

This item is the archived peer-reviewed author-version of:

Validation and optimization of the ATMO-Street air quality model chain by means of a large-scale citizen-science dataset

Reference:

Hooyberghs H., Lefebvre W., Vranckx S., Maiheu B., Trimpeneers E., Vanpoucke C., Janssen S., Meysman Filip, Fierens F., De Craemer Sam.- Validation and optimization of the ATMO-Street air quality model chain by means of a large-scale citizen-science dataset
Atmospheric environment : an international journal - ISSN 1873-2844 - 272(2022), 118946
Full text (Publisher's DOI): <https://doi.org/10.1016/J.ATMOSENV.2022.118946>
To cite this reference: <https://hdl.handle.net/10067/1886630151162165141>

Validation and optimization of the ATMO-Street air quality model chain by means of a large-scale citizen-science dataset

Hooyberghs H.¹, De Craemer S.^{1,2}, Lefebvre W.¹, Vranckx S.¹, Maiheu B.¹, Trimpeneers E.^{3,4}, Vanpoucke C.^{3,4}, Janssen S.¹, Meysman F. J. R.^{2,5}, Fierens F.^{3,4}

¹ Unit Environmental Modelling, VITO, Boeretang 200, 2400 Mol, Belgium

² Department of Biology, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk (Antwerpen), Belgium

³ VMM Vlaamse Milieumaatschappij, Kronenburgstraat 45, 2000 Antwerpen, Belgium

⁴ Belgian Interregional Environment Agency, Gaucheretstraat 92-94, 1030 Brussels, Belgium

⁵ Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands

Abstract

Detailed validation of air quality models is essential, but remains challenging, due to a lack of suitable high-resolution measurement datasets. This is particularly true for pollutants with short-scale spatial variations, such as nitrogen dioxide (NO₂). While street-level air quality model chains can predict concentration gradients at high spatial resolution, measurement campaigns lack the coverage and spatial density required to validate these gradients. Citizen science offers a tool to collect large-scale datasets, but it remains unclear to what extent such data can truly increase model performance. Here we use the passive sampler dataset collected within the large-scale citizen science campaign CurieuzeNeuzen to assess the integrated ATMO-Street street-level air quality model chain. The extensiveness of the dataset (20.000 sampling locations across the densely populated region Flanders, ~1.5 data points per km²) allowed an in-depth model validation and optimization. We illustrate generic techniques and methods to assess and improve street-level air quality models, and show that considerable model improvement can be achieved, in particular with respect to the correct representation of the small-scale spatial variability of the NO₂-concentrations. After model optimization, the model skill of the ATMO-Street chain significantly increased, passing the FAIRMODE model quality threshold, and thus substantiating its suitability for policy support. More generally, our results reveal how a “deep validation” based on extensive spatial data can substantially improve model performance, thus demonstrating how air quality modelling can benefit from one-off large-scale monitoring campaigns.

Keywords

Air pollution, Spatial Variation, Dispersion modelling, Street level modelling, Citizen Science, Model validation, Model optimization, FAIRMODE model quality objective, Semi-variogram analysis

41 1. Introduction

42

43 Air pollution remains a key environmental problem in most European cities (WHO 2016; EEA 2019),
44 and so an accurate assessment of air pollution patterns and abatement strategies is vitally important
45 to reduce the impact on human health. Many of the associated policy questions are addressed using
46 air quality models: models have been successfully applied to interpolate pollution levels in between
47 measurement locations (Thunis et al. 2016), estimate the population exposure on regional and urban
48 scales (Jerrett et al. 2005; Hoek 2017; Xie et al. 2017), and quantify the health impact related to long-
49 term exposure (Hoek et al. 2013; Faustini et al. 2014; EEA 2019). Additionally, air quality models are
50 essential tools to develop and evaluate policy scenarios (Miranda et al. 2015; Brusselen et al. 2016;
51 Thunis et al. 2016).

52

53 Nitrogen dioxide (NO₂) is one of the important air pollutants in urban environments. More than 90%
54 of the urban population in the EU is exposed to concentrations that exceed the guidelines put forward
55 by the World Health Organization (WHO), leading to approximately 70.000 premature deaths every
56 year (EEA 2019). When quantifying the population exposure and health impacts of NO₂, a particular
57 challenge is the spatial heterogeneity of the concentration field. Because of street-canyon effects and
58 the proximity to main emission sources, the NO₂-concentrations vary strongly over short distances
59 (Marshall et al. 2008; Cyrus et al. 2012; Lefebvre et al. 2013b; Jensen et al. 2017). To attain suitable
60 model skill, air quality models should adequately capture this short-scale spatial variation, and reliably
61 predict the concentration field on a scale of tens of meters.

62

63 Validation (and subsequent model improvement) are essential when models are used for regulatory
64 purposes. Model simulated pollution maps need to be validated at the proper spatial and temporal
65 scales. Street level models that target prediction of within-street variation of NO₂ at high spatial
66 resolution, should hence be validated using measurement campaigns that have a suitably dense
67 sampling grid. Measurements in streets with different traffic loads are required to capture the small-
68 scale spatial variability of NO₂-concentrations. Because of logistical and financial constraints, such a
69 high sampling density cannot be obtained using official telemetric stations (Vardoulakis et al. 2011).
70 As an alternative, wind tunnel experiments have been used (Ketzler et al. 2000; Baker and Hargreaves
71 2001). Although these validation campaigns provide an opportunity to validate air quality models in a
72 controlled environment (e.g. controlled boundary conditions) (Vardoulakis et al. 2003), one of the
73 main challenges in field campaigns is to handle all the variability in boundary conditions and the way
74 long-term averages are achieved.

75

76 Mass-scale citizen science offers an innovative way to generate the large datasets required for such a
77 validation campaign (Irwin 2018; Van Brussel and Huyse 2019; De Craemer et al. 2020a; Meysman et
78 al. 2020; Bo et al. 2020), but it is presently unclear to what extent such datasets can truly generate
79 improved model performance. There is an important trade-off in this respect. While citizen science
80 has the advantage of generating data at high spatial resolution, one typically uses passive sampler
81 measurements, and so the resulting data is generally less accurate and of lower quality than those
82 collected via official telemetric stations. Citizen science has clear benefits in terms of raising awareness
83 about air pollution (Van Brussel and Huyse 2019), but to what extent can the resulting high-resolution
84 data truly support the improvement of air quality models?

85

86 To address this question, we validate and optimize the ATMO-Street model chain (Lefebvre et al.
87 2013b) using the extensive NO₂-dataset collected within the Curieuzeneuzen citizen science project
88 (<https://2018.curieuzeneuzen.be/>). While this article makes a case study of one particular model
89 chain, many of our findings, methods and techniques are readily and generically applicable to other
90 (street-level) models, and so the conclusions are highly relevant for air quality models in general.

91

92 ATMO-Street is an integrated model chain (Lefebvre et al. 2013b) that models air quality at high,
93 street-level resolution (i.e. 10 m), and hence representative for the class of high-resolution, state-of-
94 the-art models that is operated by Environment Agencies across the world for planning and policy
95 purposes. ATMO-Street is used by the Flanders Environment Agency (Vlaamse Milieumaatschappij,
96 VMM) to assess the air pollution at the street level scale for Flanders, a densely populated region in
97 Northwestern Europe (13,522 km², 485 inhabitants km⁻²; total population 6,552,000). In addition,
98 ATMO-Street is the default tool used for planning purposes, evaluating the impact of regional and
99 local air quality plans and health impact assessments in Flanders. The model chain has been previously
100 validated via several dedicated measurement campaigns, focusing both on spatial patterns and time
101 series (Lefebvre et al. 2011; Lefebvre et al. 2013b). However, these validation campaigns focused on
102 a relatively small number of sampling locations, with at most a few dozens of locations distributed
103 among a single urban region.

104
105 In 2018, the citizen science project *CurieuzeNeuzen Vlaanderen* project engaged 20.000 citizens across
106 Flanders to measure NO₂ concentrations in front of their house using a low-cost sampler design
107 (Meysman et al. 2020). This measurement campaign was internationally unprecedented in terms of
108 coverage and spatial density: 20.000 sampling kits containing NO₂ diffusion samplers were distributed
109 (~1% of all households in Flanders), thus allowing measurements across a wide urbanized region (~
110 250 km x 50 km) at high spatial density (~1.5 sites on average per km²). The resulting extensive dataset
111 is used here for a detailed validation case study of the ATMO-Street model chain.

112
113 We develop a generic three-step methodology to validate and optimize the model chain by means of
114 the CurieuzeNeuzen dataset. Firstly, the original ATMO-Street model chain is validated against the
115 NO₂ data using validation plots and statistical techniques. The extensiveness of the measurement
116 dataset allows us to perform an in-depth model performance analysis by evaluating the
117 concentrations based on different aspects (type of location, concentration class etc.). In the second
118 step, we introduce improvements and optimizations to the model chain based on the findings of the
119 validation. The effectiveness of the optimization is again verified by validating the results of the
120 optimized model chain against the NO₂ data. Finally, we evaluate the remaining discrepancies
121 between the modelled and measured concentrations and provide an outlook for further improvement
122 of air quality modelling. In this last step, we pay special attention to the ability of ATMO-Street to
123 capture the short-scale spatial variation of the NO₂-concentrations.

124

125 2. Methods

126 2.1. Measurement dataset

127

128 The measurement campaign will be only briefly summarized here, and is discussed in more detail in
129 (De Craemer et al. 2020b; Meysman et al. 2020). In the CurieuzeNeuzen Vlaanderen citizen science
130 campaign, 20.000 sampler kits were distributed to individual citizens, schools, companies, social
131 organizations and municipalities to measure outdoor NO₂ concentration at streetside locations. At the
132 front of the house (facing the street), two passive NO₂ samplers of the Palmes diffusion tubes type
133 were strapped to a real estate sign panel and attached to a window pane. This set-up standardized air
134 turbulence conditions near the Palmes tubes across all sampling points. Measurements were
135 conducted preferably on the first floor or otherwise ground floor to constrain the height effect on NO₂
136 concentrations.

