and artificial fertiliser during the period and is normally mowed several times per year. Emission strengths were estimated using a webtool. Further details are not reported.

The campaign was not selected for the validation study for the following reasons:

- There was debate about the reliability of the measurement data. Deposition fluxes derived with biomonitors were much higher than had been expected. It was uncertain whether the 'upscaling assumptions' were correct [10]. Bulk samplers measure adsorption of ammonia, but not absorption (by plants) or re-emission (by plants). As such, the relevance of bulk samplers to study dry deposition for vegetation appears be very limited.
- Little information is available on the sources of emissions (stables and production grassland). Emission strengths and other source characteristics have not been reported. The accuracy of the input data and output data for calculating the emission strength is unknown.

12.3.3.6 GRAMINAE experiment

The GRAMINAE experiment ([11], [12]) in Braunschweig, Germany, was designed to quantify ammonia fluxes over intensively managed grassland near farm buildings housing cattle and pigs. Vertical fluxes were derived from concentration measurements from multiple heights at two locations, using multiple independent techniques. The distance to the farm buildings was about 600 m and 800 m. The magnitude of advection error could be derived from the measurements at the two locations [13]. The measurement campaign took place between 22 May and 15 June 2000, with a total duration of 25 days. The grassland was

initially covered with tall grass, which was mowed after 7 days and again 7 days later was fertilised with calcium ammonium nitrate. During the first seven days (before mowing), average fluxes were negative, indicating net ammonia deposition. After mowing, average fluxes changed sign, indicating net ammonia emissions. These net emissions increased significantly after applying the fertiliser to the field. The observed fluxes were characterised as small bi-directional fluxes prior to mowing the grass, larger diurnally varying emissions after cutting, and much larger emissions following the application of fertiliser [12].

The campaign was not selected for the validation study for the following reasons:

- The campaign duration was under four weeks.
- Measured fluxes were positive (indicating emissions) for a significant number of hours. The scope of the current study was to validate deposition processes away from the source, not emission processes at the source.

12.3.4 Campaigns relating to emissions from road traffic

Measuring the contribution from a specific road to total deposition is a challenge, because of the lower deposition velocity of NO_x compared with NH_3. Two campaigns ([14], [15]) were found that tried to link measured deposition to a specific nearby road motorway. In both campaigns, deposition was measured with bulk samplers. Dry deposition was not measured. As such, the measurements did not fulfil the essential features for selecting validation campaigns. Both campaigns are discussed in further detail below.

12.3.4.1 Bavaria motorway campaign (2002-2003)

This campaign involves concentration and deposition measurements along two transects perpendicular to a motorway in Bavaria, Germany. The aim of the study was to estimate the influence of vehicle emissions on the input of reactive nitrogen to ecosystems. The campaign had a duration of one year and was carried out in 2001-2002. Details on the campaign were reported in a paper by Kirchner et al. [14].

Measurements were carried out in a coniferous forest and over extensively farmed grass land. Bulk samplers measured wet deposition to the grassland, while deposition in the forest was measured with the throughfall method. At both sites, deposition was measured at 4 separate locations with a maximum distance of 400 (grassland) to 500 m (forest) to the motorway. Dry deposition was not measured but estimated, assuming fixed deposition velocities depending only on vegetation type. The variability of plant vegetation along the transect in the forest was also investigated.

The campaign was not selected for the validation study for the following reasons:

- Dry deposition was not measured but calculated, using a fairly simple method. Accurate measurements of dry deposition are important for model validation because the models used in this study predict that the contribution of dry deposition to total

deposition is much larger than the contribution of wet deposition.[42]
- Measured outcomes have not been compared with model outcomes. It is, therefore, uncertain whether all required model input data is available. In particular, traffic intensity and corresponding vehicle emissions were not reported in [14] and appear not to have been investigated.

12.3.4.2 Massachusetts moderately trafficked road campaign (2004-2006)
This campaign involves deposition measurements near two moderately trafficked roads in Massachusetts, USA. The campaign was run between 2004 and 2006 and is described in a paper by Bettez & Howard [15]. At both sites, deposition was measured with bulk samplers along two transects perpendicular to the road: one transect in forest measured throughfall and another transect in open field measured bulk deposition. Samplers were placed at three or four locations along each transect, with a maximum distance to the road of 300 m. The contents of the samplers were collected after each precipitation event during the measurements (three summers and one full year) and then analysed.

The campaign was not selected for the validation study for the following reasons:
- Dry deposition has not been measured.
- Ambient concentrations have not been measured.
- Measured outcomes have not been compared with model outcomes. It is, therefore, uncertain whether all required model input data is available. Average traffic intensities on the two roads are provided in [15], but further details (e.g. vehicle mix, average speed, expected emission factors) are missing.

12.3.5 Campaigns relating to emissions from industries
No campaigns were found that tried to link measured NO_x and/or NH_3 deposition to a specific industrial source. In some cases, soil samples have been taken and investigated for contaminants (e.g. polycyclic aromatic hydrocarbons, dioxins, etc.). The measurements of dioxins and furans around the Columbus Municipal Solid Waste-to-Energy Facility in Columbus, Ohio (USA) [49], may be a good example of such measurements. It is, however, difficult to use such measurements for model validation: the exposure duration is often long, emissions and source characteristics are normally not known in great detail, contaminants may have washed away, vegetation may have been removed, etcetera.

12.4 Campaigns for validating concentrations

The aim of this work was to find datasets that could be used for validating concentration outcomes at local scale. Local scale is, again, loosely interpreted as 'observed concentrations and deposition rates predominantly stemming from a known source at a short distance (typically less than 1 km away)'.

[42] This was investigated for the motorway case in the model intercomparison study [2]. The dry deposition ratio at distances ranging from 50 m to 5 km from the source was close to 100% for all models except STACKS-D, regardless of the vegetation type (grassland, forest, or mixed). STACKS-D calculated wet deposition ratios between 5% and 25% (depending on distance and vegetation type).

Datasets were found by studying dedicated websites for validating air quality models, by studying literature and by consulting the project partners. Out of the websites used, those from the EPA [17], HARMO [22], and ADMLC [23] were considered to be most useful. All partners had already used at least some campaigns to validate their models. Out of the models in this project, AERMOD and ADMS had been validated against the largest number of datasets.

12.4.1 Criteria for selecting datasets

Criteria were defined to differentiate between more useful and less useful datasets. To a large extent, these were identical to the criteria that were used to select deposition measurement campaigns (see section 12.3.2). Moreover, it was specified that concentration measurements were obtained over a terrain that was not hilly or mountainous (essential feature) and that background concentrations were specified when relevant (strongly desired feature).

The criteria were applied to datasets near an industrial stack (section 12.4.2), a livestock farm (section 12.4.3), and road traffic (section 12.4.4).

12.4.2 Campaigns relating to emissions from industrial stacks

Several websites provide good overviews of measurement campaigns that can be used to validate air quality models. In particular, the EPA [17], HARMO [22] and ADMLC [23] websites were (mostly) up-to-date and useful for gathering information and data.

At least thirty experiments and field campaigns were investigated for relevance. Some of these were quite old (e.g. the Prairie Grass experiment in 1956) while others were relatively new (e.g. the Balko campaign in 2023). A large number involved tracer experiments from (mostly) high stacks, often using SF_6 as a tracer. These tracer experiments were considered to be less relevant, because emission durations are too short and emissions of reactive nitrogen compounds were prioritised.

The above websites also provided information on at least thirteen field campaigns. In these campaigns, emissions from industrial sources were monitored for a long time, typically ranging from several weeks to approximately one year. In the early 1980s, several campaigns had been run to measure SO_2 concentrations around power plants. Newer datasets relate to other types of sources and substances. Terrain ranges from very smooth to hilly or even mountainous.

Using the selection criteria from section 12.4.1, and particularly focusing on a sufficiently long duration, good spatial coverage, and relevant terrain (not hilly or mountainous), three campaigns were considered to be most useful: SO_2 measurements around the Kincaid power plant (1980-1981), SO_2 measurements around the Baldwin power plant (1982-1983) and NO_2 and NO_x measurements around a compressor station in Balko, Oklahoma (2023). Out of these campaigns, the data for Balko was regarded as most useful, for the following reasons:

- The campaign is relatively new and measurements are expected to be of better quality than those for Kincaid and Baldwin.

- NO_2 and NO_x were measured (not SO_2).
- The site has four sources and several buildings. The source conditions varied over time, making it a particular good case for studying model performance for real, complex, cases.

A limitation of the Balko campaign is the limited number of measurement stations (only 4) and the relatively short distances from sources to measurement stations (maximum 400 m).

12.4.3 Campaigns relating to emissions from livestock farms

About fifteen campaigns were found in which NH_3 or PM_{10} concentrations were measured around livestock farms. The following publications and reports were particularly useful.

- A paper by Sommer et al. [6] describes concentration and deposition measurements around a poultry farm in Ringsted, Denmark (see section 12.3.3). This campaign had been used to validate OML.
- A paper by Theobald et al. [41] discusses two measurement campaigns that were used to validate ADMS, AERMOD, LADD, and OPS-ST. The campaigns were run around a pig farm in Falster, Denmark (2006) and around a pig farm in Green County, North Carolina (2003-2005). The latter had a very good campaign duration (about two years). Later on, these measurements have also been used to validate STACKS-D.
- The Atmospheric Dispersion Modelling Liaison Committee (ADMLC) in the UK commissioned a review of the limitations and uncertainties of modelling pollutant dispersion from non-point sources. The work focused on agricultural and bioaerosol sources and results were published by Stocker et al. in 2016 [44]. A detailed investigation was carried out to find measurement data. Obtaining the datasets was often difficult. Eventually three datasets (Farm F, Farm G, and Site B) were obtained and used to compare several modelling options. These three datasets are available on the ADMLC website [44]. However, the measurement durations are short (44 hours, 48 hours, and 6 separate days).
- The Sniffer E26 final report [45] describes datasets that have been used to validate SCAIL-Agriculture. SCAIL-Agriculture is a UK screening tool for assessing the impact from pig and poultry farms on human health and on semi-natural areas, in particular whether impact limits for human health or habitats are exceeded or not [46]. The report describes at least 25 datasets that were potentially useful for validating SCAIL-Agriculture. Seven of these datasets were subsequently selected and used: six datasets from the UK and the Republic of Ireland, and one campaign in Falster, Denmark (see above). The ADMLC review for non-point sources highlighted difficulties in obtaining these datasets.

Eventually, the dataset for the Ringsted poultry farm was selected to validate the models, primarily because it had both deposition and concentration measurements. The spatial and temporal coverage of that campaign was also adequate.

