Reviewer #1

Comments from the reviewer are in blue, and answers in black (text citations and modifications are highlighted in italics). Note that following the recommendation of the other reviewer, we added three new meteorological features (surface net solar radiation, surface solar radiation downwards, downward UV radiation at the surface) and updated all the figures, tables and corresponding text. The impact on the results is relatively small so the discussion remains essentially the same.

The article under review here aims to quantify the impact of the Covid-19 lockdown measures in Spain on air quality. The topic is interesting from the point of view of air quality practitioners and the general public, but it also raises substantial scientific challenges. Even if economic activities were substantially reduced during the lock down period, the impact of meteorological factors on air quality precludes a simple comparison with previous years. Instead, the authors mobilize innovative machine learning approaches to tackle the issue. The quality of the presentation, scientific quality, and societal relevance are excellent, and publication in ACP is therefore recommended. I am nevertheless proposing the following minor suggestions that could help further strengthen the paper.

We are thankful to the reviewer for his/her positive feedbacks and comments.

General comment:

The authors should be encouraged to extent the coverage of their study. Applying the method over the whole of Europe is certainly the scope for another paper. But an extension of the temporal coverage up to the end of the lockdown in Spain would be interesting.

We agree that an extension over Europe is interesting, and we are currently collaborating on another study addressing the question at this larger scale (focusing on the largest European cities). Concerning the extension of the temporal coverage of the present study, we took into account the time period with data available at the time of preparation/submission of this study. Although it would have been nice to cover the entire period of the lockdown, we are here considering a period already quite extended (41 days), comprising the most stringent phase of the lockdown. To our opinion, although interesting, extending the study would require to substantially reshape the first draft, without bringing much more scientific knowledge. In addition, even at the time of this revision (August 25th), the situation cannot be considered as normal since many people across Spain are still working from home in Spain (and some parts of the country have been recently confined again).

Specific comments:

- L24, L403: the coronavirus is SARS-COV-2 not COVID-19. Indeed, the reviewer is right, according to the World Health Organization, COVID-19 designates the coronavirus disease, while SARS-COV-2 refers to the virus itself. To be consistent with this terminology, we added the term "*disease*" in the text.
- L36: without supporting reference, it is wiser to state that "the impact on industry is *presumably* more contrasted". Corrected.

- L50: in the motivation of the work, the authors could add that this type of analysis will serve to validate the model-based assessment using emission scenarios derived from activity data during the lockdown. We added in the conclusion : "*The results of the present study provide a valuable reference for validating similar assessments of the impact of the COVID-19 lockdown on air quality based on chemistry transport models and emission scenarios derived from activity data during the lockdown (e.g. Guevara et al., 2020a; Menut et al., 2020).*" with the corresponding references :
    - Menut, L., Bessagnet, B., Siour, G., Mailler, S., Pennel, R., and Cholakian, A.: Impact of lockdown measures to combat Covid-19 on air quality over western Europe, Science of The Total Environment, 741, 140 426, https://doi.org/10.1016/j.scitotenv.2020.140426, https://linkinghub.elsevier.com/retrieve/pii/S0048969720339486, 2020.
    - Guevara, M., Jorba, O., Soret, A., Petetin, H., Bowdalo, D., Serradell, K., Tena, C., Denier van der Gon, H., Kuenen, J., Peuch, V.-H., and Pérez García-Pando, C.: Time-resolved emission reductions for atmospheric chemistry modelling in Europe during the COVID-19 lockdowns (in review), Atmospheric Chemistry and Physics Discussions, https://doi.org/10.5194/acp-2020-686, 2020a.
- L69: where is the GHOST data available ? If GHOST database is not publicly open, the reference of the availability of the data should remain EEA's AQ e-reporting database. GHOST is a BSC internal on-going project currently not publicly available and a publication describing the dataset is in preparation. As explained in the text, GHOST is not another database, it ingests different air quality publicly available databases (including the EEA AQ eReporting database used in this study) and provides consistent and extended metadata to ensure the quality of the observational data. Although neglected by many studies, we consider that this quality assurance screening is an essential part of the data preprocessing. This is why we consider that it is worth mentioning and explaining it in detail in the manuscript, while to our opinion, the reference to the use of the EEA AQ eReporting database is already clear enough in the text.
- L75: the formal deadline for 2019 AQ e-reporting data to be delivered as E1a is September 2020, what is the fraction of 2019 data already E1a at the date of submission? Regarding the September deadline, it seems that many countries are actually delivering E1a data earlier (sometimes bit by bit through the year). We added the following text : "*The fraction of E1a data is 0% in 2020, 99% in 2019 and 100% in 2013-2018.*"
- L125: please clarify what you mean by "unique values", is the date index the Julian day, and if so why would it be unique? There is here a misunderstanding. As explained in L119, the date index is the number of days since 2013/01/01 (i.e. unique values going from 0 for 2013/01/01 to 2677 for 2020/04/30), while the Julian date (going from 1 to 365) is another feature. We added this to the sentence : "*Including such a feature with unique values (going from 0 for 2013/01/01 to 2677 for 2020/04/30) is not expected* […]"