137

138 A four-week measurement was performed from 11 AM April 28th to 1 PM May 26th, 2018. Note that
139 the duration of the Palmes tubes campaigns should be limited to approximately one month, to avoid
140 saturation of the tubes. The exact time period for the campaign has been chosen to maximize the

141 legitimacy of the validation results: background concentrations in May closely resemble the annual
142 mean background concentrations, and May is one of the few months without a long vacation period,
143 eliminating the need for time-specific corrections to the traffic data. The meteorological conditions
144 during the measurement period were, however, somewhat atypical. The average temperatures during
145 May 2018 were significantly higher than during a typical month May in the climatological baseline
146 considered by the National Meteorological Agency (1981 – 2010), with a profound gradient in the bias
147 from the west (coastline) to the east of the region (bias of approximately 3°C in the east, and
148 approximately 1.5°C at the coastline). Moreover, there has been much less precipitation (30% less on
149 average), much smaller wind speeds, and also the prevailing wind direction was clearly different.
150 During May 2018, the prevailing wind directions were north-north-west and north-east, while on
151 average winds from the southwest are dominant in Flanders.

152

153 Duplicate samplers showed good precision (root mean square error 1.7 µg/m³ between replicates,
154 relative standard deviation <5%). These raw NO₂ data were calibrated by simultaneous deployment of
155 passive samplers at 24 EPA reference monitoring stations dispersed across the measurement region,
156 and averaged across the two duplicates, thus resulting in mean NO₂ concentration over the 4-week
157 measurement period. After a quality control, 17886 measurement locations were retained for the
158 model validation campaign (Meysman et al. 2020). Assuming errors are random and uncorrelated,
159 the addition of the standard deviations of the passive sampler measurement (1.7 µg/m³) and
160 calibration (2.2 µg/m³) resulted in a total standard deviation of 3.9 µg/m³, thus providing a relative
161 uncertainty of 10% at the WHO-guideline value of 40 µg/m³.

162

163 2.2. The ATMO-Street model chain

164 2.2.1. General overview

165

166 Street-level nitrogen dioxide concentrations are modeled using a model chain that captures the
167 different scales of urban air quality. The ATMO-Street model chain (Lefebvre et al. 2013b) consists of
168 the land-use based interpolation model RIO determining background concentrations (Janssen et al.
169 2008a), the bi-gaussian plume dispersion model IFDM accounting for the impact of local emissions
170 from traffic and industry (Lefebvre et al. 2011), and the street-canyon module OSPM that calculates
171 the in-street increment resulting from street-canyon effects (Berkowicz et al. 1997). Road traffic
172 emissions are computed by the traffic emission model FASTER (Veldeman et al. 2016). The model
173 chain calculates hourly concentrations at a number of irregularly spaced receptors, which are
174 subsequently gridded to a regular raster with a 10m resolution. The flowchart of the model chain is
175 provided in Figure 1.

176

177 For verification purposes, the simulations by the full ATMO-Street model chain were compared to
178 versions that use only part of the model chain. In one type of sensitivity analysis, only the background
179 concentrations from the RIO-model were considered, thus evaluating the predictive capability of only
180 using wide-scale land use regression. In another sensitivity analysis, we used the RIO-IFDM
181 combination, which combines the background concentrations with local contributions from traffic and
182 industry, but neglects the street-canyon increment. The remainder of this section explains the three
183 model components and their coupling in more detail.

184

185 2.2.2. Components

186

187 FASTER is a traffic emission model that calculates geographically explicit emissions for road
188 transport, based on (1) emission factors (i.e. emissions per vehicle type per speed per kilometer), (2)

189 fleet data (i.e. number of vehicles and mileages) and (3) mobility data (i.e. vehicle counts on a
190 network). Emission factors were obtained from region specific calculations with the COPERT-tool,
191 which is EU-wide used to calculate emission inventories for road transport (Ntziachristos et al. 2009).
192 FASTRACE calculates yearly total emissions for each road segment, which are subsequently combined
193 with daily, weekly and monthly traffic intensity profiles, to obtain hourly emissions for each road
194 segment.

195
196 Background concentrations are modelled using RIO (Hooyberghs et al. 2006; Janssen et al. 2008b), a
197 land use regression model for the interpolation of hourly pollutant concentrations as measured by the
198 official telemetric network. The model is based on a residual kriging interpolation scheme using a land
199 use derived covariate. A polynomial regression determines the statistical relationship (trend
200 functions) between the long-term averaged concentrations at each hour of the day and the underlying
201 land use parameter. RIO produces hourly concentration maps for NO₂, NO, and O₃ on a 4x4 km² grid,
202 which are subsequently used as background concentrations for the IFDM and OSPM components of
203 ATMO-Street chain.

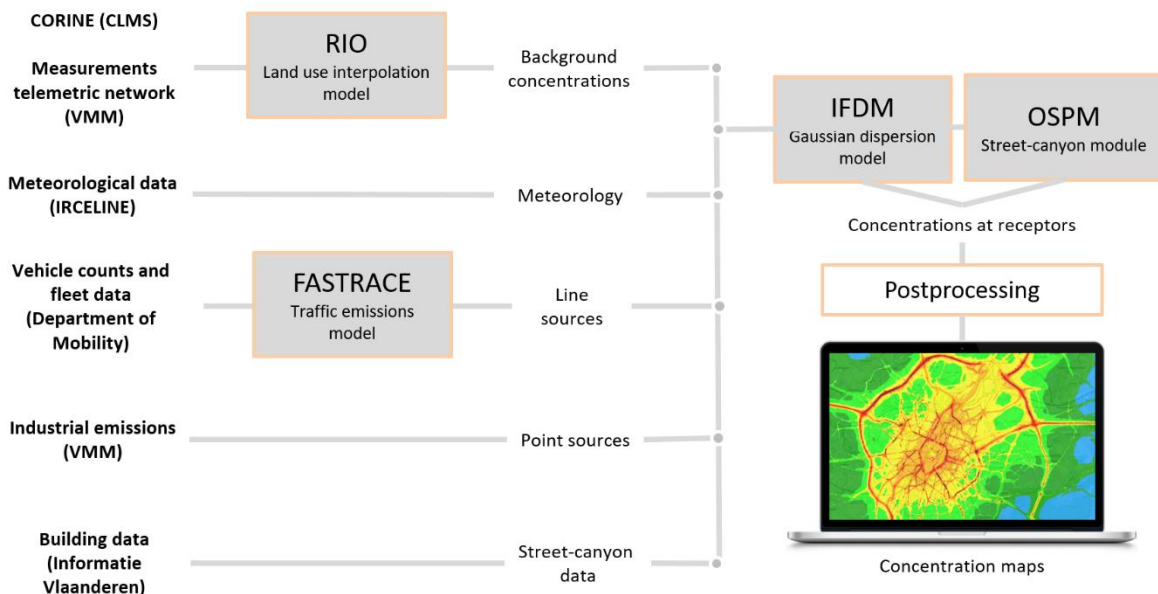
204
205 Local open-street concentrations due to traffic emissions and industrial point sources are modelled by
206 the bi-Gaussian plume model IFDM (Immission Frequency Distribution Model) (Lefebvre et al. 2013a).
207 IFDM is a receptor grid model: air pollutant concentrations are computed for an abundance of
208 receptor locations. Instead of a regular grid, we use a point source and road-following grid. This
209 approach ensures that more receptor points are available where the largest concentration gradients
210 are expected (Lefebvre et al. 2011). Since the model uses an hourly time resolution, we assume that
211 the chemical equilibrium in the NO_x-O₃ reaction is reached. We take this chemical reaction into
212 account using the fast-ozone-chemistry scheme (Berkowicz et al. 1997; Berkowicz et al. 2008), which
213 relies on temperature and solar height data. To avoid double-counting of the emission sources, a
214 specific coupling between the regional model and the urban-scale model has been developed
215 (Lefebvre et al. 2011).

216
217 To calculate the effect of buildings on the street level concentrations, the IFDM model is coupled to
218 the Operational Street Pollution Model (OSPM) (Berkowicz et al. 1997; Ottosen et al. 2015; Jensen et
219 al. 2017). Street level concentrations due to road traffic emissions are calculated using a combination
220 of a plume model for the direct contribution and a box model for the recirculating part of the
221 pollutants in the street. In the current set-up for OSPM, a receptor location is placed every 20m on
222 each road with a row of buildings adjacent to the road (i.e. at a maximum distance of 50m to the
223 middle of the road). The concentrations at the receptor locations of the IFDM and OSPM models are
224 eventually combined and gridded via a three-step postprocessing module. At first, IFDM results are
225 gridded using Delaunay triangulation to obtain gridded open street concentrations. Secondly, we grid
226 the OSPM results using nearest-neighbour interpolation. In the final step, both gridded maps are
227 combined into a map with a 10m resolution, by using the OSPM results at locations where buildings
228 are adjacent to the road, and the IFDM results at all other locations.

229
230 A priori, we expect large deviations between the measurements and the modelled data for the
231 background model RIO. Because of the coarse resolution, there will be a lot of scatter, a large
232 underestimation of the results (especially close to busy roads) and not much correlation between the
233 measurements and model values. Adding the Gaussian dispersion model IFDM should improve the
234 results, especially for open locations, which should also significantly improve the scatter and the
235 correlation. However, the RIO-IFDM model chain neglects the recirculation of pollution at locations
236 with buildings adjacent to the road, hence a large bias is still expected. Adding the OSPM module
237 should resolve this issue, but it also increases the susceptibility of the model chain to input errors.
238 Because the concentration field at locations with recirculation is very sensitive to many parameters
239 describing the setting (traffic emissions, vehicle speed and numbers influencing the traffic-induced

240 turbulence, detailed building configuration in the immediate surroundings of the location) and many
 241 of these parameters are only approximatively known, we expect a lot of scatter for the locations where
 242 the OSPM model is applied.

