EXTENDED ABSTRACT

Abstract title: A validation study for eight operational models regarding concentration and deposition of ammonia and nitrogen oxides at a local scale

Gudrún Thorkelsdottir, National Institute for Public Health and the Environment (RIVM)
gudrun.thorkelsdottir@rivm.nl

Eelke Kooi [1], Pim Meijer[1], Jenny Stocker[2], Wouter Lefebvre[3], Audric Vigier[4], Anne Sofie Lansø[5], Hans Hooyberghs[3], Cor Jacobs[1], Marina Sterk[1], Roy Wichink Kruit[1]

[1] National Institute for Public Health and the Environment (RIVM)
[2] Cambridge Environmental Research Consultants (CERC)
[3] Flemish Institute for Technological Research (VITO)
[4] UK Centre for Ecology and Hydrology (UKCEH)
[5] Aarhus University, Department of Environmental Science

Introduction

Deposition of reactive nitrogen species on natural ecosystems can have negative effects on environmental quality and biodiversity. Nitrogen comes from different sources such as traffic, agriculture and industry. Estimates of nitrogen deposition are used to inform policy and for permitting of economic activities. The Dutch government has initiated a research program that aims to improve the national reactive nitrogen modelling and measurement strategy. As part of this research programme, simulations from eight commonly used atmospheric dispersion and deposition models were compared to observations from three local scale measurements campaigns. The key goals of the study were to assess the accuracy of the modelled concentrations and deposition fluxes in the vicinity of sources, and to identify if an ensemble outcome of multiple models compares better with observations than individual models. All results from this study will be shared in a public report, which will be published in the near future.

Selection of campaigns

The primary focus of this study was to validate modelled deposition. Availability of deposition measurement data was investigated with the following criteria in mind: 1) total campaign duration of at least 1 month, 2) measurements in close range (roughly 50-500 m) from the dominant local source in terms of deposition, 3) well-defined source emission characteristics, and 4) total deposition measurements. Three types of sources of reactive nitrogen were of interest: a livestock farm with ammonia (NH_3) emissions, an industrial stack with emissions of nitrogen oxides (NO_x) and a motorway with NO_x emissions.

It was challenging to find measurement campaigns fulfilling all requirements. Of all the measurement campaigns reviewed, only one fulfilled the above criteria: a campaign measuring ammonia deposition and concentrations near a poultry farm in Ringsted, Denmark (Sommer et al. 2009). Despite fulfilling the criteria, the accuracy of the deposition measurements was not as good as had been hoped for.

As deposition strongly correlates with concentration, two further campaigns with concentration measurements only were selected: one campaign measuring NO_2 concentrations around a motorway in Affligem, Belgium (VMM 2013), and another campaign measuring NO_x and NO_2 in the vicinity of a compressor station with multiple exhaust stacks in Balko, Oklahoma, USA (Panek et al. 2020). For all campaigns, deposition and/or concentration were measured close to the source, with most measurements located within 250 m distance.

Methodology

The validation study included eight commonly used atmospheric dispersion and deposition models. Five models provided hourly results: ADMS, AERMOD, IFDM, OML-Multi and OPS-ST (hourly models), and three provided average results over the total campaign period: OPS-LT, STACKS-D and SRM2 (annual models). When estimating total nitrogen deposition in sensitive areas, annual averages are commonly used. Therefore, total period averages are considered to be more important than hourly or (bi)weekly values and comparisons between the model outcomes and the measured variables are thus made for the total average over the campaign period. Biweekly concentration measurements were used for the Ringsted campaign because the total period average concentration could not be derived from the available measurements.