12.4.4 Campaigns relating to emissions from road traffic

A small number of datasets for emissions from motorways was investigated. The most relevant campaigns were:

- The Life+ ATMOSYS highway campaign near Affligem, Belgium [16]. In this campaign, run from 2012 to 2013, multiple components (PM_{10}, $PM_{2.}5$, black carbon, NO_2, NH_3 and VOC) were measured at three to six distances from the motorway. At least NO_2 was measured during 36 consecutive weeks.
- Measurements near Svogerslev, Denmark in 2016 [21]. In this campaign, NO_2 was measured with passive samplers at ten separate locations during a period of six weeks. The largest distance to the motorway was 420 m. This campaign had already been used to validate OML-Highway (a different but related OML version, specific for motorway traffic).
- Measurements near Koge Bugt, Denmark in 2003 [21]. In this campaign, NO_x, NO and NO_2 were measured during three months in transects perpendicular to the motorway. The largest distance to the motorway was about 100 m. This campaign had already been used to validate OML-Highway.

The Life+ ATMOSYS campaign was selected because the campaign was well described in a detailed research report, because of the large number of available measurements, because of the adequate spatial and temporal coverage, and because the data could easily be obtained from the Flemish Environmental Agency (VMM).

12.4.5 Campaigns relating to emissions from fields with manure application

These campaigns were not investigated in much detail, because time was too limited to add more cases to the validation study.

13 Appendix 2 Detailed results for Ringsted measurements

13.1 Detailed comparison with measured concentrations

This appendix provides more detailed information on the Ringsted campaign, including the measured concentrations and deposition fluxes, and detailed individual model results. Aarhus University provided RIVM with the measurement outcomes and locations. More information on the processing of raw measurement data is presented in section 16.2.1 of Appendix 5.

13.1.1 Overview of measurement outcomes

Measured concentrations for different receptors and different measurement periods are listed in Table 13.1. The Geometric Mean (GM) and Geometric Standard Deviation (GSD) of the corresponding individual model outcomes (all models that provided data for this dataset), are also shown. The start and end times of each measurement period are provided in Table 13.2.

Table 13.1 Measured concentrations and the Geometric Mean (GM) and Geometric Standard Deviation (GSD) of all individual model outcomes

Receptor	Distance to nearest source (m)	Measurement period	Average observation (µg/m³)	GM (µg/m³)	GSD (-)
rcp_132_NVP	119	1	3.07	3.02	1.3
rcp_132_NVP	119	3	3.67	3.48	1.3
rcp_140_SVP	106	1	6.39	4.29	1.4
rcp_140_SVP	106	3	3.72	3.15	1.3
rcp_143_NV	124	1	2.45	2.59	1.2
rcp_143_NV	124	3	5.5	5.35	1.5
rcp_179_NO	149	1	2.86	3.26	1.3
rcp_179_NO	149	3	2.03	2.08	1.1
rcp_192_NOP	162	1	2.58	3.04	1.3
rcp_192_NOP	162	3	1.98	2.05	1.1
rcp_215_SV	180	3	2.42	2.42	1.2
rcp_228_NV	209	2	2.81	2.54	1.2
rcp_228_NV	209	3	3.88	3.65	1.4
rcp_289_NO	258	2	2.82	2.52	1.1
rcp_289_NO	258	3	1.75	1.90	1.1
rcp_344_SV	310	3	1.78	2.03	1.1
rcp_465_NV	446	2	2.45	1.94	1.1
rcp_465_NV	446	3	3.64	2.30	1.2
rcp_47_NVP	40	1	13.27	7.30	1.9
rcp_47_NVP	40	2.5	16.04	9.61	2.1
rcp_47_NVP	40	3	11.29	5.27	1.7
rcp_54_NOP	26	1	12.21	8.06	2.1
rcp_580_NO	550	2	2.07	1.95	1.1
rcp_580_NO	550	3	1.78	1.77	1.0
rcp_60_SVP	30	1	17.59	8.97	1.9
rcp_60_SVP	30	3	13.79	8.20	1.9
rcp_601_SV	567	3	1.4	1.83	1.0

Table 13.2 Start and end times of the various periods

Measurement period	Start of period	End of period
1	5-9-2005 12:55	15-9-2005 12:45
2	15-9-2005 12:45	29-9-2005 09:35
2.5	15-9-2005 12:45	3-10-2005 09:35
3	3-10-2005 09:35	17-10-2005 12:00

13.1.2 Receptor contributions to total NMSE and total VG

Model performance was measured with several indicators, including Normalised Mean Square Error (NMSE) and Geometric Variance (VG). The calculated outcomes apply to the whole measurement dataset used for validating the models. The indicators do not make clear which receptors have the largest differences between measured and modelled outcomes.

The Normalised Mean Square Error (NMSE) is a dimensionless metric for the average absolute square deviation between model outcomes and measured outcomes:

$$NMSE = \frac{1}{\bar{O} \cdot \bar{M}} \cdot \overline{(O - M)^2} \qquad 13$$

For *N* measurements, the mean square deviation is the sum of N individual measurement contributions $\frac{1}{N} \cdot (O_i - M_i)^2$:

$$\overline{(O - M)^2} = \frac{1}{N} \cdot \sum_{i=1}^{i=N} (O_i - M_i)^2 \qquad 14$$

Then, the contribution of an individual measurement to the total NMSE is equal to:

$$\frac{1}{N} \cdot \frac{(O_i - M_i)^2}{NMSE} \qquad 15$$

In order to limit the amount of information, contributions to receptors for different measurement periods were accumulated. The result is the contribution of each receptor (all time periods included) to the total NMSE. The total NMSE per model was reported in section 3.2.1, Table 3.4.

Table 13.3 shows how much individual receptors contribute to the total NMSE. The table only shows contributions to the total NMSE of at least 5%. The three locations closest to the source (26 m NE, 30 m SW, and 40 m NW) determine the total NMSE to a large extent. Together, these three locations contribute 89% to 97% to the total NMSE. The other thirteen locations together contribute between 3% and 11%.

Table 13.3 Receptor contributions to the total NMSE (cutoff: 5%)

Direction	Distance	ADMS	AERMOD	IFDM	OML	OPS-LT	OPS-ST	STACKS
NE	26		9%	8%	10%	9%	39%	11%
SW	30	42%	48%	45%	36%	40%	12%	40%
NW	40	47%	40%	40%	50%	48%	39%	45%
NW	124						5%	

NW=north-west, SW=south-west, NE=north-east

The Geometric Variance (VG) is a geometric alternative to NMSE. It measures the mean square factor deviation between modelled outcomes and measured outcomes:

$$VG = e^{\overline{(\ln(O)-\ln(M))^2}} = e^{\overline{(\ln(O/M))^2}} \quad 16$$

VG has an exponential form; it multiplies terms from individual measurements. This makes it difficult to identify the contribution of a single measurement to total VG as a percentage. The solution is to take the logarithm of VG and identify the individual contributions to ln(VG).

$$ln(VG) = \overline{\ln(O/M)^2} = \frac{1}{N} \cdot \sum_{i=1}^{i=N} ln\left((\frac{O_i}{M_i})^2\right) \quad 17$$

Then, the contribution of an individual measurement to total ln(VG) is:

$$\frac{1}{N} \cdot \frac{ln\left((\frac{O_i}{M_i})^2\right)}{ln(VG)} \quad 18$$

Contributions for different time periods can be summed per receptor in order to get the total contribution by individual receptors to the logarithm of VG.

The results are shown in Table 13.4, again focusing on receptor contributions of at least 5%. The table shows more diversity than Table 13.3: for some models, nearby receptors are the most important for the total VG, while for other models (IFDM, OPS-ST) distant receptors are more important. The number of receptors that 'provide' at least 5% to total VG is also larger than before for NMSE.

Table 13.4 Receptor contributions to the logarithm of VG (cutoff: 5%)

Direction	Distance	ADMS	AERMOD	IFDM	OML	OPS-LT	OPS-ST	STACKS
NE	26		8%	5%	9%	8%	18%	12%
SW	30	20%	28%	18%	18%	26%	6%	30%
NW	40	37%	36%	25%	49%	44%	16%	44%
SW	106	16%	6%	9%	9%	5%		5%
NW	119						7%	
NW	124						17%	
NE	149			5%			12%	
NE	162			5%			10%	
SW	180			6%				
NW	209						8%	
SW	310			5%				
NW	446	12%	12%	9%	9%	8%		
SW	567			6%				

13.2 Detailed comparison with measured deposition

13.2.1 Overview of measurement outcomes

Measured deposition for different receptors are listed in Table 13.5. This only includes those receptors that were used by Sommer et al. [6] to validate OML (outliers are not included) The Geometric Mean (GM) and Geometric Standard Deviation (GSD) of the corresponding individual model outcomes (all models that provided data for this dataset) are also shown. The deposition measurements started on 1 September 2005 and ended on 25 October 2009. In the model runs, it was assumed that measurements started and ended at noon (12:00).

Table 13.5 Measured deposition and the Geometric Mean (GM) and Geometric Standard Deviation (GSD) of all individual model outcomes

Receptor	**Distance to nearest source (m)**	**Average observation (g N/m^2)**	**GM (g N/m^2)**	**GSD (-)**
rcp_15_80	26	0.091	0.22	1.7
rcp_20_80	28	0.18	0.2	1.6
rcp_40_80	47	0.05	0.17	1.5
rcp_160_80	170	0.13	0.049	1.7
rcp_225_80	230	0.065	0.032	1.8
rcp_60_280	11	0.5	0.13	7.1
rcp_80_280	25	0.026	0.11	6.1
rcp_100_280	42	0.028	0.15	1.5
rcp_120_280	61	0.11	0.11	1.5
rcp_160_280	100	-0.033	0.058	1.5
rcp_320_280	260	-0.025	0.017	1.9
rcp_30_370	41	0.41	0.27	1.7
rcp_40_370	48	0.29	0.2	2.1
rcp_80_370	83	0.19	0.15	1.7
rcp_100_370	100	0.048	0.12	1.7
rcp_160_370	160	0.053	0.065	1.7

13.2.2 Receptor contributions to total NMSE and total VG

Table 13.6 shows the contributions by individual receptors to total NMSE (with a cutoff of 5%), using the method described in section 13.1.2. For each model, several receptors contribute significantly to the total NMSE. The contributions by the two closest measurement locations are particularly relevant.

Table 13.6 Receptor contributions to total NMSE (cutoff: 5%)

Direction	Distance	ADMS	AERMOD	IFDM	OML	OPS-LT	OPS-ST	STACKS
West	11	21%	50%	13%		60%	27%	47%
West	25	24%		26%	22%	6%	21%	
East	26	15%		13%	7%	6%	14%	
East	28	7%						
North	41		24%		27%	7%		32%
West	42	7%		16%	9%		6%	
East	47	8%		10%	5%	6%	11%	
North	48	6%	10%		6%			10%
North	83		5%					
West	101			7%	8%			
North	102						6%	

Table 13.7 shows the contributions by individual receptors to the logarithm of total VG (with a cutoff of 5%). Again, for each model, several receptors contribute significantly to the total outcome.