- L145: hyperparameters should be defined and discussed either in the main text or in the annex. Further details would be appreciated in the annex on how the choice of those hyperparameters are related with the problem at hand (density and spread of observations, number and diversity of predictors etc.). The tuning strategy is explained in detail in Appendix C. The hyperparameters selected here are very common to any ML exercise with the gradient boosting machine and are not tailored to our specific problem. For each of these hyperparameters, we defined a reasonably large range of possible values to be tested through a randomized search, following again the idea we have about the common practices in the field (and the computational resources available for these calculations). We are not arguing here that this tuning strategy optimizes the best the performance but the performance obtained was found to be acceptable for the present study.
- L245: include the value of the uncertainty interval, it is difficult to compare percentages in 3.2 and ppbv intervals in 2.3.3. Actually, both should not be compared because they are not directly comparable. There is here a misunderstanding since the uncertainty intervals of Sect. 2.3.3 correspond to the uncertainties of the ML predictions at the daily and weekly scales (i.e. the uncertainties of the daily or weekly average NO2 concentrations).
- L255: the impact of the LEZ could actually be an increase of NO2 at stations in the outskirts of that zone. As also explained in our answer to the first reviewer, although the reviewer is right in principle, to our opinion, the 3 reasons already mentioned here in the text (namely the very limited area of this LEZ zone (5 km^2), the rather large distance to the stations selected and last but not least, the expected progressive transition to a new traffic pattern, given the absence of fines before April 1st, now postponed to September 15th 2020), combined together, reasonably justify our assumption that only a "*limited impact is expected*" in Madrid.
- Figure 2: N seems to be missing from the plot. Thanks, we corrected it (this was an old version of the legend).
- L266: clarify if the confidence interval is taken from the distribution of daily differences. We are not sure to properly understand what should be clarified here. The uncertainties used here correspond to the uncertainties at weekly scale (computed based on the differences between NO2 observations and predictions weekly averaged, as explained in Sect. 2.3.3). If the reviewer is talking about the uncertainties at daily scale, they are indeed obtained from the distribution of the daily differences.
- L325 and L344: could there be a role of background ozone in the relation between NOx emission changes and NO2 concentrations that would appear through this latitudinal gradient? The NO2 reductions obtained tend to be stronger in the southern half of Spain, but there is not a very clear latitudinal gradient that apply to all provinces. For instance, relatively lower NO2 reductions are found along the southern coast of Spain. Ozone and other chemical compounds may in principle impact the

NO2 concentrations (directly or indirectly) but we do not have any clear evidence for this at this stage.

- L365: clarify which reduction is for urban and traffic stations. We modified the text as follows : "*On average over this set of provinces, the NO2 reduction is -44 and -53% at the urban background and traffic stations, respectively* […]"
- L412: also mention day of the week in the predictors, which is presumably very important for NO2. We modified the sentence as follows : "*To tackle this issue, we used ML models fed by meteorological data and time variables (Julian date, day of week and date index) to estimate* […]"

Reviewer #2

Comments from the reviewer are in blue, and answers in black (text citations and modifications are highlighted in italics).

This work by Petetin et al., deals with the hot topic of variation of pollutants during the lockdown measures against the COVID19 pandemic. More specifically it focuses on the NO2 and the area of the Spanish state. Transports are the main source of NO2 in the troposphere, thus the reduction of traffic is estimated to lower significantly the emissions. Though the decrease of the emissions was very clear during the lockdown, the actual concentration in various areas is also dependent on meteorological parameters that rule the dispersion and the chemical processes of the gas. In order to better estimate the expected concentrations, based on meteorology, authors have trained a machine learning algorithm, to simulate the business as usual conditions, using as input meteorological variables. The work is generally well presented and should be accepted for publication in ACP after minor revisions.