The study investigates the ability of models to predict concentrations and deposition at varying distances from the source. Model performance was investigated by comparing modelled outcomes with observations pairwise and using statistical performance metrics from Chang & Hanna (2004). Of the statistical metrics, logarithmic measures (Geometric mean bias, MG, and Geometric variance, VG)[1] were considered more appropriate than non-logarithmic measures (Fractional bias, FB, and Normalised mean square error, NMSE) because concentrations and deposition fluxes tend to reduce rapidly with distance from the source. The logarithmic metrics provide a better balance between all measurement locations (COST ES1006, 2015). When comparing model simulations and measured outcomes, deviations can occur due to limited accuracy of model input, limited precision of the model itself and measurement inaccuracies. The statistical performance metrics do not solely reflect model accuracy.

In addition to comparing individual model performance, different model ensembles were evaluated, using consistent statistical metrics. The ensemble outcome was defined as the

[1] The following formulae apply: $MG = e^{\overline{\ln(M)}-\overline{\ln(O)}}$ and $VG = e^{\overline{(\ln(O)-\ln(M))^2}}$. For MG, values larger than 1 indicate average overprediction by a model in this study, contrary to the initial definition from Chang & Hanna (2004).

geometric mean (GM) of the individual model outcomes[2], in order to balance both large and small outcomes from individual models.

Results

Figure 1 summarises the statistical performance of individual models and ensembles for each campaign and measurement type, using the geometric mean bias (MG) and the geometric variance (VG). The black solid line represents the minimum variance related to bias, i.e. the variance in absence of random scatter around mean bias.

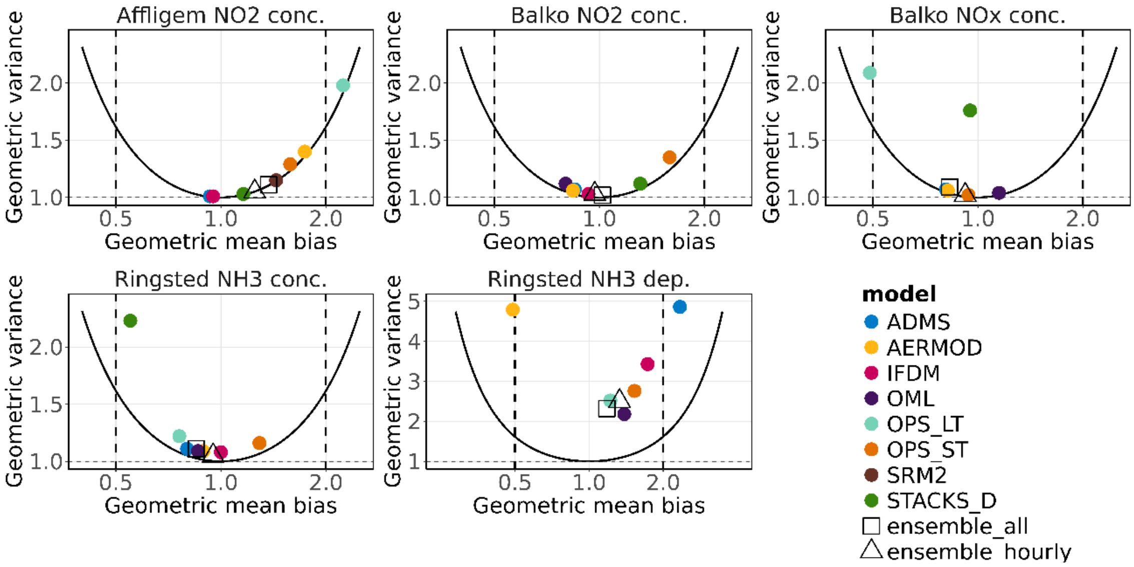

Figure 1 Geometric variance and geometric mean bias for the different models and ensembles per campaign and measurement type. STACKS-D outcomes for the Ringsted NH_3 deposition is off the scale. Note that the vertical axis for the Ringsted NH_3 deposition plot differs from the other plots. The black dashed lines indicates a model bias towards under/overprediction of a factor of two.