Table 13.7 Receptor contributions to the logarithm of total VG (cutoff: 5%)

Direction	Distance	ADMS	AERMOD	IFDM	OML	OPS-LT	OPS-ST	STACKS
West	11		18%			14%		49%
West	25	30%	6%	31%	35%	21%	30%	11%
East	26	8%		7%			7%	
North	41		7%					
West	42	18%		24%	22%	17%	17%	
East	47	12%		11%	9%	11%	12%	
North	48		5%					
North	83		7%					
North	102	9%		5%	6%	6%	9%	
East	166		17%			7%	5%	
East	231		14%			5%		

13.3 Combining concentration and deposition outcomes

Models that accurately represent the deposition processes should predict consistent model outcomes for both concentrations and deposition; that is, if the concentration was being under-predicted, the deposition should also be under-predicted, and vice versa.

Calculated biases for modelled concentrations can be compared with calculated biases for modelled deposition. However, the comparison is only fair if concentrations and deposition were measured at the same locations and for the same time periods, or at least to a reasonable

extent. For the Ringsted campaign, this is not the case: concentration and deposition measurements differ in number, locations, and durations. Combined with the limited accuracy of the deposition measurements in particular, it is uncertain whether the comparison of biases can reveal strengths or weaknesses of the modelling of deposition processes. The comparison of Fractional Biases (Figure 13.1) and Geometric Mean Biases (Figure 13.2) is added for the purpose of completeness.

Figure 13.1 Comparison of the Fractional Biases (FB) of the models regarding concentration and deposition measurements

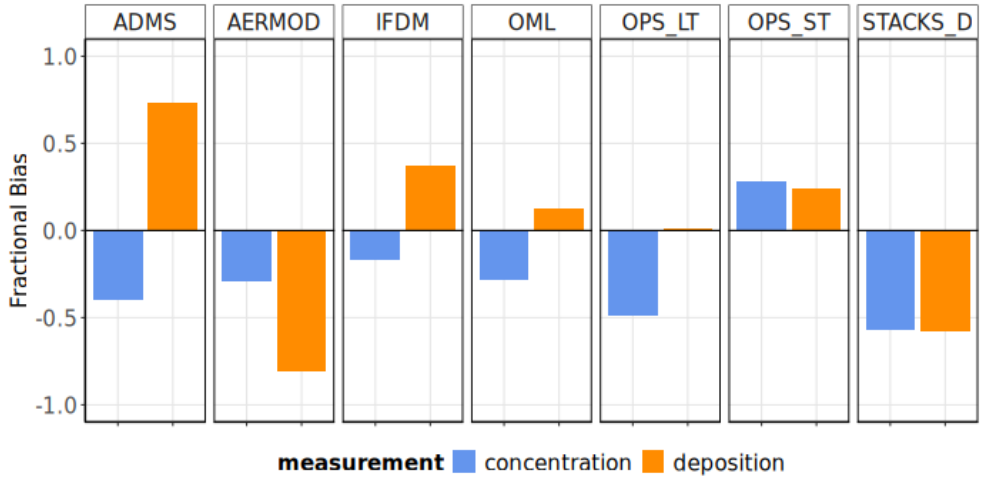
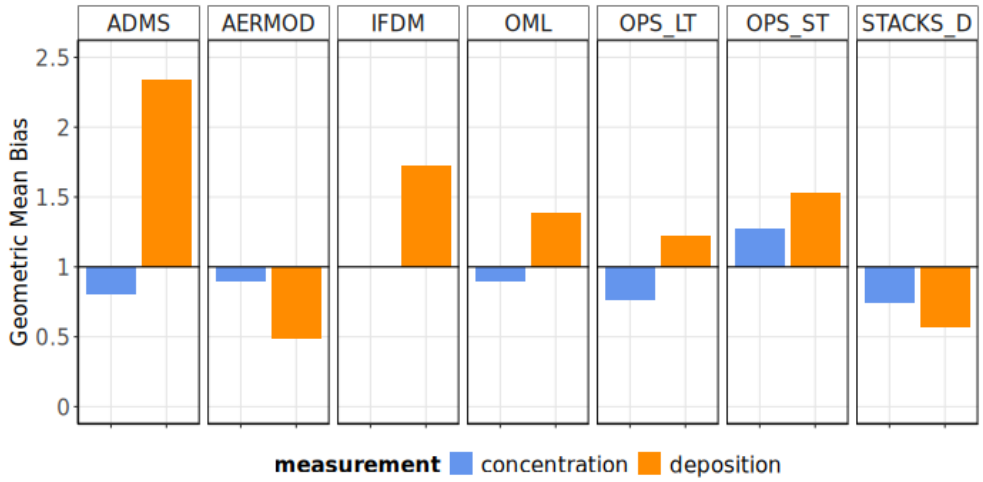


Figure 13.2 Comparison of the Geometric Mean Biases (MG) of the models regarding concentration and deposition measurements



14 Appendix 3 Detailed results for the Affligem campaign

This appendix provides more detailed information on the Affligem campaign, including period-averages for measured NO_2 concentrations and detailed individual model results.

14.1 Overview of period-average measurement outcomes

Period-average concentrations for the various receptors are listed in Table 14.1. The Flemish Environmental Agency (VMM) provided RIVM with the period-average concentrations. Additional information on the processing of these measurement data is provided in section 16.2.2 of Appendix 5. The Geometric Mean (GM) and Geometric Standard Deviation (GSD) of the corresponding individual model outcomes (all models that provided data for this dataset) are also shown in Table 14.1.

Table 14.1 Period-average concentrations and the Geometric Mean (GM) and Geometric Standard Deviation (GSD) of all individual model outcomes

Receptor	Distance (m)	Average observation (µg/m³)	GM (µg/m³)	GSD (-)
AF02	6.2	69	91	1.5
AF03	28	47	669	1.4
AF04	55	38	55	1.4
AF05	96	31	44	1.4
AF06	150	29	36	1.3
AF07	-14	35	47	1.4

14.2 Receptor contributions to total NMSE and VG

The outcomes for Normalised Mean Square Error (NMSE) and Geometric Variance (VG) were reported in section 4.2 Both scores apply to the full set of measurement data. In order to identify how important individual receptors are for these numbers, we have reported contributions from individual receptors to total NMSE and (the logarithm of) VG in the following tables. The mathematics needed for this analysis are provided in section 13.1.2 of Appendix 2.

The results for contributions to NMSE and the logarithm of VG are shown in Table 14.2 and Table 14.3. For most models, there is no receptor that determines the total score (either NMSE or VG) for more than 50%. Exceptions are AERMOD and IFDM, for which AF02 is very important, and ADMS for which AF06 has a contribution to VG larger than 50%. Compared with the Ringsted and Balko measurements, contributions from several receptors are more in balance.

Table 14.2 Contributions from the receptors to total NMSE

Receptor	ADMS	AERMOD	IFDM	OPS-LT	OPS_ST	SRM2	STACKS-D
AF07	3%	11%	13%	10%	3%	8%	8%
AF02	21%	64%	84%	31%	1%	35%	35%
AF03	3%	17%	1%	29%	23%	27%	35%
AF04	16%	6%	0%	19%	27%	16%	15%
AF05	13%	2%	1%	10%	26%	10%	6%
AF06	43%	1%	1%	1%	19%	4%	0%

Table 14.3 Contributions from the receptors to the logarithm of total VG

Receptor	ADMS	AERMOD	IFDM	OPS-LT	OPS_ST	SRM2	STACKS-D
AF07	3%	23%	34%	17%	5%	13%	14%
AF02	5%	31%	59%	14%	0%	15%	16%
AF03	2%	21%	2%	22%	15%	22%	32%
AF04	12%	12%	0%	22%	23%	20%	22%
AF05	15%	8%	2%	20%	30%	19%	15%
AF06	64%	3%	3%	5%	27%	11%	1%

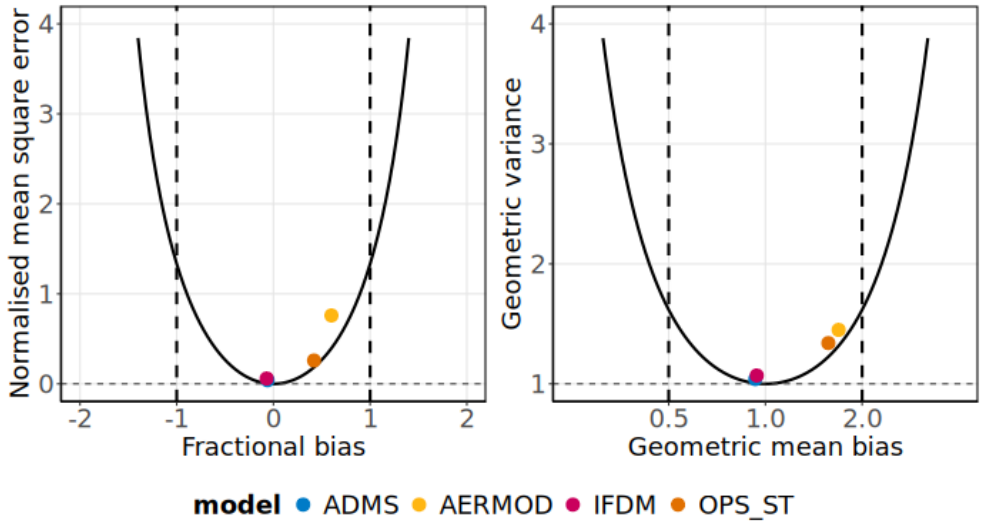
14.3 Results for weekly measurements

Performance indicator outcomes regarding weekly concentrations are shown in Table 14.4, for those models that provided hourly output for this campaign. The same results are visualised in Figure 14.1. In the analysis of weekly measurements, all models exhibited slightly higher NMSE and VG values compared with the full-campaign average, while FB and MG remained similar. FAC2 values for all models were lower than the full-campaign average but still remained above 50%.

Table 14.4 Model performance for weekly measured concentrations during the Affligem campaign

Model	FB	NMSE	MG	VG	FAC2
ADMS	-0.06	0.04	0.93	1.04	0.98
AERMOD	0.60	0.76	1.69	1.45	0.74
IFDM	-0.07	0.06	0.94	1.07	0.98
OPS-ST	0.42	0.26	1.57	1.34	0.77

Figure 14.1 Deviations and biases of weekly measured concentrations for the Affligem campaign. Left pane: FB and NMSE, right pane: MG and VG. Outcomes for ADMS and IFDM overlap in both panes.



15 Appendix 4 Detailed results for the Balko campaign

This appendix provides more detailed information on the Balko campaign, including period averages for measured NO_2 and NO_x concentrations and detailed individual model results.

15.1 Overview of period-average measurement outcomes

Period-average NO_2 and NO_x concentrations for several receptors are listed in Table 15.1. The period-average concentrations were derived from the hourly measurements that were downloaded from the EPA website [17]. Additional information on the processing of these measurement data is provided in section 16.2.3 of Appendix 5. The Geometric Mean (GM) and Geometric Standard Deviation (GSD) of the corresponding individual model outcomes (all models that provided data for this dataset), are also shown in Table 15.1.