243
 244



245
 246 *Figure 1: Flowchart of ATMO-Street model.*

247

248 **2.2.3. Set-up for the validation campaign**

249

250 In this study, the ATMO-Street model chain was applied to the same 4-week period as the citizen
 251 science measurement campaign. Input data stem from official datasets of the regional authorities. The
 252 regional background model RIO has been set up using the data from the telemetric network of the
 253 Flanders Environment Agency (VMM) and the Corine Land Cover of the Copernicus Land Monitoring
 254 Service (CLMS) as land-use input. Vehicle fleet and traffic data for the major roads in Flanders are
 255 provided by the Flemish Department for Mobility. Minor roads are only sparsely represented in these
 256 traffic data, and these roads are thus not considered in the present air quality assessment. There are
 257 also some known issues with the traffic data for urban locations, as recent mobility plans (e.g. low-
 258 traffic zones in city centers) are not always correctly represented. Point sources stem from the official
 259 emission inventory for industry of VMM. Building data has been retrieved from the official building
 260 dataset for Flanders (Informatie Vlaanderen).

261

262 The Gaussian dispersion model internally computes stability classes based on the Bultynck-Malet
 263 parametrization (Bultynck and Malet 1972), and thus only requires surface temperature, wind speed
 264 and wind direction as meteorological input. These parameters have been composed by the Belgian
 265 Interregional Environment Agency (IRCEL - CELINE) by assimilating Copernicus C3S ERA5 reanalysis
 266 data (Copernicus Climate Change 2017) with measurements at several meteorological stations,
 267 yielding surface wind and temperature fields with a 1km resolution. Because the model uses input
 268 data for the actual time frame of the measurements, we do not expect an influence of the atypical
 269 meteorological conditions during the measurement period on the final conclusions of the study.

270

271

272 The coordinates of the measurement locations were recorded with a precision of 2 meter. The
 273 corresponding model concentrations are determined by the concentration for the pixel (10m x 10m)
 274 in the gridded map that contains the measurement location. Note that in this way, the coordinates of
 275 the measurement locations are not used when defining the receptor grid. Model results are always
 276 reported at 1.5m height, even in cases where the measurements were done at higher locations.
 277 Differences between the measurements and the model result at their location in this manuscript thus
 278 include uncertainties in the measurements, errors in the model input data, model errors and errors
 279 from the postprocessing.
 280

281 3. Model optimization

282 3.1. Initial validation

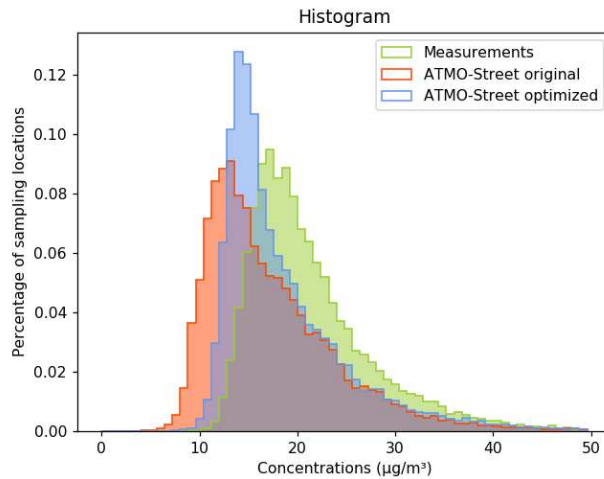
283
 284 Firstly, we focus on the validation statistics of the original setup of the ATMO-Street model chain
 285 before optimization (the so-called “original model”). Table 1 provides the validation statistics obtained
 286 by comparing the Curieuzeneuzen data with the model values (see the appendix for a mathematical
 287 definition of the statistical quantities). The Pearson correlation coefficient (0.58) points at a
 288 reasonable correlation between the measurements and the model results, and is in line with results
 289 obtained in previous validation campaigns (Lefebvre et al. 2013b). The bias of the original model is
 290 substantial and negative (-4.1 $\mu\text{g}/\text{m}^3$, -20%). This indicates that the model underestimates the NO_2
 291 concentrations in general, which is also reflected by the shift in histograms (see Figure 2). Especially
 292 for the lower concentrations (<25 $\mu\text{g}/\text{m}^3$), the modelled distribution is shifted to lower concentrations
 293 in comparison with the measured contribution. The other statistics, such as the Bias Corrected Root
 294 Mean Square Error (BCRMSE 4.6 $\mu\text{g}/\text{m}^3$) and fraction of model values within a factor of two of the
 295 observation (Fac2: 99%) are more in line with the results of previous validation studies.
 296

297 *Table 1: Validation statistics for the original and the optimized ATMO-Street model chain. The*
 298 *validation statistics for the separate buildings blocks of the optimized model chain (RIO and RIO-IFDM)*
 299 *are also provided. Statistics include the bias, the Root Mean Square Error (RMSE), the bias-corrected*
 300 *RMSE (BCRMSE), the Pearson R^2 coefficient, the FAIRMODE model quality indicator (MQI) and the*
 301 *fraction of model values within a factor of two of observations (Fac2). A mathematical definition of the*
 302 *statistical quantities is provided in the Appendix.*

303

Statistic	ATMO-Street Original model	ATMO-Street Optimized model	RIO-IFDM Optimized model	RIO Optimized model
Bias ($\mu\text{g}/\text{m}^3$)	-4.1	-2.7	-4.3	-4.5
RMSE ($\mu\text{g}/\text{m}^3$)	5.5	5.2	6.2	7.0
BCRMSE ($\mu\text{g}/\text{m}^3$)	4.6	4.4	4.4	5.3
Pearson R^2	0.58	0.58	0.51	0.33
MQI	0.96	0.80		
Fac2 (%)	99.0	99.4	98.0	96.2

304



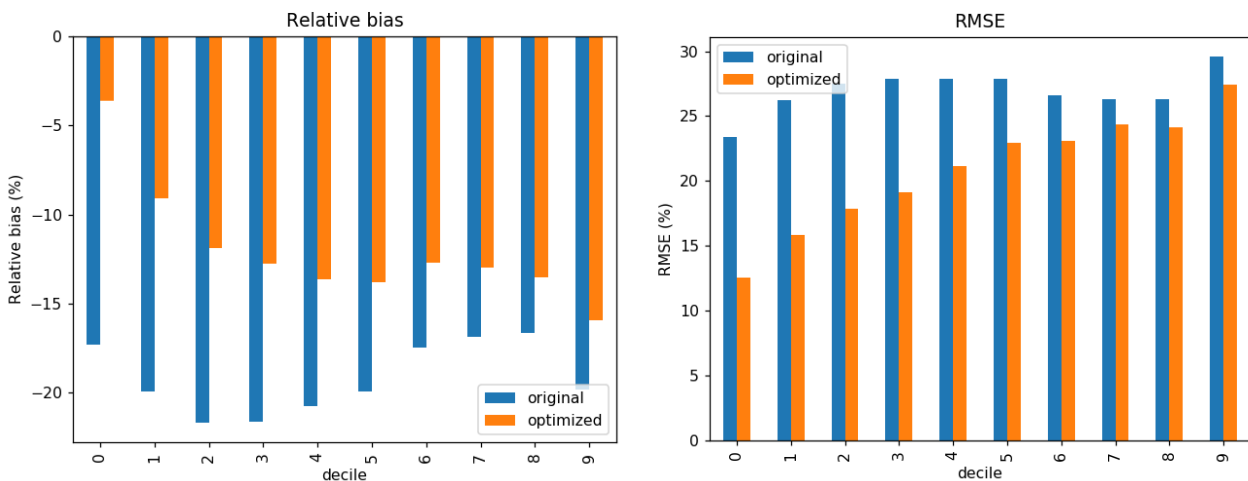
305

306 *Figure 2: Comparison between measured and modelled NO₂ concentrations at the 17.886*
 307 *measurement locations with high quality data. Histograms for the measurement data (green), the*
 308 *updated model chain (blue) and the original model chain (red).*

309

310 Additional insight into the discrepancy between modelled and measured concentrations is obtained
 311 by considering the bias and RMSE per concentration class. For this purpose, we grouped the locations
 312 in ten classes according to the deciles for the measured concentrations (Figure 3). Apart from the 10th
 313 decile, the original model shows the highest relative biases (up to -22%) for the lower deciles. Similarly,
 314 the (relative) RMSE is larger for the third to sixth decile than for the seventh to ninth decile. As
 315 locations with higher-than-average concentrations are the most sensitive to issues with the Gaussian
 316 dispersion model or the street canyon module (or one of their input datasets), which would introduce
 317 larger (relative) deviations for the higher deciles, these findings therefore point at issues with the
 318 background model, which underestimates the background concentrations for the lower deciles. If we
 319 plot the bias across the spatial domain (Figure 4), we indeed observe that underestimations are mainly
 320 occurring in rural locations (i.e. in the less densely populated areas), whereas the bias is much smaller
 321 for the urban locations. Finally, this underestimation of the background concentrations is also
 322 observed in the histogram (Figure 2).