In general, modelled concentrations compare well with measurements for all campaigns. The VG is often close to the solid line, indicating limited random scatter in the model outcomes. For the Affligem campaign, the majority of the models have a positive bias compared to measurements. The Balko and Ringsted concentration measurements do not show consistent signs for biases between the models. In general, according to the metrics used here, the hourly models tend to perform better than the annual models.

For the deposition measurements in the Ringsted campaign, the vertical distance from model outcomes to the solid black line in Figure 1 is much larger than for any of the concentration measurements (Ringsted, Affligem and Balko). The range of model biases is also larger for deposition than for concentration. Six out of the seven models calculate more deposition than measured; explanations for this behaviour have not been investigated during the current study.

---

[2] The following formula applies: $GM = exp\left[\frac{1}{n}\sum_{i}^{n}\ln(M_i)\right]$.

The increased complexity of modelling deposition processes, compared to concentrations, is likely to partly explain why models perform worse for deposition than for concentration. The accuracy of deposition calculations depends on the accuracy of calculated concentrations and of predicted deposition velocities, and can therefore not be expected to be better than the accuracy of calculated concentrations alone. In addition, deposition measurement uncertainty is larger than concentration measurement uncertainty. For Ringsted, triplet deposition measurements at identical locations showed significant variation, and mean outcomes of deposition measurements per location did not show the same smooth trends with distance as concentrations measurements.

For concentrations, an ensemble of the hourly models in this study had on average a slightly better performance score than most individual models, for all campaigns. Including the annual models to the ensemble did not significantly improve or worsen the performance of the ensemble. Both the ensemble of the hourly models and the ensemble of all models outperformed most of the individual models for the deposition measurements in the Ringsted campaign, with the ensemble of all models performing slightly better between the two. However, the ensembles still have a significant bias and variance from the observations. This is because for these measurements, six out of the seven models had positive bias. When all models show a similar over- or underprediction to measurements, the resulting ensemble outcome will be somewhere between the individual models outcomes. Therefore it cannot become (significantly) closer to measured outcomes than an average model.

Conclusions

Outcomes of eight atmospheric dispersion and deposition models were compared to three campaigns with concentration measurements and to one campaign with deposition measurements. All measurements were nearby sources. Model predictions for deposition fluxes deviated more from measurements than modelled concentrations. This is expected to be the result of the larger complexity of modelling and measuring deposition. Regarding concentrations, hourly models generally compared better to measurements than annual models.

Using an ensemble of the hourly models for predicting concentration resulted in similar or slightly better performance compared to the individual models, while adding the annual models did not significantly improve or worsen the ensemble performance. For the deposition measurements, the ensemble performed similarly to the best performing individual models, but not better, possibly because outcomes of models with relatively small deviations to measurements were mixed with outcomes of models with relatively large deviations.. The benefits of an ensemble of models therefore depends, to some extent, on the performance of the individual models. In cases where most models show a consistent bias compared to measurements, the ensemble will show a similar bias.

The outcomes of performance metrics for comparing modelled and measured outcomes depend on the accuracy of the model input, model processes and measurements. Information on which of these contributions is dominant was not further investigated in this study. The study therefore provides insight into any of the individual model performances compared to measurements and the other models, but it does not explain the reason behind any of the differences between modelled and measured outcomes.

References

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 18-7-2025).

Vlaamse Milieumaatschappij (VMM). 2013. Life+ ATMOSYS snelwegcampagne: Luchtkwaliteit nabij de E40-snelweg in Affligem (in Dutch). 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 18-7-2025).

Panek, J.A., McCarthy, J.M., Huth, A.Z., Krol, A.J. and 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).

Chang, JC & Hanna, SR, 2004. Air quality model performance evaluation, Meteorology and atmospheric physics 87, pp. 167-196. Available from https://link.springer.com/article/10.1007/s00703-003-0070-7

Chang, J.C. and Hanna, S.R., 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).

COST ES1006 – Model Evaluation Case Studies. COST Action ES1006. April 2015.