Table 15.1 Period-average concentrations and the Geometric Mean (GM) and Geometric Standard Deviation (GSD) of corresponding model outcomes

Receptor	**Comp.**	**Observation ($\mu g/m^3$)**	**GM ($\mu g/m^3$)**	**GSD (-)**
East fence	NO_2	6.5	6.6	1.5
East fence	NO_x	12	11	1.7
North fence	NO_2	10	9.1	1.6
North fence	NO_x	36	20	1.8
Tower	NO_2	5.6	5.1	1.2
Tower	NO_x	7.5	7.0	1.3
Field	NO_2	4.7	6.1	1.2
Field	NO_x	9.5	8.9	1.3

15.2 Receptor contributions to total NMSE and VG

The outcomes for Normalised Mean Square Error (NMSE) and Geometric Variance (VG) were reported in section 5.2 Both scores apply to the full set of measurement data. In order to identify how important individual receptors are for these numbers, we have reported contributions from individual receptors to total NMSE and (the logarithm of) VG in the following tables. The mathematics needed for this analysis are provided in section 13.1.2 of Appendix 2.

Table 15.2 (NO_2 measurements) and Table 15.3 (NO_x measurements) show that outcomes for North Fence are often dominant for the total NMSE score; 'delivering' between 61% and 97% of the total NMSE outcome. Scores for STACKS-D relating to NO_2 are the only exception. Receptor contributions to VG (Table 15.4 and Table 15.5) are somewhat more in balance.

Table 15.2 Contributions of the receptors to total NMSE regarding NO_2

Receptor	ADMS	AERMOD	IFDM	OML	OPS_ST	STACKS-D
East fence	18%	14%	7%	12%	8%	88%
Field	4%	3%	20%	3%	9%	7%
North fence	71%	73%	61%	80%	82%	1%
Tower	7%	10%	12%	5%	0%	4%

Table 15.3 Contributions from the receptors to total NMSE regarding NO_x

Receptor	ADMS	AERMOD	OML	OPS-LT	OPS_ST	STACKS-D
East fence	3%	2%	2%	4%	4%	33%
Field	0%	0%	34%	1%	0%	1%
North fence	97%	97%	64%	95%	89%	65%
Tower	0%	1%	0%	1%	7%	1%

Table 15.4 Contributions from the receptors to the logarithm of total VG regarding NO_2

Receptor	ADMS	AERMOD	IFDM	OML	OPS_ST	STACKS-D
East fence	29%	22%	8%	20%	18%	76%
Field	8%	6%	35%	5%	31%	16%
North fence	50%	52%	35%	66%	51%	0%
Tower	13%	20%	22%	9%	0%	7%

Table 15.5 Contributions from the receptors to the logarithm of total VG regarding NO_x

Receptor	ADMS	AERMOD	OML	OPS-LT	OPS_ST	STACKS-D
East fence	18%	15%	5%	18%	11%	37%
Field	2%	0%	82%	3%	0%	5%
North fence	78%	74%	13%	75%	31%	53%
Tower	2%	12%	0%	4%	58%	5%

15.3 Results for hourly measurements

Performance indicator outcomes regarding hourly NO_2 and NO_x concentrations are shown in Table 15.6 (NO_2) and Table 15.7 (NO_x), for those models that provided the relevant hourly output. The same results are visualised in Figure 15.1. The hourly analysis only included impact hours, as defined by the filtering criteria outlined by PRCI in [20]. Model performance in the hourly analysis was lower than the full-campaign average for both NO_2 and NO_x. The cause was not investigated further. Additionally, all models showed better results when modelling NO_2 on an hourly basis compared with NO_x.

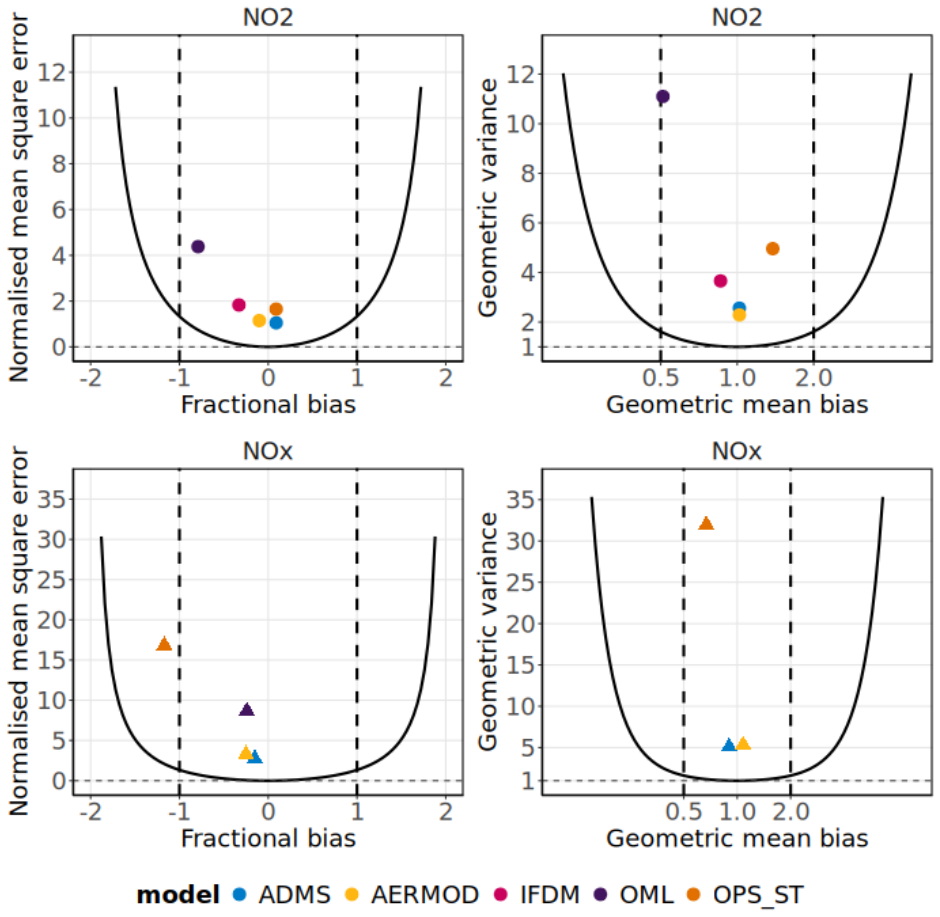
Table 15.6 Model performance for hourly NO_2 concentrations during the Balko campaign

Model	FB	NMSE	MG	VG	FAC2
ADMS	0.09	1.0	1.0	2.6	0.62
AERMOD	-0.10	1.2	1.0	2.3	0.63
IFDM	-0.33	1.8	0.86	3.7	0.49
OML-Multi	-0.79	4.4	0.51	11	0.48
OPS-ST	0.09	1.6	1.4	5.0	0.40

Table 15.7 Model performance for hourly NO_x concentrations during the Balko campaign

Model	FB	NMSE	MG	VG	FAC2
ADMS	-0.15	2.7	0.90	5.1	0.54
AERMOD	-0.25	3.3	1.10	5.3	0.49
OML-Multi	-0.24	8.6	0.50	290	0.36
OPS-ST	-1.2	17	0.67	32	0.28

Figure 15.1 Deviations and biases of hourly NO_2 and NO_x concentrations for the Balko campaign. Left-hand pane: FB and NMSE, right-hand pane: MG and VG. Note that OML-Multi is off the scale for NO_x in the right pane.



16 Appendix 5 Data processing for the validation study

This appendix provides a summary of the data processing of the three cases used for the validation study.

16.1 Processing of meteorological input data

The majority of the meteorological data that was used for all three validation cases was measured during the campaign and provided to RIVM along with other measurement data. Some missing input data was obtained from the hourly ERA5 meteorology data in the Copernicus dataset [18], [19].

16.1.1 Rain occurrence

Some of the models used for this validation study required information on rain occurrence. This refers to whether or not it rains within an hour or not and is a Boolean value expressed in 0 (FALSE) or 1 (TRUE). The variable was determined on the basis of the total precipitation for the hour, provided by the ERA5 data. If the total precipitation within an hour is larger than 0, this rain occurrence term is set to TRUE.

16.1.2 Rain duration

A related variable required by some models is rain duration. This refers to the duration of rain within each hour and is expressed as a ratio of an hour. Since this variable was not reported anywhere, an assumption was needed. To this end, a dataset including meteorological information over 10 years, from 2010 – 2020 in the Netherlands, including the rain duration, was analysed. The average rain duration per hour was calculated and was reported to be 23 minutes, or 0.38 hour. Therefore, we assume that if rain occurs within an hour (that is rain occurrence = 1), the rain duration is 0.38 hour.

16.1.3 Snow occurrence

Similar to the variable rain occurrence that was described before, some models required a Boolean variable for snow occurrence as well. To this end, we used the information on snowfall over the period, provided by the ERA5 data. If the total snowfall within an hour is larger than 0, the snow occurrence term is set to TRUE.

16.2 Data processing for analysis

16.2.1 Ringsted poultry farm

16.2.1.1 Observations

There were a total of 15 concentration measurement locations for the Ringsted campaign, measuring over 4 separate time periods that ranged from 10 to 17 days. Unfortunately, there was a lot of missing data for several periods during the campaign. For this reason, it was deemed unreasonable to group all concentration measurements together over the total campaign. Instead, we focused on the performance of the models within each of the available periods. Table 16.1 shows the number of locations that provided concentration measurements per period.

Table 16.1 The number of locations that include concentration measurement for each period during the Ringsted campaign.

Period	Start	End	Number of receptors
1	5 September	15 September	8
2	15 September	3 October	4
2.5	15 September	29 September	1
3	3 October	17 October	14

The original concentration observations, shared with us by Aarhus University, were reported in µg N/m^3. This unit was converted to µg NH_3/m^3 before comparing it with model outputs. Additionally, the concentration observations had been pre-processed to exclude the background concentration, but for a more realistic representation of the concentration values we wanted to include it. The background concentration for each period was described in another file provided to RIVM by Aarhus University (Passive-Per-nyBagg2.xls). To include the background in the measurements again, these values were added to the reported concentration measurements for each period. The background concentration for each period is described in Table 16.2.

The Ringsted campaign comprised a total of 25 deposition measurement locations. However, in his paper [6], Sommer excludes some receptors, regarding them as outliers. This includes all receptors located in the southerly direction (5 receptors) due to another NH_3 source south of the farm, as well as three receptors in the northerly direction and one in the easterly direction, due to a reported average deposition of below 0. There were two additional receptors in the westerly direction that reported negative deposition fluxes, but these were not reported as outliers. For the purpose of consistency with the approach of the original campaign organisers, we use the same terms for outliers for the deposition measurements. The final count of deposition receptors is, therefore, 16.