323

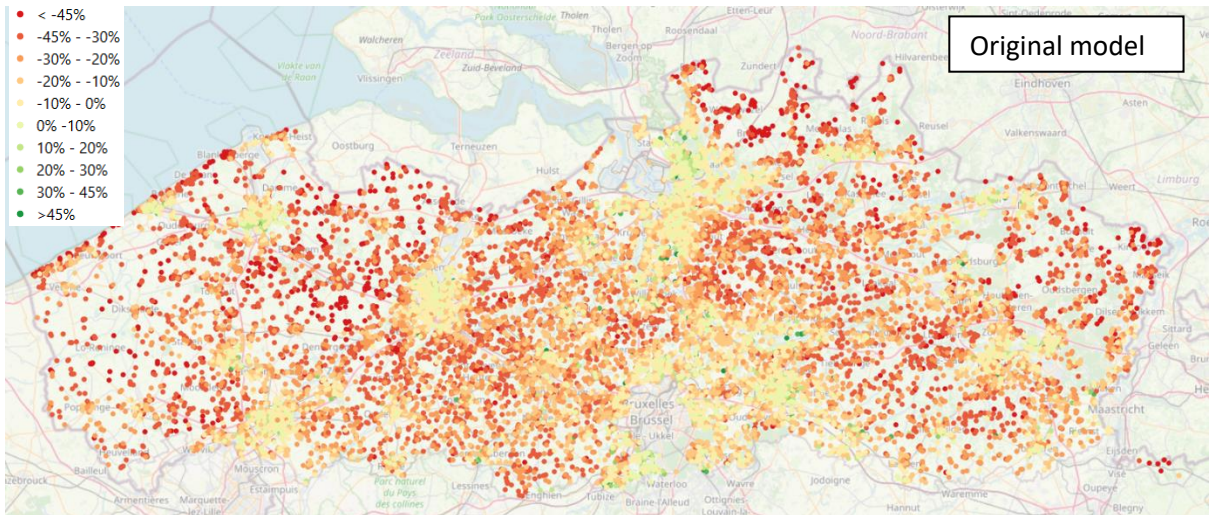


324

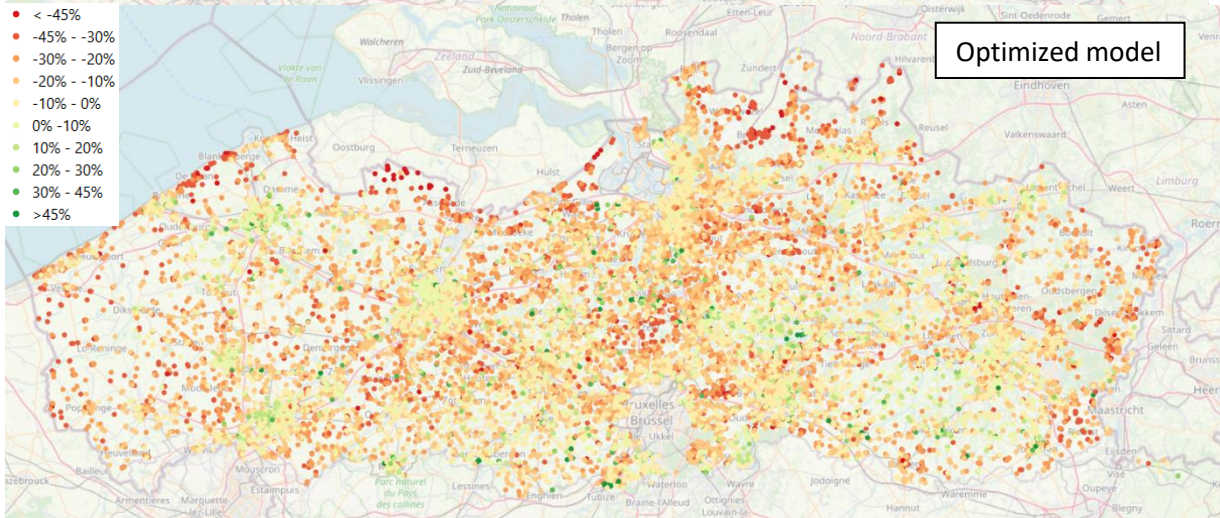
325 *Figure 3: Comparison between measured and modelled NO₂ concentrations per concentration class (10*
 326 *deciles of the measured concentrations – decile 0 contains lowest concentrations). Relative bias (left)*
 327 *and RMSE (right) per decile. The panels provide the results for the original model chain (blue) and the*
 328 *updated model chain (orange).*

329

330



331



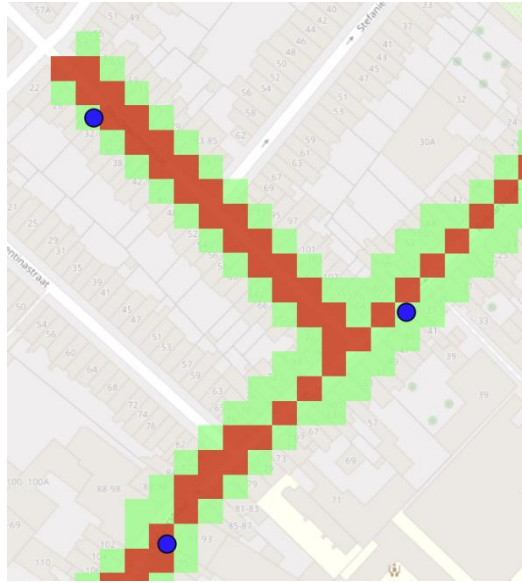
332
333

334 *Figure 4: Map of the relative difference between the measured and the modeled concentrations (in %)*
 335 *across the region of Flanders (17.886 measurement locations with high quality data). Negative (red)*
 336 *values indicate model underestimation, positive (green) values signify model overestimation. The top*
 337 *panel shows results for the original model, the bottom panel for the optimized model.*

338

339 An in-depth analysis of the deviations between models and measurements uncovered a second issue
 340 with the original model, which concerned to the coupling of the Gaussian dispersion model (IFDM)
 341 and the street-canyon module (OSPM). When gridding the final concentration maps (i.e. during the
 342 postprocessing as discussed in Section 2.2), the grid cells with street-canyon module increments do
 343 not always correspond to the street side location where citizens put their diffusive samplers. This was
 344 especially apparent for streets that are diagonal with respect to the north-south axis, as illustrated in
 345 Figure 5. These locations are therefore incorrectly assigned the results of the Gaussian dispersion
 346 model, instead of the results of the street-canyon module. Although this situation only occurred at a
 347 limited number of locations, large underestimations were obtained at these locations, which hence
 348 significantly influence the bias for the largest decile in Figure 3.

349
350
351
352
353



354
355

356 *Figure 5: Illustration of the issue concerning the coupling between the Gaussian dispersion and the*
 357 *streetcanyon model. In the original model chain, the streetcanyon concentrations are only used for the*
 358 *red colored grid cells. For some of the sampling locations (blue dots), the Gaussian dispersion results*
 359 *are hence applied. In the optimized model, the streetcanyon contribution is also used for the green grid*
 360 *cells, and hence for all three sampling locations in the domain of the figure.*

361

362 Note that the two issues discussed above (background underestimation, incorrect street canyon
 363 postprocessing assignment) could only be detected due to the extensiveness of the Curieuzeneuzen
 364 dataset. While ATMO-Street has previously been validated with smaller datasets, this analysis has
 365 been unable to reveal the background concentration issue, as a large and spatially widespread dataset
 366 is required for the type of analysis presented in Figure 3 and Figure 4. Additionally, the postprocessing
 367 issue was only observed for a small number of locations, and so the issue can only surface in suitably
 368 large datasets (the probability to include such locations in a dataset increases with the sampling size).
 369

370 3.2. Model optimization

371

372 Guided by the results of the initial validation, model optimizations were implemented. A first
 373 improvement targeted the rural background concentrations, as the results of the Curieuzeneuzen
 374 campaign clearly indicated an overestimation at these locations in the original model.
 375

376 After a detailed analysis, relying on land use data for the Curieuzeneuzen sampling locations, we found
 377 that the problem was linked to the land-use parameterization that is used in the RIO module. In this
 378 module, NO₂ data are assimilated from reference stations of the environmental monitoring agencies
 379 across the whole of Belgium. Yet, there is a strong north-south difference in urbanization in Belgium,
 380 which makes that there are relatively few reference stations in rural areas of Flanders (the northern
 381 part of Belgium). As a result, the land use parameterization applied in the original RIO module was
 382 heavily influenced by the observations at EPA reference stations in rural areas of Wallonia (the
 383 southern part of Belgium). Since background NO₂ values are lower in Wallonia (which is less densely
 384 populated and less industrialized), this caused an underestimation of background concentrations in
 385 Flanders, which was uncovered for the first time thanks to the Curieuzeneuzen sampler data.
 386

387 Guided by the citizen science data, the RIO module was adapted by improving the parameterization
388 of the different rural land use classes, yielding an optimized relation between the concentrations and
389 the land use parameters (trend function). These optimizations principally consisted of a decoupling of
390 the rural land uses classes (namely forests, natural areas and arable land) in the northern urbanized
391 part and southern non-urbanized part of Belgium. Note that the finetuning required the availability of
392 an abundance of measurement data at many rural locations with different land uses in their
393 surroundings, and the Curieuzeneuzen measurements have thus been indispensable.

394

395 The citizen science data were only used to determine an improved land use parameterization, but
396 they are not directly used in the spatial interpolation itself. The model results hence remain
397 independent of the measurements, and thus independent validation of the optimized model using the
398 citizen science data is still possible.

399

400 A second correction targets the incorrect street canyon postprocessing assignment, by adjusting the
401 GIS-tools that determine the locations where the street-canyon concentrations are applied. To this
402 end, we modified the parameter that determines the maximal extent of the street canyon
403 concentrations (expressed as the distance to the middle of the road), to make sure the OSPM results
404 are used for all locations where the concentration is significantly influenced by the presence of the
405 buildings. Moreover, in the optimized model version, the concentrations of the OSPM street-canyon
406 module are now also used for half-open locations, i.e., for roads with a continuous row of houses at
407 one side of the street¹. Due to these two modifications, the (higher) street-canyon contributions are
408 attributed to more sampling locations, leading to an increase in the mean NO₂ concentration across
409 all sampling locations.