We thank the reviewer for his/her constructive comments.

Specific comments

- L10 It would be better to provide some quantitative measure of the performance of the model. We modified the sentence : "*The ML predictive models were found to perform remarkably well in most locations, with overall bias, root-mean-squared error and correlation of +4%, 29% and 0.86.*"
- L77 Please provide some bibliographical reference for the uncertainty of these NO2 measurements. We added some information regarding the measurement uncertainties : "*All NO_2 measurements taken into account here are operated using chemiluminescence with an internal Molybdenum converter. Although predominantly used over Europe for measuring NO_2, this measurement technique is well known to be have strong positive artifacts due to interferences of NO_z compounds (e.g. nitric acid, peroxyacetyl nitrates, organic nitrates), especially during daytime when these species are photo-chemically formed, up to a factor of 2-4 as observed during summertime in urban atmospheres (e.g. Dunlea et al., 2007; Villena et al., 2012). In our case, the positive artifacts at urban background stations are probably lower since the period of study (late winter and early*

spring) is less photo-chemically active than summertime. Even lower interferences are expected at traffic stations where the NO_2/NO_x ratio is typically lower due to the proximity to fresh NOx emissions. In any case, the present study focuses on the relative changes of NO_2 due to the lockdown, so biases in the NO_2 measurements are of lower importance." with the corresponding references are :

  - Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., San Martini, F., Sheehy, P. M., Zahniser, M. S., Shorter, J. H., Wormhoudt, J. C., Lamb, B. K., Allwine, E. J., Gaffney, J. S., Marley, N. A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Ramos Villegas, C. R., Kolb, C. E., Molina, L. T., and Molina, M. J.: Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment, Atmos. Chem. Phys., 7, 2691–2704, doi:10.5194/acp-7-2691-2007, 2007.
  - Villena, G., Bejan, I., Kurtenbach, R., Wiesen, P., and Kleffmann, J.: Interferences of commercial NO2 instruments in the urban atmosphere and in a smog chamber, Atmos. Meas. Tech., 5, 149–159, doi:10.5194/amt-5-149-2012, 2012.

- L100 The selection of variables to feed the ML algorithm is very crucial and implies the physical and chemical processes that should be associated with the gas' concentration. My thought is that the photochemical cycle is implied by cloud coverage, which indirectly influences the irradiance which drive the photolysis. Since daily values are used, it is imperfectly fed to the algorithm, since nighttime cloud coverage would no affect NO2 concentration. Thus, some irradiance related variable from ERA-5 seems a better choice (SSI is a good one to investigate first). Since the results are satisfactory even using the cloud coverage proxy, I suggest to add some discussion on the selection of the variables and probable investigate other ones in the future. The reviewer here raises an interesting point, and we agree that including such information is susceptible to improve the ML-based predictions. We thus re-run our analysis adding the ERA5 surface net solar radiation, surface solar radiation downwards and the downward UV radiation at the surface to the set of features. The impact on the statistical results is generally positive although relatively small (error and correlation very slightly improved, and bias very slightly increased). On average, the importance of these new features is 4, 4 and 5%, respectively, which demonstrates their usefulness for predicting NO2 concentrations. We updated the entire document (figures, tables and text) with the results obtained with this new set of features. Note that most changes are minor, so the discussion remains the same. We thank the reviewer for helping us further improving the results.
- Figure 1. I think it is somehow difficult to understand the map, probably a different selection of color bar would make it easier to figure out the conditions. The *viridis* default color bar in Python *matplotlib* library presents a number of well recognized advantages over most of the existing color bars (e.g. color-blind friendly, perceptually uniform when printed in black and white). We thus decided to keep it but we modified the number of colors in order to make Figs. 1 and 6 easier to read.