The original unit for the deposition in the dataset provided to RIVM by Aarhus University was in g N/m^2. Before comparing the observation with the modelled output, the units were converted to mol/m^2/yr. This was done using the following equation:

$$dep\ \left[\frac{mol\ N}{m^2 \cdot yr}\right] = \frac{dep\ \left[\frac{g\ N}{m^2}\right] \times 365\ \left[\frac{days}{year}\right]}{MW\ \left[\frac{g\ N}{mol}\right] \times t_{campaign}\ [days]}$$

where MW is the molecular weight of nitrogen (14 g/mol) and $t_{campaign}$ is the duration of the deposition measurements (54 days). For the deposition, no background concentration is added to the observations. We assume that the deposition that is modelled is dominated by the concentration contribution from the source.

16.2.1.2 Data processing of model outputs

To start with, all model outputs were consistently converted to the following units: µg/m^3 NH_3 for concentration and mol/ha/yr for deposition.

The Ringsted measurement campaign reports concentration measurements over four different periods, each ranging from ten to seventeen days. These periods are referred to as period 1, 2, 2.5, and 3 (see Table 16.1). To be able to compare the hourly models with the measurements, the modelled concentration was averaged over the hours within each period (Table 16.2). The concentration was additionally calculated for the full period of concentration measurements (‘period 4’). However, due to a large portion of missing data between the periods during the campaign, this value was not deemed accurate enough for usage in further analysis of the data.

For an accurate comparison of model outputs with observations, the two values were compared using background concentration. While most of the models (ADMS, AERMOD, IFDM, and OML-Multi) include background concentration in their model outputs, some models (OPS-ST, OPS-LT, STACKS-D) do not. For the outputs from these models, the average background concentration for each period was added to the modelled concentration. The values for the background concentration can be found in file Passive-Per-nyBagg2.xls, but are also described in Table 16.2.

The Ringsted campaign also reported results on deposition, but these results are only reported for one period, which has been marked as ‘period 5’. The modelled outcomes were processed here in a similar way as for the concentration. That is, for the hourly models, the deposition was averaged over the hours within period 5 and for the long-term models a special run from the start until the end of period 5 was made. Since there is no provided information on the hour on which the deposition measurements are started and ended, it was assumed that it started and ended at noon (12:00). No background concentration is included for the deposition measurements.

Table 16.2 Measurement periods for the Ringsted campaign. Star-marked periods cover the full campaign.

Type	Period	Start of period	End of period	Start of models	End of models	Bg. conc. [µg/m^3]
Con.	1	05-09-05 12:55:00	05-09-15 12:45:00	05-09-05 13:00:00	05-09-15 12:00:00	1.8
Con.	2	05-09-15 12:45:00	05-10-03 09:35:00	05-09-15 13:00:00	05-10-03 09:00:00	1.7
Con.	2.5	05-09-15 12:45:00	05-09-29 09:35:00	05-09-15 13:00:00	05-09-29 09:00:00	1.7
Con.	3	05-10-03 09:35:00	05-10-17 12:00:00	05-10-03 10:00:00	05-10-17 12:00:00	1.7
Con.	4*	05-09-05 12:55:00	05-10-17 12:00:00	05-09-05 13:00:00	05-10-17 12:00:00	1.7
Dep.	5*	05-09-01 12:00:00	05-10-25 12:00:00	05-09-01 12:00:00	05-10-25 12:00:00	-

16.2.2 Affligem motorway

16.2.2.1 Observations

For the Affligem case, the NO_2 concentration measurements were used. NO_2 concentrations were measured weekly with passive sensors over a period of 36 weeks. There were six measurement locations in total, and the concentration was measured in triplets at each location. To derive the single value that we use for the comparison with our model outcomes, the average value of the triplets was taken at each location, for each measurement period.

Within the observations from the Affligem campaign, the receptor AF03 had missing data for the 12-007 period. Additionally, the data at receptor AF06 in the 12-027 period is reported as an outlier in the dataset, due to very high measured concentrations compared with any of the other periods. Both of the values for these receptors at the correlating period are taken as non-applicable values (NA).

16.2.2.2 Data processing of model outputs

The outputs for NO_2 concentration from all models was converted to the unit µg NO_2/m^3 when needed. Each measurement period for the Affligem campaign spanned 14 days, but no start or end time was reported. For the purpose of data analysis, we assume for each of the measurement periods a starting time of 11:00 UTC and an ending time of 10:00 UTC. The hourly modelled concentrations within these timestamps are averaged for each measurement period. Additionally, the average concentration over the total period is calculated, using the starting time of the first period (period 12-001) and the ending time of the final period (period 12-036). The long-term models that are used in this validation study (OPS-LT and STACKS-D) only provide one average value over the total period and therefore, they can only be compared with the average value of the total campaign. Due to the consistency in measurements over the total period, we assume that comparing the average results over the total period is possible.

Out of the seven models used for this validation case, four provided model outputs without background concentration (OPS-ST, OPS-LT, STACKS-D, and SRM2). Since the measurements include background concentrations, the measured background concentration for each period is added to the model outputs. This is done by using the hourly, data-filled background concentration that was also used as model input. The background concentration over each period is, therefore, the average of the hourly outcomes for each period.

16.2.3 Balko compressor station

16.2.3.1 Observations

In the Balko campaign, two components were measured and can, therefore, be compared with the outputs, NO_2 and NO_x. The background concentration for both components was measured as well.
Out of the 13-month measurement period, 9.8% of the hours had a non-applicable measured value. These hours were not used for a comparison of the models. Additionally, there was a reported issue regarding the measured outcomes for NO_2 for the East fence (EF) receptor. The measurement team reported an ‘unidentified NO_2 measurement interference’ at both the East fence and, to a lesser

extent, the North fence (NF). This apparently resulted in much higher concentrations of NO_2 than was to be expected [50] (document itself not consulted in this study). In order to achieve more representative data for the East fence monitor, the measured ambient ratio between NO_2 and NO_x over the impacted hours was calculated and the NO_2 concentration at the East fence receptor was assumed on the basis of the measured NO_x concentration. This resulted in NO_2 concentrations for the EF receptor that were more in line with the measured values of the other receptors at the same hours. For the comparison of the observations and model results in this validation study, we use the adjusted NO_2 concentration for the EF receptor, according to the recommendations from PRCI [50].

The observations are reported in ppb. In order to compare the results with the model outcomes we need to convert the observations to μg/m^3. This can be achieved by using the equation:

$$C\left[\frac{\mu g}{m^3}\right] = C\,[ppb] \times \frac{P\,[Pa] \times MW\left[\frac{g}{mol}\right]}{R\left[\frac{m^3 \cdot Pa}{K \cdot mol}\right] \times T\,[K]}$$

where C is the concentration, P is the pressure in Pa, MW is the molecular weight (46.01 g/mol for NO_2), R is the gas constant (8.1314 m^3 Pa K^{-1} mol^{-1}), and T is the temperature in K. For the conversion, the pressure was assumed to be 1 atm or 101300 Pa, and temperature was taken from the measured temperature alongside the measured concentration. For some of the measured hours, the measured temperature is not recorded. In these cases, the temperature was assumed to equal the average temperature over the total period, which is 23.86 °C. Measured background concentrations for NO_2 were also reported in ppb. The same conversion is used to convert this concentration from ppb to μg/m^3.

As has been described in the previous chapter, an average value for concentration over the whole measurement campaign was calculated by grouping all available hours together and then calculating the average value.

16.2.3.2 Data processing of model outputs

Since hourly data was available for the Balko measurement campaign, there was no need to group model outputs together in the post processing. The model outputs for Balko observations have a timestamp for the start of the measurement hour, and the end time of the measurement output. The timestamp of the models, which shows the end time of the measurement hour, is matched to the ending hour of each measurement.

For the long-term models and the short-term models, an average concentration at each receptor was calculated and given the period mark of 9999. Here, the output for each hour of the models was averaged together, by excluding any NA values. The same was done for the observation values. The output from the long-term models was compared with the averaged outcome from the observations.

17 Appendix 6 Assumptions used for the modelling

This Appendix provides a summary of the assumptions used for the modelling. Detailed model descriptions have been provided in Appendix 6 of the model intercomparison study [2].

17.1 ADMS

ADMS runs were carried out by CERC. All campaigns were modelled using the latest available and most suitable ADMS family software. These share the same underlying model formulations, however, different software have different specialisations.

The Ringsted and Balko campaigns were modelled using ADMS 6.0, specialised for industrial sources, while Affligem was modelled using ADMS-Urban 5.0, specialised for urban areas. All settings not specifically mentioned below are left as default, which are specified in online ADMS user guides.

17.1.1 Specific assumptions for the Ringsted campaign

For Ringsted, the input meteorological dataset was used as supplied by RIVM (using underlying data from Aarhus University), as defined in section 3.1.2.

The point source emissions were varied hourly for each of the eight chimneys modelled. Data were used as provided by RIVM, although units were converted: normalised volume flux at 0 ℃ was converted to volume flux at the actual temperature (T) in Kelvin by multiplication by (T+273)/273; emission rate (g N/s) was converted to (g NH3 /s) by multiplication by (atomic mass of NH3)/(atomic mass of N) = (17.03)/(14.01).

A constant, average surface roughness value of 10 cm was used for concentration calculations, as described in section 3.1.3.

A constant background value of NH_3 = 1.7 µg/m^3 was used as described in section 3.1.5. Explicit NH_3 chemistry calculations are not included for this study but the influence of concentration magnitudes on deposition velocities is accounted for.

A concentration dependent spatially varying deposition ADMS input file was generated. This was created by first modelling the period average concentration of NH_3 across the domain (output at a ground level) and using the concentration at each modelled point to determine a deposition velocity at that location, according to UK government guidance. The land use type was taken to be short vegetation, which is consistent with the vegetation type at the deposition measurement locations (grass). The washout coefficient (A) was set to be 0.005, as recommended for NH_3 in the publicly available ADMS 6 user guidance.

Building effects were included for this study. Information about the four main buildings are provided in section 3.1.1. Satellite imagery was used to revise the building location relative to the sources.

17.1.2 Specific assumptions for the Affligem campaign

Meteorology data were used as provided by RIVM. A unit conversion was performed for cloud cover, from percentage to oktas. The minimum Monin-Obukhov length was set to 30 m, to account for site location near a city.

The emission source term varied in time, using hourly data provided, as described in section 4.1.1.

A constant surface roughness of 28 cm, corresponding to grass as the land use type, was used as described in section 4.1.3.

An hourly background was input, as provided by RIVM and described in section 4.1.5.

The model set up included the in-built 'Chemical Reaction Scheme' and the 'Local night-time chemistry' option included in ADMS-Urban, details can be found in the user guide.

17.1.3 Specific assumptions for the Balko campaign

The received processed meteorological datasets from AERMOD, as described in section 5.1.2, were reformatted for ADMS. This included meteorological parameters recorded at 9 m, and data from other heights, for a subset of parameters.

Hourly varying source emissions were used as provided, as described in section 5.1.1. The exit velocity of the source Clark 9 (C9) was modelled as hourly varying, while for all other sources this model input parameter remained constant.

A constant surface roughness of 6 cm was used as described for the control file in section 5.1.3.2. This corresponds to a land use type of cultivated land.