410

411 3.3. Analysis of the optimizations

412 3.3.1. Basic analysis

413

414 Table 1 provides the validation statistics for the optimized model. Figure 2 shows the histogram and
415 Figure 3 depicts the bias and RMSE per decile. The optimized model outperforms the original model
416 in many aspects. The bias and RMSE are markedly lower for the optimized model. The largest
417 improvements in the bias and RMSE are observed for the lower deciles, as shown by Figure 3, and as
418 expected because of the nature of the optimizations. Moreover, also the Fac2 increases from 99% to
419 99.4%, which implies that the number of sampling sites for which the modeled data deviates more
420 than a factor of two of the observations decreases with 40% from 1% to 0.6%. The relative difference
421 map of the original and the optimized model (Figure 4) highlights moreover the reduced (relative) bias
422 in rural locations. For many of the locations in the rural areas, the relative difference between the
423 modeled and measured data is reduced to less than 10%.

424

425 To facilitate the interpretation of these results, we compare the statistics of the validation reported in
426 this Paper (based on more error-prone citizen science data) with those of more traditional studies
427 (with more controlled measurements), for ATMO-Street and similar street-level models used for policy
428 support in Europe. Typically, the bias is somewhat higher for the study at hand, compared to a bias of
429 -0.7% for the ADMS-Urban map in London (Hood et al. 2018) and 2% for the DEHM/UBM/OSPM map
430 in Copenhagen (Jensen et al. 2017). On the other hand, the correlation, RMSE and Fac2 are more in
431 line with those found in the traditional campaigns (e.g. correlation 0.6 in Aarhus and 0.7 in
432 Copenhagen (Jensen et al. 2017), correlation 0.7 in London (Hood et al. 2018), and RMSE 6 µg/m³ in

¹ Because the size of the recirculation vortex is only dependent on the *upwind* building in the OSPM model, the model can also be used for locations with a continuous row of buildings at one side of the road.

433 Antwerp (Lefebvre et al. 2013b)). A detailed comparison is, however, complicated, as both the set-up
434 of the measurement campaigns and the model chains vary significantly among the studies.

435

436 The optimization does, however, not remove all differences between measurements and modeled
437 concentrations. The histogram (Figure 2) indicates that the underestimation for many locations with
438 low-to-middle concentrations is reduced, but has not completely disappeared, and that the
439 distribution of the model values for the optimized model still deviates from the distribution for the
440 measured values. In addition, also, the Pearson R^2 is the same for the optimized and the original model
441 (see Table 1), indicating that the correlation between the measured and modeled data does not
442 improve.

443

444 3.3.2. MQI

445 As an additional benchmark for the model quality, we focus on the model quality index (MQI) as
446 proposed by the Forum for Air Quality Modelling in Europe (FAIRMODE). This indicator describes the
447 discrepancy between measurements and modelling results linked to the RMSE (Thunis et al. 2012;
448 Pisoni et al. 2019; Janssen et al. 2020). The MQI is a quality indicator that is specifically designed to
449 assess the performance of a model as a policy support tool for official assessments and EU reporting.
450 The FAIRMODE model quality objective (MQO) states that air quality models can be used for official
451 assessment purposes if the MQI is less or equal to one.

452

453 The original ATMO-Street model just meets the MQO objective, as the MQI is equal to 0.96 (Table 1).
454 The ensuing model optimization however decreased the RMSE and bias, which resulted in a
455 substantial decrease of the MQI to 0.80, and the optimized model thus satisfies the objective with a
456 far greater margin, indicating that the optimized model is better suited for policy support.

457

458 3.3.3. Spatial variation

459 Street-level air quality models are designed to simulate street-level concentration fields with high
460 spatial resolution. We use a semi-variogram analysis to test how well different models represent the
461 spatial heterogeneity of the concentration field. The semi-variogram visualizes the degree of spatial
462 variation of a set of observations by quantifying the differences between observations at a given
463 distance through the semi-variance (Cressie 1992)

$$464 \gamma(h) = \frac{1}{2N(h)} \sum_{N(h)} (c_i - c_j)^2$$

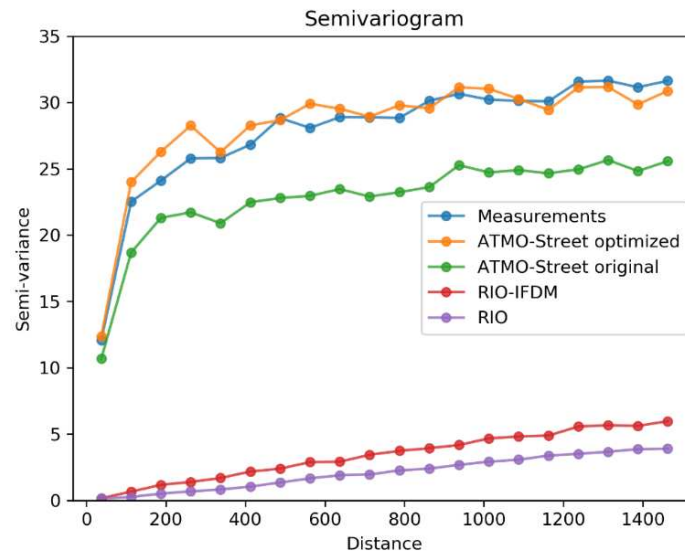
465 Here, the sum is over all pairs of locations that are a certain distance h apart, c_i and c_j are the
466 concentrations at these two locations, and $N(h)$ is the number of pairs that are considered. The semi-
467 variance $\gamma(h)$ quantifies the difference between two observations separated by a distance h , with
468 larger values indicating larger spatial variations. The technique is only effective if the (spatial)
469 resolution of the sampling dataset is in line with the actual spatial scale of the gradients in the
470 concentration field. An application for NO_2 concentrations thus requires a dataset with a dense
471 sampling, like the citizen science dataset analyzed here.

472

473 Figure 6 compares the semi-variogram for the citizen science data to the ATMO-Street model results,
474 for both the original and the optimized model chain. As NO_2 tends to vary over short spatial scales, we
475 only consider measurement locations that are less than 2 km apart. The figure highlights how the
476 short scale spatial variations are clearly better represented by the optimized model. The underlying
477 reason is the optimization in street canyon postprocessing (i.e., improvement of the coupling between
478 the Gaussian dispersion and the street canyon module), which yields a much better representation of
479 the spatial gradients on a local scale. The analysis shows how the optimized model chain adequately
480 captures the short-scale spatial variation of the concentration field, whereas the original model

481 underestimates the spatial variation. Although over- and underestimations of the spatial
 482 heterogeneity can still occur at specific locations, the mean heterogeneity of the NO₂-concentrations
 483 in Flanders is trustworthily explained by the optimized model chain.

484
 485
 486



487
 488 *Figure 6: Semi-variogram. The blue line indicates the semi-variance for the measurements of NO₂, while*
 489 *the other lines provide model results for NO₂ for three different types of models. The purple line*
 490 *provides results for the optimized RIO, the red line for the optimized RIO-IFDM and the orange line for*
 491 *the optimized ATMO-Street chain. The green line shows the results for the original ATMO-Street chain.*

492

493 3.3.4. Conclusions regarding the optimization

494

495 In summary, we conclude that the optimization resulted in substantial improvement of model
 496 performance, as substantiated by increased validation statistics, an improved MQI and a better
 497 representation of the short-scale spatial variations. This extensive analysis was made possible by
 498 exploitation of the large-scale data of the citizen science campaign. Although the optimization
 499 procedure presented here is specific to the ATMO-Street model, the underlying methodology and
 500 resulting conclusions are of wider interest for the air quality modelling community. Our “in-depth”
 501 validation of the ATMO-Street model relies on a statistical analysis that is applicable for any large-
 502 scale model validation, whereas the techniques to improve the RIO-model are applicable to any land
 503 use regression (LUR) model. Moreover, the observation that model shortcomings remain hidden when
 504 validation is done with limited data and only revealed through suitable large spatial datasets, is
 505 particularly relevant to the whole field of air quality monitoring and modelling.

506

507 Although the optimizations greatly improve many aspects of the model chain, they do not remove all
 508 differences between measurements and modeled concentrations, as e.g. indicated by the unchanged
 509 correlation coefficient and the updated histogram. In the next sections, we elaborate further on the
 510 validation of the optimized model and focus on the remaining discrepancies between the modelled
 511 and the measured concentrations.

512

513 4. In-depth validation of the optimized model

514

515 4.1. Analysis of the submodels

516

517 Environmental agencies use a range of different air quality model types for policy purposes, with
518 different spatial resolution. Some models are solely based on land use regression, while others include
519 more computationally intense approaches that explicitly include point and line emissions and simulate
520 the ensuing atmospheric dispersion of the emitted pollutants and / or account for street canyon
521 effects. The three different model components of the ATMO-Street chain reflect this cumulative
522 complexity and increasing spatial detail. To examine the importance of the different components,
523 Figure 7 and Table 1 compare the validation statistics of the full optimized ATMO-Street chain with
524 the background model only (RIO), and the combination of the background model with the Gaussian
525 dispersion model (RIO-IFDM, i.e. ATMO-Street without street-canyon increments).

526

527 The background model only substantially underestimates the measured concentrations (Figure 7). As
528 substantiated by the linear regression coefficient and the scatterplot, the highest concentrations are
529 particularly underestimated. The RIO model provides background concentrations on a 4 by 4 km
530 resolution, and the highest roadside peaks in traffic dense streets are clearly missed. The addition of
531 the Gaussian dispersion model IFDM considerably decreases the model-data discrepancy at these
532 locations, and, consequently, the correlation and linear coefficient substantially improve. There is
533 however still a significant bias, which is due to an underestimation of the street-canyon locations. Only
534 the complete ATMO-Street model chain appropriately captures the recirculation of the pollution at
535 these locations, yielding a much smaller bias.

