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

Authors: G. Thorkelsdottir[1], E. Kooi[1] , P. 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. RIVM, National Institute for Public Health and the Environment, Bilthoven, the Netherlands
2. CERC, Cambridge Environmental Research Consultants
3. VITO, Flemish Institute for Technological Research
4. UKCEH, UK Centre for Ecology and Hydrology
5. Aarhus University, Department of Environmental Science

**Contact:** gudrun.thorkelsdottir@rivm.nl

Background

Deposition of reactive nitrogen species on natural ecosystems can have negative effects on environmental quality and biodiversity. The Dutch government has initiated a research program that aims to improve the reactive nitrogen deposition modelling and measurement strategy. As part of this research program, concentration and deposition estimates from eight commonly used atmospheric dispersion and deposition models were compared to observations from three local scale measurements campaigns. The main aims of the study were 1) to assess the accuracy of the modelled concentrations and deposition fluxes; 2) to identify if an ensemble outcome of multiple models compares better with observations than individual model outcomes.

Selection of measurement campaigns

Overall it was challenging to find suitable deposition measurements. Ideally, the campaigns should have a runtime of at least a month, include measurements in a close range from a local defined source, and provide total deposition measurements. Only one campaign (partly) fulfilled these criteria. However, the measurement accuracy of the deposition measurements was relatively low compared to the concentration measurements. Additionally, two campaigns that only provided concentration measurement were included. For all campaigns, most measurements were located within 250 m distance.

The measurement campaigns:

1. **Poultry farm in Ringsted, Denmark (2005)**
   - **Component:** NH_3
   - **Measurements:** Concentration and deposition

2. **Motorway in Affligem, Belgium (2012)**
   - **Component:** NO_2
   - **Measurements:** Concentration

3. **Compression station in Balko, Oklahoma, USA (2015 – 2016)**
   - **Component:** NO_x and NO_2
   - **Measurements:** Concentration

Methods

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, and three provided average results over the total campaign period, OPS-LT, STACKS-D and SRM2. Model performance was investigated using two logarithmic statistical metrics; The Geometric Mean Bias indicates if a model, on average, over- or underpredicts compared to observations, and the Geometric Variance indicates the deviation of model outcomes from observations.

The performance of two different model ensembles was also investigated for each campaign: one ensemble containing only models with hourly outputs and the other containing all models. The ensemble average was defined as the geometric mean.

Table 1. The two ensembles that were investigated included the following models.

Ensemble with hourly models	Ensemble with all models
ADMS, AERMOD, IFDM, OML-Multi, OPS-ST	ADMS, AERMOD, IFDM, OML-Multi, OPS-ST, OPS-LT, STACKS-D, and SRM2

Results

For all campaigns, concentration outcomes from individual models tend to compare well with measured concentrations. For deposition, larger deviations between models and measurements are found. These larger deviations are a combined result of the increase complexity of modelling and measuring deposition.

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. The black dashed lines indicates a model bias towards under/overprediction of a factor 2. The black solid line shows the minimum Geometric Variance for each Geometric Mean Bias.
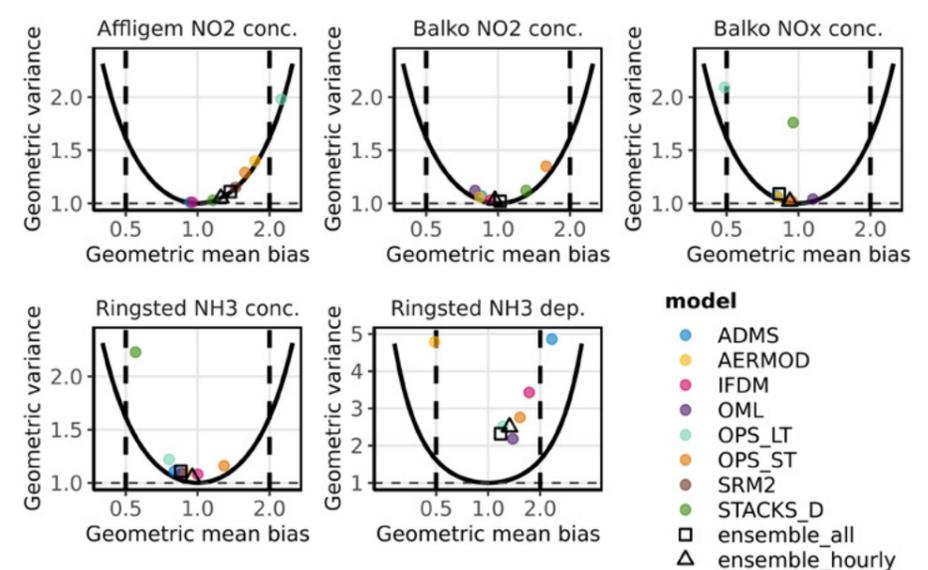
For concentration, the ensemble of hourly models had on average a slightly better performance score than the individual models. Adding the annual models did not improve the performance of the ensemble. Neither an ensemble of the hourly models nor the ensemble of all models outperformed the best performing individual models for the deposition measurements in the Ringsted campaign.

Conclusions

- Model predictions for deposition fluxes deviated more from measurements than modelled concentrations.

- The performance metric used to compare modelled and measured outcomes depends on the accuracy of the model input, model processes and measurements. These metrics provide insight into overall performance of individual models compared to measurements and other models, but do not provide information on what is the cause of any differences.

- Hourly models had smaller deviations than annual models with concentration measurements. The ensemble of hourly models had similar or better performance scores than the individual models. Adding annual models to the ensemble did not improve the performance of the ensemble.

- For deposition, the hourly models did not perform better than the annual models. The performance of both model ensembles did not clearly outperform the better performing individual models.