Hourly varying background concentrations for NO_x, NO_2 and O_3 were used as provided (see section 5.1.5), where hourly value is taken from the upwind monitor for a given hour, determined from the meteorological data.

Model set up included the in-built 'Chemical Reaction Scheme' and the 'Local night-time chemistry' option included in ADMS 6, details can be found in the user guide.

Building effects were included. All significant on-site buildings were included, as described in section 5.1.3.1, apart from the cooling tanks (COOLER1 and COOLER2).

17.2 AERMOD

AERMOD runs for Ringsted and Affligem were carried out by the UK Centre for Ecology & Hydrology (UKCEH). For Balko, existing AERMOD results from the EPA website for Air Quality Dispersion Modeling relating to the validation of the Generic Reaction Set Method for NO2 conversion in AERMOD were used. These validation runs had been carried out by CERC in 2023 [25].

17.2.1 Specific assumptions for the Ringsted campaign

AERMOD version 23132 for Windows and AERMET 21112 within BREEZE AERMET have been used for this case.

The raw meteorology file provided by RIVM contained hourly data between 01/09/2005 and 25/10/2005 for temperature (Celsius degrees), wind speed (m/s), wind direction, precipitation rate (mm/hr) solar radiation (W/m^2) and relative humidity.

Emissions: hourly variations were modelled, using the hourly measurements provided by RIVM. They were converted from gN/s to gNH_3/s (multiplication by 17/14).

Albedo, Bowen ratio, and surface roughness were set respectively at 0.28, 0.75 and 0.1 m for the whole domain. Land use was set to arable land.

Receptors were set using the coordinates provided by RIVM. Concentration receptors were placed at a 2 m height, and deposition receptors on the ground (0 m).

Building effects were modelled, merging buildings 1 to 3 into a single building (building 4 was ignored, since physically separated from the other buildings), as AERMOD couldn't accommodate for several buildings per source.

The ammonia background concentration was constant and set at 1.7 $\mu g/m^3$.

Deposition was modelled using the GASDEPOS card, with the following parameter values:

- Diffusivity in air: 0.1978 cm/s (Massman et al., 1998).
- Diffusivity in water: $1.54 \cdot 10^{-5}$ cm^2/s at 20° Celsius (Rives et al., 2022).
- Cuticular resistance: 10 s/m (Fléchard et al., 2010).
- Henry's law constant: 1.68875 $m^3 \cdot Pa/mol$ (Burkholder et al., 2019).

17.2.2 Specific assumptions for the Affligem campaign

AERMOD version 23132 for Windows and AERMET 21112 within BREEZE AERMET have been used for this case.

The raw meteorology file provided by RIVM contained hourly data between 01/04/2012 and 31/12/2012 for temperature (Celsius degrees), wind speed (m/s), wind direction, precipitation rate (mm/hr)

solar radiation (W/m^2) and relative humidity. Only the subset from 20th April 2012 10:00 UTC to 28th December 2012 11:00 UTC was kept.

Emissions varied with time (hourly), using provided variations factors for months, days and hours. The motorway was split in segments and lanes: a total of 56 LINE objects have been modelled in AERMOD. Despite the motorway having 2×3 lanes on the studied area, we modelled a 2×1 lane on most of it, and 2×3 lanes close to the receptors. Northernmost lanes were extended by 3 m to the North, and southernmost lanes by 3 m to the South. This was done to account for the mechanical turbulences caused by travelling vehicles (see below).

For 2×1 cases, a single lane was considered to have 3 sub-lanes (approx. 3.5m each), and an extra 3 m was added on the outer side: each 2×1 lane was 13.5 m wide, with its centre shifted away by 1.5 m (due to the extra 3 m). For 2×3 cases, a single lane was considered to have one 3.57 m sub-lane. If the lane was the outermost, 3 extras were added and its centre was shifted away by 1.5 m. hence 2×3 lanes were either 3.57 or 6.57 wide.

To deduce the initial vertical dimension coefficient and the release height, we used data from VMM website (downloaded on 16/08/24). On the period of interest, 88.6% non-trucks (mix of cars and vans shorter than 6.9 m; we assumed a height of 1.6 m) and 11.4% trucks (mix of vehicles longer than 6.9 m, with mostly coaches, buses, trucks, articulated vehicles; we assumed a height of 3.8 m) used the highway in both directions; we hence assumed an average vehicle height of 1.85 m overall. Using a guide on modelling roads with AERMOD, we deduced the initial vertical dimension coefficient szinit at 1.4615 m and the release height at 1.5725 m.

Land use was set to "Grassland" in AERMET and AERMOD. Albedo was fixed at 0.29 and Bowen ratio at 0.925 in AERMET. For surface roughness, spatial variations were introduced along 12 30-degree sectors in AERMET (see table below; angles start at the North (0 degree) and go clockwise).

Sector number	Sector	Surface roughness (m)
1	[0;30[	0.47
2	[30;60[	0.31
3	[60;90[	0.49
4	[90;120[	0.37
5	[120;150[	0.34
6	[150;180[	0.41
7	[180;210[	0.22
8	[210;240[	0.38
9	[240;270[	0.35
10	[270;300[	0.5
11	[300;330[	0.27
12	[330;360[	0.46

The above values were provided by RIVM, averaging from a 250 m resolution map in a 3 km radius around the centroid the z_0 values in each 30-degree sector.

Background concentration was fixed, using an annual average value of 10.6 µg/m^3 for NO_2.

Receptors coordinates were defined using the file provided by RIVM; no alteration was made to their locations.

We used the Ambient Ratio Method (ARM2 keyword, which adjusts NO_x concentrations basing on empirical relationships between NO_x and NO_2), as part of a 3-tiered approach (EPA: NO2 modeling techniques White Paper). Only tiers 1 and 2 could be performed, and tier 2 was selected for validation (ARM method).

Deposition was modelled using the GASDEPOS card, with the following parameter values:

- Diffusivity in air: 0.1361 cm/s (Massman et al., 1998).
- Diffusivity in water: $1.30 \cdot 10^{-5}$ cm^2/s at 10° Celsius (Schwartz et al., 1983).
- Cuticular resistance: 9999 s/m, as cuticular deposition is assumed to be negligible.
- Henry's law constant: 8333.33 m^3·Pa/mol (Burkholder et al., 2019).

17.2.3 Specific assumptions for the Balko campaign

For Balko, existing AERMOD results from the EPA website for Air Quality Dispersion Modeling relating to the validation of the Generic Reaction Set Method (GRSM) for NO_2 conversion in AERMOD were used. These validation runs had been carried out by CERC in 2023 [25].

The AERMOD output data for Balko were retrieved from https://gaftp.epa.gov/Air/aqmg/SCRAM/models/preferred/aermod/eval_databases/balko_oklahoma_no2.zip (accessed 19-2026). NO_2 concentrations are reported in the file Balko_GRSM.pst, while NO_x concentrations are reported in Balko_NOx.pst.

17.3 IFDM

IFDM runs were carried out by VITO, using IFDM version 7.1.

The conversion of NO_x into NO_2 is derived from fast-ozone chemistry, along the method of Hertel and Berkowicz, which is in detail described in (Berkowicz et al. 2008).

Wet deposition is calculated by the programme (IFDM) directly.

Dry deposition is calculated in after finishing the IFDM runs, normally using output from IFDM and VLOPS (note: VLOPS is the Flemish version of OPS-LT). Typically dry deposition is calculated on annual basis, using (i) annual-average IFDM concentrations and (ii) ratios of annual average deposition and annual average concentration from VLOPS. In the current project, OPS-LT output was used instead of VLOPS output. In addition,

the time averages in this project relate to the full durations of the campaigns (eight weeks for Ringsted, 36 weeks for Affligem and 13 months for Balko). The resulting formula for calculating 'IFDM' period-average dry deposition fluxes per location ($\overline{D_\iota}$) is:

$$\overline{D_\iota} = \overline{C^{IFDM}}_\iota \times {\overline{D^{OPS}}_\iota}\Big/{\overline{C^{OPS}}_\iota}$$

where $\overline{C^{IFDM}}_\iota$ is the period-average concentration per receptor location from IFDM, and $\overline{D^{OPS}}_\iota / \overline{C^{OPS}}_\iota$ the ratio of period-average deposition and concentration per receptor from OPS-LT. The required OPS-LT output was provided by RIVM.

17.3.1 Specific assumptions for the Ringsted campaign

The Ringsted case was run as follows:

- The required hourly meteorological data were provided by RIVM (using underlying data from Aarhus University), as defined in section 3.1.2.
- Regarding emissions, hourly ammonia emission data were used for each of the eight source locations (see section 3.1.1). These emission data from Aarhus University had been provided to VITO by RIVM.
- The surface roughness was varying in space, using data from Sommer et al., that were provided to VITO by Aarhus University via RIVM (see section 3.1.3).
- The NH_3 background concentration was not used because the model does not model any chemical reactors for NH_3. The background concentration was only added in the postprocessing of the data by RIVM.
- Building effects were not modelled in the IFDM run for the Ringsted campaign, because the effects of a small building in heterogeneous terrain are expected to be small.

17.3.2 Specific assumptions for the Affligem campaign

The Affligem case was run as follows:

- The required hourly meteorological input data were provided by RIVM (using underlying data from VMM and RMI), as defined in section 4.1.2.
- For emissions, hourly varying emissions strengths were used (see section 4.1.1).
- Surface roughness varied in space, using the surface roughness map with resolution 250×250 m^2 (see section 4.1.3).
- The conversion of NO to NO_2 was calculated specifically per hour and location, using ozone concentrations and other parameters for input. For NO, NO_2 and O_3, hourly varying background concentrations were used.

17.3.3 Specific assumptions for the Balko campaign

The Balko case was run as follows:

- The required hourly meteorological input data were provided by RIVM (using underlying data from EPA), as defined in section 5.1.2.
- For emissions, hourly varying emissions strengths were used (see section 5.1.1).

- Surface roughness was assumed to be 0.06 m across the entire domain (see section 5.1.3).
- The conversion of NO to NO_2 was calculated specifically per hour and location, using ozone concentrations and other parameters for input. For NO, NO_2 and O_3, hourly varying background concentrations were used.
- Building effects were not modelled.

17.4 OML-Multi

OML-Multi version 7.1 was used for the Ringsted campaign and the Balko campaign. The Affligem case was not run.

17.4.1 Specific assumptions for the Ringsted campaign

The Ringsted case had already been run with OML prior to this work. Input data for these runs had been provided to RIVM.

The original OML file with meteorological data was reused.

The emissions varied in time in the OML-Multi calculations where a daily cycle is included using temporal emission factors. The chosen emissions factors are based on a temporal analysis of the hourly emission measurements (see section 3.1.1).