536

537 The results for the RMSE, bias and correlation are in line with the expectations regarding the model
538 components (see Section 2.2.2). The (absolute) bias and RMSE are large and the correlation low for
539 the RIO-model. The RIO-IFDM model substantially improves on the RMSE and the correlation, but still
540 has a large bias. When OSPM is added, the bias and the RMSE further decrease, but the improvement
541 in RMSE is exclusively due to the decrease in bias, as indicated by the BCRMSE.

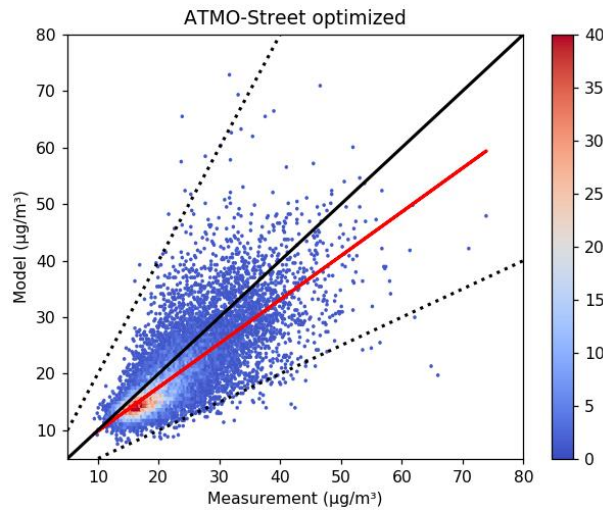
542

543 We conclude that only a model chain that takes the street-canyon increments explicitly into account
544 manages to adequately assess the NO₂-concentrations. These findings emphasize the importance of
545 the street-canyon contributions, and are in line with the results observed in other studies concerning
546 modelling of air quality at street-level scale (Vardoulakis et al. 2003; Lefebvre et al. 2013b; Jensen et
547 al. 2017).

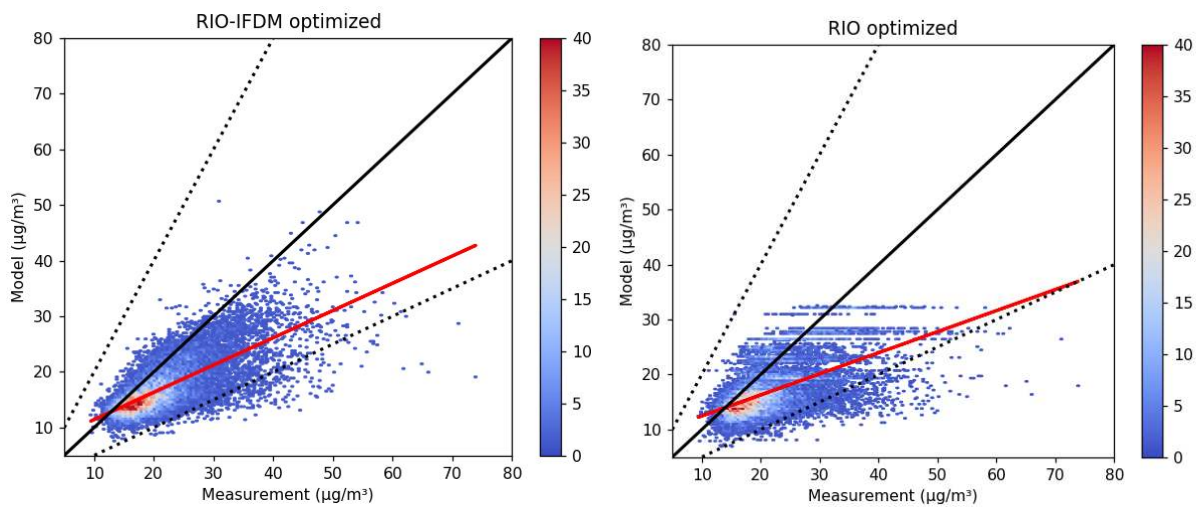
548

549

550



551



552

553 *Figure 7: Scatterplots showing the modeled concentration as a function of the measured*
 554 *concentrations for all sampling locations. Different panels depict the full ATMO-Street model chain*
 555 *(top), the Gaussian dispersion model (RIO-IFDM, bottom left) and the background model only (RIO,*
 556 *bottom right). To improve the visibility, the scatter points have been binned per $0.5 \mu\text{g}/\text{m}^3$. The*
 557 *colorscale indicates the number of points in each bin. The dashed lines indicate the upper and lower*
 558 *boundary of the interval $[0.5 * \text{measurements} ; 1.5 * \text{measurements}]$.*

559

560 We furthermore analyze the effect of the different submodels on the spatial variation of the modeled
 561 concentration field. In addition to the results of the full ATMO-Street chain, Figure 6 also shows the
 562 spatial variation modeled by the RIO and RIO-IFDM submodels. Clearly, the RIO-background model
 563 largely underestimates the spatial variation observed in the citizen science data. This is not
 564 unsurprising given the coarse resolution of the RIO model ($4 \text{ km} \times 4 \text{ km}$). The Gaussian dispersion
 565 model IFDM adds the open-street concentrations due to the road traffic and point sources, and as a
 566 result, the spatial variation increases. However, the semi-variance of the RIO-IFDM model still falls
 567 widely below that of the citizen science data. Adding the street canyon module OSPM greatly improves
 568 the representation of spatial variation: the spatial heterogeneity now closely approximates that
 569 observed by the measurements. The semi-variogram analysis thus demonstrates that street level air
 570 quality models like ATMO-Street are capable of capturing the general spatial heterogeneity of the NO_2
 571 concentration field, if street canyon increments are included in the model chain.

572

4.2. Breakdown by location type

To gain some further insight in the remaining discrepancies between the modelled concentrations and the measurements, and the impact of the input data on these, we analyze the validation for some specific location types.

First, we group the sampling locations based on the model that is applied at the sampler location. We divide all the locations in two groups: locations where RIO-IFDM is applied (labeled 'IFDM'), or locations where RIO-IFDM-OSPM is applied (labeled 'OSPM'). The former set are typically locations with isolated buildings, whereas the second set consists of (more complex) locations with a row of buildings adjacent to a road. The validation statistics for the full ATMO-Street model chain at these two groups of locations are provided in Table 2. The results indicate a slightly lower bias for the locations where OSPM has been used, but also a much larger scatter (RMSE) and much lower correlation for these locations, as expected, because of the larger sensitivity to input errors for the OSPM locations (see 2.2.2).

Table 2: Validation statistics for the optimized ATMO-Street model, with sampling locations clustered by location type. The table provides the bias, the relative bias, the relative bias-corrected RMSE (BCRMSE) and the Pearson R^2 coefficient. Columns 2 and 3 are related to the breakdown based on the model applied at the sampler location, columns 4 and 5 to the breakdown based on the availability of traffic data and the remaining columns to the breakdown based on the Flemish cities. More details on the binning are provided in the main text.

Statistic	IFDM locations (isolated building)	OSPM locations (multiple buildings)	Traffic data available	Traffic data unavailable	Antwerp	Ghent	Other Cities
Bias ($\mu\text{g}/\text{m}^3$)	-2.83	-2.43	-2.4	-2.9	-3.9	-1.1	-2.0
Relative Bias (%)	-14	-10	-10	-15	-12	-4	-8
Relative BCRMSE (%)	16	25	19	15	16	22	17
Pearson R^2	0.61	0.47	0.50	0.64	0.52	0.4	0.46

Secondly, we split the sampling locations according to the availability of traffic data for the nearest road to the measurement location. The traffic dataset contains traffic flows for a limited number of streets (the major roads). The (absolute value of the) bias is lower for the locations for which traffic data is available (absolute bias $-2.4 \mu\text{g}/\text{m}^3$, relative bias -10%) compared to the locations without known traffic data (bias $-2.9 \mu\text{g}/\text{m}^3$, -15%) (see Table 2). Note that the mean measured concentration is higher for the locations for which traffic data is available ($23.4 \mu\text{g}/\text{m}^3$ versus $19.6 \mu\text{g}/\text{m}^3$). As the relative bias increases with increasing concentration, we would expect the (relative) bias to be higher for the locations close to the roads. Since we observe the opposite, we definitely detect underestimations for sampling locations near roads where traffic data is lacking. Note, on the other hand, that the scatter is larger for the locations with traffic data (as quantified by a lower R^2 and larger BCRMSE). The underlying reason is the abundance of OSPM locations for the samplers in the vicinity of roads with traffic data (for 76% of the locations with traffic data OSPM has been used, while OSPM is not used for locations without traffic data). As the results in Table 2 indicate, the scatter is much

611 larger for the OSPM locations, which is also reflected in a larger scatter for the locations with traffic
612 data.

613

614 Finally, we make a comparison between cities in Flanders. We consider three groups of locations:
615 Flanders' largest city Antwerp (500.000 inhabitants; 1002 samplers), Flanders' second largest city
616 Ghent (250.000 inhabitants; 800 samplers), and samplers located in the other 8 largest cities (60.000
617 to 120.000 inhabitants; 2549 samplers). Validation statistics are provided in Table 2. The city of Ghent
618 stands out from the other. This is because a new mobility plan has been introduced in 2017, which led
619 to the introduction of new pedestrian streets, and modified traffic flows in the nearby streets, thus
620 altering traffic flows within the historical city center. However, the available traffic data do not (yet)
621 account for this new condition, and so the traffic data used for Ghent in the model set-up are less
622 accurate than those for other cities. The validation statistics reflect these shortcomings in the traffic
623 data. The BCRMSE in Ghent (22%) is higher than in Antwerp (16%) and the other cities (17%), while
624 similarly, the correlation coefficient in Ghent (0.40) is lower than in Antwerp (0.52) and the other cities
625 (0.46). When we only focus on the 200 samplers in the inner city of Ghent, where the largest impact
626 of the new circulation plan is observed, the validation statistics become even worse. The correlation
627 coefficient decreases to 0.27, and the relative BCRMSE increases to 25%.

628

629 4.3.Open issues

630

631 As the validation indicates, there are some remaining discrepancies between the modelled
632 concentrations and the measurements, indicating some room for further improvement of the model
633 and its input data.

634

635 An important issue concerns the quality of the mobility data that is used as input. Firstly, the traffic
636 dataset only contains traffic flows for a limited number of streets. The validation substantiates that
637 the NO₂-concentrations are, as expected, more adequately modeled for sampling locations near the
638 roads that are included in the traffic data. Furthermore, the spatial pattern of the traffic data is
639 outdated, which has an impact on the model quality for locations at which new mobility plans have
640 recently been introduced (e.g Ghent). These findings clearly highlight the importance of up-to-data
641 traffic data for air quality modelling at the local scale. Our analysis hence reveals that Environmental
642 Protection Agencies should invest in the collection of traffic data, and keep these datasets also up to
643 date, in order to support their air quality policies.

644

645 The statistics per city hint at a remaining issue with the optimized model, related to the urban
646 background concentrations in Flanders' largest city, Antwerp. The bias in the largest city, Antwerp
647 (-3.9 µg/m³, -12%), is significantly larger than the bias in Ghent (-1.1 µg/m³, -4%) and the other cities
648 (-2.0 µg/m³, -8%) (see Table 2). These results hint at a strong underestimation of the urban background
649 concentration in Antwerp. Note, however, that previous validation studies have not observed the
650 current underestimation: in a dedicated campaign focusing on Antwerp, a bias of -2 µg/m³ has been
651 observed (Lefebvre et al. 2013b), which is more in line with the bias observed in this work for the other
652 urban locations. Therefore, the underlying reason of the issue is unclear, as it could either be related
653 to the measurements (e.g. the calibration of the sampling results is mostly based on mid-range
654 concentrations, whereas higher concentrations are mainly observed in Antwerp), or the model set-up
655 (e.g. because the trend function of the land use regression model RIO may be unable to adequately
656 represent the concentration in the dense urban area in Antwerp).

657

658

659

660 5. Conclusions

661

662 We have validated and optimized the high resolution ATMO-Street air quality model chain using the
663 data of a large-scale citizen science measurement campaign. The extensiveness of the measurement
664 dataset allows us to perform an in-depth model validation and optimization. We have evaluated the
665 modelled concentrations by clustering the sampling sites by different aspects (type of location,
666 concentration class etc.), thereby paying special attention to the small-scale spatial variability of the
667 NO₂-concentrations. Optimizations guided by the data increased the model performance and
668 enhanced the capability of the model to correctly capture the spatial variation of the air pollution. The
669 ATMO-Street model chain attains the FAIRMODE model quality objective, substantiating that the
670 model is suited for policy support.

671

672 Our detailed model validation and optimization study reveals methodologies and insights that are of
673 wider importance for the air quality monitoring and modelling community. Foremost, it demonstrates
674 how the availability of an extensive spatial dataset enables a “deep validation”, which can result in
675 substantially improved model skill. Secondly, the validation also highlights the importance of the
676 street-canyon contributions. Only a model chain that takes the street-canyon increments caused by
677 the recirculation of pollution explicitly into account, manages to adequately assess the NO₂-
678 concentrations in Flanders. Thirdly, a model is only as good as the input it receives. Gaussian dispersion
679 models and street-canyon modules are very sensitive to the availability and quality of the traffic data.
680 Our analysis shows that the performance of the model chain is significantly reduced at locations where
681 the traffic flows are outdated or locations which lack traffic data. Therefore, in order to improve the
682 predictive power of street-level air quality models, a clear policy recommendation is to invest in the
683 collection of accurate, up-to-date traffic data across the whole road network (i.e. not solely focusing
684 on the major roads).

685

686 Finally, the most important lesson learnt is that street-level air quality models can substantially benefit
687 from a validation using a one-off widespread spatial monitoring campaign. Such a detailed and
688 rigorous validation of air quality models with large datasets is presently not a standard practice.
689 Currently, the monitoring strategy of environmental monitoring agencies is focused on capturing
690 temporal variability (i.e., high frequency monitoring at telemetric reference stations), while devoting
691 far less attention to a profound documentation of spatial variability. As a result, model validation
692 studies must typically focus on a small number of sampling sites. The analysis presented here, such as
693 the semivariogram analysis regarding the spatial variation of concentrations, however highlights the
694 importance of such large-scale measurement datasets with a high spatial resolution. In the case of
695 NO₂, such widespread spatial data collection is possible through mass-scale citizen science using low-
696 cost passive samplers. As such, citizen science offers not only a tool to increase awareness about air
697 quality, but also removes a critical bottleneck to ascertain and improve the quality of air quality
698 models.

699

700 Acknowledgements

701 This work was supported by funding for the citizen science project CurieuzeNeuzen Vlaanderen. We
702 thank Prof. R. Blust at University of Antwerp, M. Naert and I. Renson at the newspaper De Standaard
703 and Mr. M. Van Peteghem at Vlaamse Milieumaatschappij for enabling the CurieuzeNeuzen project.
704 Foremost, we gratefully thank all 20.000 citizens of the CurieuzeNeuzen project for their enthusiastic
705 participation and data collection.

706

707 Bibliography

- 708 Baker CJ, Hargreaves DM (2001) Wind tunnel evaluation of a vehicle pollution dispersion model. *J*
709 *Wind Eng Ind Aerodyn* 89:187–200 . [https://doi.org/10.1016/S0167-6105\(00\)00061-1](https://doi.org/10.1016/S0167-6105(00)00061-1)
- 710 Berkowicz R, Hertel O, Larsen S, Sørensen N, Nielsen M (1997) Modelling traffic pollution in streets.
711 *Natl Environ Res Institute, Roskilde, Denmark* 10129:20
- 712 Berkowicz R, Ketzel. M., Lofstrom P, Rordam H (2008) NO₂ chemistry scheme in OSPM and other Danish
713 models
- 714 Bo M, Salizzoni P, Pognant F, Mezzalama R, Clerico M (2020) A Combined Citizen Science—Modelling
715 Approach for NO₂ Assessment in Torino Urban Agglomeration. *Atmos* 2020, Vol 11, Page 721
716 11:721 . <https://doi.org/10.3390/ATMOS11070721>
- 717 Brusselen D Van, Oñate WA de, Maiheu B, Vranckx S, Lefebvre W, Janssen S, Nawrot TS, Nemery B,
718 Avonts D (2016) Health Impact Assessment of a Predicted Air Quality Change by Moving Traffic
719 from an Urban Ring Road into a Tunnel. The Case of Antwerp, Belgium. *PLoS One Accepted*:1–19
720 . <https://doi.org/10.1371/journal.pone.0154052>
- 721 Bultynck H, Malet LM (1972) Evaluation of atmospheric dilution factors for effluents diffused from an
722 elevated continuous point source. *Tellus* 24:455–472 . [https://doi.org/10.1111/j.2153-](https://doi.org/10.1111/j.2153-3490.1972.tb01572.x)
723 [3490.1972.tb01572.x](https://doi.org/10.1111/j.2153-3490.1972.tb01572.x)
- 724 Copernicus Climate Change C (2017) ERA5: Fifth generation of ECMWF atmospheric reanalyses of the
725 global climate. Copernicus Climate Change Service Climate Data Store (CDS).
- 726 Cressie N (1992) Statistics for spatial data. *Terra Nov* 4:613–617 . [https://doi.org/10.1111/j.1365-](https://doi.org/10.1111/j.1365-3121.1992.tb00605.x)
727 [3121.1992.tb00605.x](https://doi.org/10.1111/j.1365-3121.1992.tb00605.x)
- 728 Cyrus J, Eeftens M, Heinrich J, Ampe C, Armengaud A, Beelen R, Bellander T, Beregszaszi T, Birk M,
729 Cesaroni G, Cirach M, de Hoogh K, De Nazelle A, de Vocht F, Declercq C, Dedele A, Dimakopoulou
730 K, Eriksen K, Galassi C, Graulevičiene R, Grivas G, Gruzieva O, Gustafsson AH, Hoffmann B,
731 Iakovides M, Ineichen A, Krämer U, Lanki T, Lozano P, Madsen C, Meliefste K, Modig L, Mölter A,
732 Mosler G, Nieuwenhuijsen M, Nonnemacher M, Oldenwening M, Peters A, Pontet S, Probst-
733 Hensch N, Quass U, Raaschou-Nielsen O, Ranzi A, Sugiri D, Stephanou EG, Taimisto P, Tsai MY,
734 Vaskövi É, Villani S, Wang M, Brunekreef B, Hoek G (2012) Variation of NO₂ and NO_x
735 concentrations between and within 36 European study areas: Results from the ESCAPE study.
736 *Atmos Environ* 62:374–390 . <https://doi.org/10.1016/j.atmosenv.2012.07.080>
- 737 De Craemer S, Vercauteren J, Fierens F, Lefebvre W, Hooyberghs H, Meysman F (2020a)
738 *CurieuzeNeuzen: monitoring air quality together with 20.000 citizens*
- 739 De Craemer S, Vercauteren J, Fierens F, Lefebvre W, Meysman F (2020b) Using large-scale NO₂ data
740 from citizen science for air quality compliance and policy support. *Environ Sci Technol* 54:11070
741 . <https://doi.org/https://doi.org/10.1021/acs.est.0c02436>
- 742 EEA (2019) Air quality in Europe 2019 . [https://www.eea.europa.eu/publications/air-quality-in-](https://www.eea.europa.eu/publications/air-quality-in-europe-2019)
743 [europe-2019](https://www.eea.europa.eu/publications/air-quality-in-europe-2019). Accessed 10 Aug 2020
- 744 Faustini A, Rapp R, Forastiere F (2014) Nitrogen dioxide and mortality: Review and meta-analysis of
745 long-term studies. *Eur. Respir. J.* 44:744–753
- 746 Hoek G (2017) Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. *Curr. Environ.*
747 *Heal. reports* 4:450–462
- 748 Hoek G, Krishnan RM, Beelen R, Peters A, Ostro B, Brunekreef B, Kaufman JD (2013) Long-term air
749 pollution exposure and cardio-respiratory mortality: A review. *Environ. Heal. A Glob. Access Sci.*
750 *Source* 12:43
- 751 Hood C, Mackenzie I, Stocker J, Johnson K, Carruthers D, Vieno M, Doherty R (2018) Air quality
752 simulations for London using a coupled regional-to-local modelling system. *Atmos Chem Phys*
753 18:11221–11245 . <https://doi.org/10.5194/acp-18-11221-2018>
- 754 Hooyberghs J, Mensink C, Dumont G, Fierens F (2006) Spatial interpolation of ambient ozone
755 concentrations from sparse monitoring points in Belgium. *J Environ Monit* 8:1129 .
756 <https://doi.org/10.1039/b612607n>
- 757 Irwin A (2018) No PhDs needed: how citizen science is transforming research. *Nature* 562:480–482