Hour of day	Emission factor	Hour of day	Emission factor	Hour of day	Emission factor
0-1	0.53	8-9	1.39	16-17	1.96
1-2	0.53	9-10	1.88	17-18	1.30
2-3	0.53	10-11	2.13	18-19	1.02
3-4	0.53	11-12	2.39	19-20	0.82
4-5	0.53	12-13	2.21	20-21	0.69
5-6	0.57	13-14	2.21	21-22	0.65
6-7	0.65	14-15	2.17	22-23	0.61
7-8	0.90	15-16	2.21	23-00	0.61

Surface roughness and vegetation and inherently linked in OML-Multi; the vegetation type/land use classes determine the surface roughness. In OML-Multi it is only possible to specify one roughness length for the entire domain, which for the Ringsted campaign was set to 10 cm.

Building effects are included in OML-Multi. Buildings were considered to influence plume dispersion only if their height exceeded one-third of the stack height and their horizontal distance to the stack was less than two stack heights. Accordingly, only buildings meeting these criteria for each individual stack were included in the assessment of building-induced impacts on plume dispersion.

The provided background concentration of 1.7 µg NH_3/m^3 was included in the calculations.

No chemistry was simulated by OML-Multi.

17.4.2 Specific assumptions for the Balko campaign

For the Balko campaign, the meteorology provided by RIVM was run through a pre-processor that calculates the micro-meteorological parameters needed for OML-Multi.

The temporal emission factors for the Balko case applied in OML-Multi were derived through an analysis of temporal emission patterns for the four sources (C9, C10, EGEN, and Boiler). The analysis was performed in R using the timeVariation and calendarPlot functions from the openair package. In OML-Multi, monthly, day-of-week, and hourly emission factors can be specified. For this study, emission factors were defined as either 0 (no emissions) or 1 (fully operational). When a source was considered operational, the corresponding operational source strength from Table 5.1 was applied individually for each source.

- C9 was operational approximately 20% of the time and exhibited little diurnal variation. Operation occurred predominantly during the autumn and winter months. The static temporal emission factors were therefore defined to reflect the observed operational pattern. C9 was assumed to be operational in January and from September through December on Mondays, Tuesdays, and Wednesdays, with no hourly emission factors applied.
- C10 was operational approximately 19% of the time. The source was rarely operational between May and September, and lower activity was observed on Mondays and Sundays. Consequently, C10 was assumed to be operational from January to April and from October to December, on Tuesdays through Saturdays, between 06:00 and 20:00. This configuration results in an operational time of approximately 19% in the OML-Multi simulation.
- EGEN was operational approximately 0.9% of the time and was most frequently active on Thursdays for a few hours in the morning throughout the year. Accordingly, EGEN was assumed to be operational for three hours each Thursday morning in the OML-Multi calculations.
- The Boiler was operational approximately 54% of the time. No operation was observed from June to October, with limited activity in May. Furthermore, only minor diurnal variability and no strong dependence on day of the week were detected during operational periods. Emissions from the Boiler were therefore included in OML-Multi for January to April and November to December, for all days of the week, without applying hourly emission factors.

The year-average surface roughness length of 6 cm was used by OML-Multi for the Balko campaign.

The simulations with OML-Multi were run with the provided background concentrations of NO_x, NO_2 and O_3 activating the simplified NO_x chemistry scheme in OML-Multi.

Building effects were likewise included in OML-Multi for the Balko campaign (see criteria in section 17.4.1).

17.5 OPS-LT

Runs for the Ringsted and Balko campaigns were performed using OPS-LT version 5.2.1.0. For the Affligem campaign, two sets of runs were performed: one set of runs using version 5.2.1.0 (2024) and another set using 5.3.1.0 (2025). The latter set was only used to test the effect of two new options for road traffic modelling in this latest version. Results presented in the main report relate to version 5.2.1.0 (2024), unless stated otherwise.

For all three campaigns, input settings were equal to the settings that are normally used for such runs.

17.5.1 Specific assumptions for the Ringsted campaign

Many receptors in this campaign only had one or two concentration measurements available and thus, total period average concentrations could not be derived. The comparison of model outcomes with observations was, therefore, carried out for individual measurement periods; between 10 and 17 days for concentrations and 54 days for deposition (see section 3.2). Each measurement period was run separately, using the average emission strength, velocity and temperature per source over each period.

The Ringsted case was run using the meteo input provided by RIVM (using underlying data from Aarhus University), as defined in section 3.1.2. The surface roughness (10 cm) was constant in space and time. Regarding concentration measurements, the type of land cover was assumed to be arable land, see section 3.1.3. For deposition measurements (rye grass in pots), the land cover was set to grass.

The assumed background concentration was fixed in time, using the average background level for the full campaign duration (1.7 µg NH_3/m^3, see section 3.1.5). Chemical conversion rates (see [53]) were assumed to be equal to NL average values from EMEP for 2005.

Regarding the variation of emissions with time, the default setting dv=4 for animal housing was used. With this setting, a generic temperature correction is applied to the emission strength combined with a diurnal variation as an estimate for the day-night cycle [53].

Building effects were included for this campaign. The various buildings around the sources were combined into one single surrounding building as illustrated in Figure 3.6. Then, the effects of a building were read from the standard reference table for building effects in OPS-LT, using the dimensions of this imaginary surrounding building as input.

17.5.2 Specific assumptions for the Affligem campaign

The Affligem case was run with two different versions: 5.2.1.0 and 5.3.1.0, see the introduction to this section. Runs with 5.3.1.0 were performed with two new options for road traffic modelling turned on: one option to account for the varying NO/NO_2 ratio near a road, and another option to account for the enhanced turbulence from motorway traffic. The runs with version 5.2.10 used neither of these options (the first option was turned off, and the second option did not yet exist in

that version). Remaining input values were identical between the two runs.

The Affligem case was run with the meteo input provided by RIVM (using underlying data from VMM and RMI), as defined in section 4.1.2. For local surface roughness, we used the map provided by VITO (see section 4.1.3). Vegetation was assumed to be grass throughout the domain. Regarding background concentrations, the average background levels for the full campaign duration (see Table 4.7) were used. Chemical conversion rates (see [53]) were assumed to be equal to NL average values from EMEP for 2012.

Regarding the variation of emissions with time, only variations with hours of the day (Table 4.1) were accounted for. OPS-LT cannot accommodate weekly (Table 4.2) or monthly (Table 4.3) variations of source strength within a single run. Default road traffic settings were used for source heat, source height, and source height spread.

Table 17.1 Default source characteristics for road traffic in OPS-LT

	Heat content	**Source height**	**Initial vertical spread**
Version 5.2.1.0	0 MW	2.5 m	irrelevant
Version 5.3.1.0	0 MW	0.3 m	2.5 m

17.5.3 Specific assumptions for the Balko campaign

The Balko case was run with the meteo input provided by RIVM (using underlying data from EPA), as defined in section 5.1.2. The local surface roughness (5.9 cm) and vegetation (arable land) were assumed to be constant in space and time. For background, the average background levels for the full campaign duration (see Table 5.7) were used. Chemical conversion rates (see [53]) were assumed to be equal to NL average values from EMEP for 2016.

Emissions were assumed to be constant in time, using average emission rates for the entire period for each source (see Table 5.1).

Consistent with the default way of modelling industrial sources, building effects were not accounted for.

17.6 OPS-ST

All runs were performed using OPS-ST version 12.1.0 and using the input settings that we normally use for such runs.

17.6.1 Specific assumptions for the Ringsted campaign

Many receptors in this campaign only had one or two concentration measurements available and thus, total period average concentrations could not be derived. The comparison of model outcomes with observations was, therefore, carried out for individual measurement periods; between 10 and 17 days for concentrations and 54 days for deposition (see section 3.2). Since the current version of OPS-ST does not allow for hourly varying emission, each measurement period was run separately, using the average emission strength, velocity and temperature per source over each period.

The Ringsted case was run with the meteo input provided by RIVM (using underlying data from Aarhus University), as defined in section 3.1.2. The surface roughness (10 cm) was constant in space and time. Regarding concentration measurements, the type of land cover was assumed to be arable land, see section 3.1.3. For deposition measurements (rye grass in pots), the land cover was set to grass.

Regarding the variation of emissions over time, the default setting dv=4 for animal housing was used. With this setting, the hourly emission strength is modified by a correction factor that depends on temperature and wind speed.

The NH_3 background concentration was set to zero in the runs, because OPS-ST currently does not allow to calculate deposition fluxes for source contributions and background together. Then, the background concentration was added in the postprocessing of the data.

Two runs were carried out for the Ringsted campaign: one including building effects and another excluding building effects. This allowed RIVM to identify whether OPS-ST outcomes improve when accounting for building effects. The module for building effects was only recently added to OPS-ST [54], and experience with that module is still limited. For the main report, the outcomes including building effects were used because it seemed to be the most appropriate way of modelling releases near buildings. For the modelling of building effects, one 'representative building' was defined for the complex of buildings at the site. The four buildings at the Ringsted farm were very close together (see Figure 3.1). It was believed that these buildings should be modelled as one large building (see red rectangle in Figure 3.6).

17.6.2 Specific assumptions for the Affligem campaign

The Affligem case was run with the meteo input provided by RIVM (using underlying data from VMM and RMI), as defined in section 4.1.2. The local surface roughness was assumed to be constant throughout the domain and equal to 0.28 m (the average surface roughness length in a 1 km square around the measurement site was 0.28 m, see section 4.1.3). Vegetation was assumed to be grass throughout the domain.

Regarding the variation of emissions over time, only variations with hours of the day (Table 4.1) were accounted for. OPS-ST cannot accommodate weekly (Table 4.2) or monthly (Table 4.3) variations of source strength within a single run. Default road traffic settings were used for source heat (0 W), source height (2.5 m), and source height spread (1.25 m).

NO_x background concentrations were set to zero in the runs and were subsequently added during the postprocessing of output data.

17.6.3 Specific assumptions for the Balko campaign

The Balko case was run with the meteo input provided by RIVM (using underlying data from EPA), as defined in section 5.1.2. The local surface roughness (5.9 cm) and vegetation (arable land) were assumed to be constant in space and time. Emissions were assumed to be constant in

time, using average emission rates for the entire period for each source (see Table 5.1). NO_x background concentrations are not used in OPS-ST and were therefore added during the postprocessing of output data.

Two runs were carried out for the Balko campaign: one including building effects and another excluding building effects. This allowed RIVM to identify whether OPS-ST outcomes improve when accounting for building effects. The module for building effects was only recently added to OPS-ST [54], and experience with that module is still limited. For the main report, the outcomes including building effects were used because it seemed to be the most appropriate way of modelling releases near buildings. For the modelling of building effects, the characteristics of the nearest building were used for each source. It was believed that it was not necessary to define a 'representative building' for the complex of buildings at the site, because the various buildings were sufficiently far away from each other.

17.7 SRM2

SRM2 is a specific model for road traffic and therefore, it was only used for the Affligem campaign. The SRM2 run for Affligem was carried out by RIVM, using the SRM2 implementation in AERIUS Calculator.