758 Janssen S, Dumont G, Fierens F, Mensink C (2008a) Spatial interpolation of air pollution measurements
759 using CORINE land cover data. *Atmos Environ* 42:4884–4903 .
760 <https://doi.org/10.1016/j.atmosenv.2008.02.043>

761 Janssen S, Fierens F, Dumont G, Mensink C (2008b) Rio: A novel approach for air pollution mapping.
762 In: *Hrvatski Meteoroloski Casopis*. pp 172–176

763 Jensen SS, Ketzel M, Becker T, Christensen J, Brandt J, Plejdrup M, Winther M, Nielsen OK, Hertel O,
764 Ellermann T (2017) High resolution multi-scale air quality modelling for all streets in Denmark.
765 *Transp Res Part D Transp Environ* 52:322–339 . <https://doi.org/10.1016/j.trd.2017.02.019>

766 Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahuvaroglu T, Morrison J, Giovis C (2005)
767 A review and evaluation of intraurban air pollution exposure models. *J. Expo. Anal. Environ.*
768 *Epidemiol.* 15:185–204

769 Ketzel M, Berkowicz R, Lohmeyer A (2000) Comparison of Numerical Street Dispersion Models with
770 Results from Wind Tunnel and Field Measurements. *Environ Monit Assess* 2000 651 65:363–370
771 . <https://doi.org/10.1023/A:1006460724703>

772 Lefebvre W, Degraeve B, Beckx C, Vanhulsel M, Kochan B, Bellemans T, Janssens D, Wets G (2013a)
773 Presentation and evaluation of an integrated model chain to respond to traf fi c- and health-
774 related policy questions. *Environ Model Softw* 40:160–170 .
775 <https://doi.org/10.1016/j.envsoft.2012.09.003>

776 Lefebvre W, Van Poppel M, Maiheu B, Janssen S, Dons E (2013b) Evaluation of the RIO-IFDM-street
777 canyon model chain. *Atmos Environ* 77:325–337 .
778 <https://doi.org/10.1016/j.atmosenv.2013.05.026>

779 Lefebvre W, Vercauteren J, Schrooten L, Janssen S, Degraeuwe B, Maenhaut W, de Vlieger I,
780 Vankerkom J, Cossemans G, Mensink C, Veldeman N, Deutsch F, Van Looy S, Peelaerts W, Lefebvre
781 F (2011) Validation of the MIMOSA-AURORA-IFDM model chain for policy support: Modeling
782 concentrations of elemental carbon in Flanders. *Atmos Environ* 45:6705–6713 .
783 <https://doi.org/10.1016/j.atmosenv.2011.08.033>

784 Marshall JD, Nethery E, Brauer M (2008) Within-urban variability in ambient air pollution: Comparison
785 of estimation methods. *Atmos Environ* 42:1359–1369 .
786 <https://doi.org/10.1016/j.atmosenv.2007.08.012>

787 Meysman F, De Craemer S, Lefebvre W, Vercauteren J, Sluydts V, Dons E, Hooyberghs H, Van den
788 Bossche J, Trimpeneers E, Fierens F, Huyse H (2020) Citizen science reveals the spatial structure
789 of NO₂ traffic-related air pollution

790 Miranda A, Silveira C, Ferreira J, Monteiro A, Lopes D, Relvas H, Borrego C, Roebeling P (2015) Current
791 air quality plans in Europe designed to support air quality management policies. *Atmos Pollut*
792 *Res* 6:434–443 . <https://doi.org/10.5094/APR.2015.048>

793 Ntziachristos L, Gkatzoflias D, Kouridis C, Samaras Z (2009) COPERT: A European Road Transport
794 Emission Inventory Model. Springer, Berlin, Heidelberg, pp 491–504

795 Ottosen T-B, Kakosimos KE, Johansson C, Hertel O, Brandt J, Skov H, Berkowicz R, Ellermann T, Jensen
796 SS, Ketzel M (2015) Analysis of the impact of inhomogeneous emissions in a semi-parameterized
797 street canyon model. *Geosci Model Dev Discuss* 8:935–977 . <https://doi.org/10.5194/gmdd-8-935-2015>

799 Thunis P, Miranda A, Baldasano JM, Blond N, Douros J, Graff A, Janssen S, Juda-Rezler K, Karvosenoja
800 N, Maffei G, Martilli A, Rasoloharimahefa M, Real E, Viaene P, Volta M, White L (2016) Overview
801 of current regional and local scale air quality modelling practices: Assessment and planning tools
802 in the EU. *Environ Sci Policy* 65:13–21 . <https://doi.org/10.1016/j.envsci.2016.03.013>

803 Van Brussel S, Huyse H (2019) Citizen science on speed? Realising the triple objective of scientific
804 rigour, policy influence and deep citizen engagement in a large-scale citizen science project on
805 ambient air quality in Antwerp. *J Environ Plan Manag* 62:534–551 .
806 <https://doi.org/10.1080/09640568.2018.1428183>

807 Vardoulakis S, Fisher BEA, Pericleous K, Gonzalez-Flesca N (2003) Modelling air quality in street
808 canyons: A review. *Atmos Environ* 37:155–182 . [https://doi.org/10.1016/S1352-2310\(02\)00857-](https://doi.org/10.1016/S1352-2310(02)00857-)

809 9

810 Vardoulakis S, Solazzo E, Lumbreiras J (2011) Intra-urban and street scale variability of BTEX, NO₂ and
811 O₃ in Birmingham, UK: Implications for exposure assessment. *Atmos Environ* 45:5069–5078 .
812 <https://doi.org/10.1016/J.ATMOENV.2011.06.038>

813 Veldeman N, Maiheu B, Lefebvre W, Viaene P, Deutsch F, Janssen S, Vanhulsel M, Janssen L, Peelaerts
814 W, Driesen G, Van Looy S, Hooyberghs H (2016) Rapport activiteiten in 2015 uitgevoerd in kader
815 van de referentietaak 12 “Kenniscentrum Luchtkwaliteitmodellering”

816 WHO (2016) Ambient air pollution: A global assessment of exposure and burden of disease

817 Xie X, Semanjski I, Gautama S, Tsiligianni E, Deligiannis N, Rajan R, Pasveer F, Philips W (2017) A Review
818 of Urban Air Pollution Monitoring and Exposure Assessment Methods. *ISPRS Int J Geo-*
819 *Information* 6:389 . <https://doi.org/10.3390/ijgi6120389>

820

821

822

823 Appendix: definition of the validation statistics

824

825 In this appendix, we provide an overview of the validation statistics that have been used. We
826 henceforth assume that the difference between data set X (with values x_i) and data set Y (with values
827 y_i) is studied. $\langle . \rangle$ indicates the mean of a dataset, f.e. $\langle x \rangle$ is the mean of X.

828

- 829 • **Bias:** The bias indicates the relative difference between both data sets. Here the bias is
830 indicated relative to the mean of the measurements.

831

$$832 \text{Bias} = \langle x \rangle - \langle y \rangle$$

833

- 834 • **Root mean square error (RMSE):** The RMSE is the sample standard deviation of the
835 differences between predicted values and observed values. Both the absolute and relative
836 RMSE are used. The absolute RMSE is

$$837 \text{RMSE} = \sqrt{\langle (X - Y)^2 \rangle}$$

838 while the relative RMSE is

$$839 \text{RMSE} = \frac{\sqrt{\langle (X - Y)^2 \rangle}}{\langle x \rangle}$$

840

- 841 • **Bias corrected root mean square error (BCRMSE):** The BCRMSE is the RMSE of the unbiased
842 data sets.

$$843 \text{BCRMSE} = \sqrt{\langle ((X - \langle x \rangle) - (Y - \langle y \rangle))^2 \rangle}$$

844

- 845 • **The Pearson correlation coefficient** quantifies is a measure of linear correlation between two
846 sets of data. We always report the square of the Pearson coefficient, R^2 , where R is defined as

847

$$848 R = \frac{\text{COV}(X, Y)}{\sigma(X)\sigma(Y)}$$

849

- 850 • **Factor2 (FAC2):** the FAC2 indicates the percentage of modeled points that lies within a factor
851 two of the measured values, i.e. the percentage of data points that satisfies $\frac{1}{2}y_i < x_i < 2y_i$.

852

853