17.7.1 Specific assumptions for the Affligem campaign

The Affligem case was run with the meteo input provided by RIVM (using underlying data from VMM and RMI), as defined in section 4.1.2. For local surface roughness, we used the map provided by VITO (see section 4.1.3). Regarding background concentrations, the average background levels for the full campaign duration (see Table 4.7) were used. Conversion of NO into NO_2 was calculated with the default conversion scheme in SRM2 and the period-average ozone background concentration (Table 4.7).

Regarding the variation of emissions with time, only variations with hours of the day were accounted for, using default SRM2 diurnal variation correction factors.

17.8 STACKS-D

The STACKS-D runs were carried out by RIVM, using an executable (dll) that was provided to RIVM by DGMR. DGMR also gave instructions on how to use that executable. During the summer of 2024, a bug in the software relating to the calculation of relative humidity was identified. This bug was subsequently fixed. The outcomes in this report used the corrected version of STACKS-D (using pcstack.dll version 19/9-2024 or 24-10-2024, both versions yield identical results).

STACKS-D normally requires runs for an entire year. For the validation study, DGMR provided RIVM with a version that can also run for shorter periods. Outcomes for STACKS-D involve average values (concentrations and deposition fluxes) for the total duration of the run.

17.8.1 Specific assumptions for the Ringsted campaign

Many receptors in this campaign only had one or two concentration measurements available and thus, total period average concentrations

could not be derived. Therefore, the comparison of model outcomes with observations was carried out for individual measurement periods; between 10 and 17 days for concentrations and 54 days for deposition (see section 3.2). Each measurement period was run separately, using the average emission strength, volume flux and emission temperature per source over each period.

The Ringsted case was run with the meteo input provided by RIVM (using underlying data from Aarhus University), as defined in section 3.1.2. The surface roughness (10 cm) was constant in space and time. Regarding concentration measurements, the type of land cover was assumed to be arable land, see section 3.1.3. For deposition measurements (rye grass in pots), the land cover was set to grass.

The assumed background concentration was fixed in time, using the average background level for the full campaign duration (1.7 µg NH_3/m^3, see section 3.1.5).

Building effects were accounted for, using one 'representative building' defined for the complex of buildings at the site. The four buildings at the Ringsted farm were very close together (see Figure 3.1). It was believed that these buildings should be modelled as one large building (see red rectangle in Figure 3.6).

17.8.2 Specific assumptions for the Affligem campaign

The Affligem case was run with the meteo input provided by RIVM (using underlying data from VMM and RMI), as defined in section 4.1.2. The local surface roughness was assumed to be constant throughout the domain and equal to 0.28 m (the average surface roughness length in a 1 km square around the measurement site was 0.28 m, see section 4.1.3). Vegetation was assumed to be grass throughout the domain. Regarding background concentrations, the average background levels for the full campaign duration (see Table 4.7) were used.

Regarding emissions, STACKS-D requires the definition of numbers of traffic movements for different vehicle types and emission factors for these vehicle types. Two vehicle types were used to define the Affligem case: passenger cars and medium and heavy-duty cars. The traffic movements were derived from the VMM traffic counts, only using the daily (24 hour) cycle and, therefore, not accounting for the weekly and seasonal patterns in vehicle movements. The emission factors per vehicle class were tuned to the desired total emission strength.

17.8.3 Specific assumptions for the Balko campaign

The Balko case was run with the meteo input provided by RIVM (using underlying data from EPA), as defined in section 5.1.2. The local surface roughness (5.9 cm) and vegetation (arable land) were assumed to be constant in space and time. For background, the average background levels for the full campaign duration (see Table 5.7) were used.

Emissions were assumed to be constant in time, using average emission strength, temperature and volume flux during operational hours for the full period per source (see Table 5.1).

For the modelling of building effects, the characteristics of the nearest building were used for each source. It was believed that it was not necessary to define a 'representative building' for the complex of buildings at the site, because the various buildings were sufficiently far away from each other. This approach is identical to the modelling in OPS-ST.

The receptor height was set to the maximum allowed value: 2 m. In reality, measurements were carried out at 2.5 m height.

STACKS-D only provides output for NO_2 concentrations (not NO_x concentrations). In order to get estimates for NO_x concentrations, additional runs were carried out with NO_2/NO_x ratios set to 1. The main limitation of this workaround is that the calculated deposition and depletion are too high (because the NO_2 fraction is too high). As a result, the calculated concentrations will be too low. However, it was expected that these limitations are insignificant, because the mass fraction that deposits within 425 m is sufficiently small, even when the emission is modelled as pure NO_2.

18 Appendix 7 Order of models in the two studies

The model intercomparison study [2] showed that some models calculated relatively high or low outcomes for some cases. It was unknown whether these findings apply more generally. The outcomes of the current study can be used to see whether the order of models is consistent between the two studies, for similar cases. For this analysis, we compared outcomes for individual models with the outcomes for other models. Model outcomes are regarded as 'quite in the middle' if the outcomes are within a 1 GSD range of the GM of the other outcomes.[43] If model outcomes deviate by about 1 GSD, the outcomes are said to be 'slightly above (or below) average', if the deviation ranges between 1 and 2 GSD, outcomes are regarded as 'above or below average', and if the deviation is larger than 2 GSD, outcomes are 'well above (or below) average'.

The order of models was compared between the two studies by using visual inspection of dedicated plots per individual model in which these individual model outcomes were set out against the average outcomes of other models, distinguishing between different directions and distances, for both studies. Model outcomes of the current study were averaged out over the full campaign duration, in order to provide a fair comparison with the annual average output of the model intercomparison study. For the current study, all predefined receptors were used (not only those receptors with measurement data).

18.1 Ringsted poultry farm – concentration

The concentration outcomes for the Ringsted poultry farm (see Figure 3.9 for the subset of locations with available measurement data) can be compared with the concentration outcomes for the livestock farm with a building in the model intercomparison study (Figure 3-10 in [2]). The distances used for this comparison range between 26 m and 570 m for the Ringsted case and between 50 and 500 m for the livestock farm case in the model intercomparison study.

The results of the comparison are summarised in Table 18.1. In general, the order of models compares well between the two studies: models that produced above-/below-average concentrations in the intercomparison study, also produce relatively high/low concentrations for the Ringsted case. IFDM is an exception: relatively high concentrations previously, and average concentrations now.

[43] The concerned model is excluded when calculating GM and GSD in order to get more contrast.

Table 18.1 Order of models regarding concentration outcomes for the livestock farm with building in the model intercomparison study and for Ringsted

Model	**Concentrations for Livestock farm with building**	**Ringsted concentrations**
ADMS	In the middle or slightly below average	In the middle or slightly below average
AERMOD	In the middle or slightly below average	Quite in the middle
IFDM	Well above average	In the middle or slightly above average
OML-Multi	Quite in the middle	Quite in the middle
OPS-LT	Quite in the middle	In the middle or slightly below average
OPS-ST	Above average	Well above average
STACKS-D	In the middle or slightly below average	Well below average close to the source, in the middle further away

18.2 Ringsted poultry farm - deposition

The deposition outcomes for the Ringsted poultry farm (see Figure 3.11 for the subset of locations with available measurement data) can be compared with the deposition outcomes for the livestock farm with a building in the model intercomparison study (Figure 3-12 in [2]). The distances used for this comparison range between 11 m and 260 m for the Ringsted case and between 50 and 200 m for the livestock farm case in the model intercomparison study.

The results of the comparison are summarised in Table 18.2. The order of the models for the two studies is mostly similar.

Table 18.2 Order of models regarding deposition outcomes for the livestock farm with building in the model intercomparison study and for Ringsted

Model	**Deposition for Livestock farm with building**	**Ringsted deposition**
ADMS	Slightly above average	(Slightly) above average
AERMOD	(Well) below average	Well below average
IFDM	In the middle or slightly above average	In the middle or slightly above average
OML-Multi	Quite in the middle	Quite in the middle, above average at the largest distance
OPS-LT	Below average	Quite in the middle
OPS-ST	Quite in the middle	Quite in the middle
STACKS-D	Below average close to the source, in the middle or slightly above average farther away	Below average close to the source, in the middle farther away

18.3 Affligem motorway

The concentration output from the Affligem motorway (see Figure 4.5 for the subset of locations with available measurement data) can be compared with the concentration outcomes for the Motorway without noise barrier in the model intercomparison study (Figure 3-27 in [2]). Some caution is required when comparing data from these two studies, because outcomes of the model intercomparison study involve NO_x, while output for the Affligem case involves NO_2. Results for OPS-LT relate to its 2024 version, that is, not using the two recent improvements for modelling motorway traffic. The distances used for this comparison range between 6 and 500 m for the Affligem case and between 50 m and 1 km for the motorway case in the model intercomparison study.

The results of the comparison are summarised in Table 18.3. Except for ADMS, the order of models is similar for the two studies.

Table 18.3 Order of models for the motorway case in the model intercomparison study (NO_x) and for Affligem (NO_2)

Model	**Concentrations for the motorway without barrier (NO_x)**	**Affligem concentrations (NO_2)**
ADMS	(Slightly) above average	Below average
AERMOD	Quite in the middle	Quite in the middle
IFDM	Well below average	(Slightly) below average
OPS-LT	Quite in the middle	Above average close to the source, in the middle at farther distance
OPS-ST	Above average close to the source, well above average farther away	In the middle close to the source, well above average farther away
SRM2	Quite in the middle	Quite in the middle
STACKS-D	Quite in the middle	Quite in the middle

18.4 Balko compressor station

The concentration outcomes for the Balko compressor station (see Figure 5.4 for the subset of locations with available measurement data) can be compared with the concentration outcomes for the tall buoyant stack in the model intercomparison study (Figure 3-23 in [2]). Only NO_x concentrations are compared since NO_2 concentration was not modelled in the intercomparison study. The distances used for this comparison range between 100 m and 1 km for the Balko case and between 50 m and 1 km for the industrial case in the model intercomparison study.

The results of the comparison are summarised in Table 18.4. For most models, its order relative to other models is quite similar for the two studies. OML-Multi and STACKS-D show different patterns between different directions in the Balko study. Underlying causes have not been investigated.

Table 18.4 NO_x concentration outcomes for the industrial case in the model intercomparison study and for the Balko compressor station

Model	**Concentrations for the tall industrial stack**	**Balko NO_x concentrations**
ADMS	Quite in the middle	Quite in the middle
AERMOD	Quite in the middle	Quite in the middle (only output for the four receptors with measurement data available)
IFDM	Quite in the middle	No NO_x output available, quite in the middle for NO_2.
OML-Multi	Quite in the middle	In the middle in easterly and westerly directions, above average in northerly and southerly directions
OPS-LT	Quite in the middle	Below average
OPS-ST	Above average close to the source, in the middle farther away	Quite in the middle
STACKS-D	Well below average	Above average East and West, below average North and South

Published by

National Institute for Public Health and the Environment, RIVM
P.O. Box 1 | 3720 BA Bilthoven
The Netherlands
www.rivm.nl/en

March 2026

Committed to health and sustainability