- L119 ERA-5 spatial resolution is around 30km. Are there stations that correspond to the same grid point of the database? Please discuss the uncertainty introduced by the problem of non-colocation of ERA-5 and actual measuring stations. Given the ERA5 spatial resolution, urban background and traffic stations within a same city typically belong to the same ERA5 grid cell. We are not sure to perfectly understand the point raised here by the reviewer given that ERA5, as gridded data, can always be collocated with any measuring stations. After that, considering numerical meteorological data over a volume (the grid cell) as a proxy of the meteorological conditions occurring at a point (the air quality station) indeed necessarily comes with some uncertainties. The uncertainties (e.g. of representativeness) related to the relatively coarse resolution of ERA5 for representing accurately the meteorological conditions at the different stations are already discussed (L216-226 in the first version) in the initial manuscript, so we think that there is not much more useful information to add concerning this point.
- L130 Is that the case in any of the data used here? Are there any stations with significant trends in the training period? To our opinion, the 3-years training period is too short to compute meaningful trends. Over the period 2013-2019, a simple linear trend analysis on annual mean NO2 mixing ratios indicates that 21 over 75 stations show significant trends, with a median of -5%/year.
- L141 Following the arguments deployed in previous paragraphs, it seems preferable to test the validity in the same period of the year, as the one of interest (March-May), than in January -February. The reviewer is raising here an important point that deserves more discussion. In the revised version of the paper, we greatly reshaped Table 1 and the corresponding discussion.

  As explained in the text, at each station, several ML experiments have been conducted, including the reference one with training over 2017-2019 and testing in 2020 (hereafter referred to as the EXP_{2020} experiment), and the four other experiments based on past data and used for quantifying the uncertainties of our NO2 predictions (hereafter referred to as the EXP_{2016}, EXP_{2017}, EXP_{2018}, and EXP_{2019} experiments).

  Only the ML models obtained from the reference EXP_{2020} experiment are used for estimating the business-as-usual NO2 during the COVID-19 lockdown, which explains why we initially focused on them for the statistical evaluation. Since the lockdown period in 2020 can evidently not be used for evaluation, this constrained us to restrict the evaluation to the period 01/01/2020-13/03/2020. However, we agree that the performance of the ML models may be different during the lockdown period. In the revised version of the paper, we now also discuss the performance obtained with the four other experiments ($EXP_{2016\text{-}2019}$), which allows to check the performance during the period of the year of the lockdown. Besides Table 1, the text in this section is modified as follows :

  "*The performance of the ML predictions in each Spanish province and station type is shown in Fig. 2, and the statistics over all Spanish provinces reported in Table 1. Statistical results in Table 1 are given for both the*

reference ML experiment (EXP2020) and the other experiments combined together (EXP2016, EXP2017, EXP2018 and EXP2019, hereafter referred to as EXP2016–2019). Besides providing a broader view of the performance of our modeling strategy, considering these past experiments also allows assessing the performance of the ML predictions during the period of the year of the lockdown (14/03-30/04, for years 2016 to 2019), which may be important given the potential seasonality of prediction errors. Statistics obtained at urban background and traffic stations are given in Table A2 in Appendix. Results are evaluated using the following metrics, calculated based on daily NO2 mixing ratios : mean bias (MB), normalized mean bias (nMB), root mean square error (RMSE), normalized root mean square error (nRMSE) and Pearson correlation coefficient (PCC).

For information purposes, we included the statistical results obtained over the training dataset (2017/01/01-2019/12/31 in EXP_{2020}). Checking results over the training data may be useful for highlighting obvious situations of overfitting, when the performance is almost perfect. At both urban background and traffic stations, results show no bias, low nRMSE (always below 35%, 19% when considering all provinces), and a high PCC of 0.96. Similar results are obtained when considering the ensemble of all past experiments ($EXP_{2016–2019}$). Although such a performance obtained is very good, there are no clear signs of too prejudicial overfitting at this stage.

On the test dataset of the EXP_{2020} reference experiment (2020/01/01-2020/03/13, before the lockdown), the performance remains reasonably good in most provinces. Over all Spanish provinces, the nMB increases to +4%, the nRMSE to 29% and the PCC is reduced to 0.86, in very close agreement with the performance obtained with $EXP_{2016–2020}$ (nMB of +1%, nRMSE of 28% and PCC of 0.86). In comparison, the performance obtained in $EXP_{2016–2019}$ during the period of the year of the lockdown (14/03-30/04) is a bit lower but remains reasonable, with a nMB of +4%, a nRMSE of 37% and a PCC of 0.80. Although moderate, such a deterioration of the performance after mid-March might reflect some seasonality in the ML model errors and/or could be related to the presence of trends in the NO2 concentrations. Concerning this last point, as previously discussed in Sect. 2.3.2, including the date index feature in the ML model aims at limiting this potential issue but likely cannot completely solve it. Generally, only minor differences of performance are found between urban background and traffic stations.

Results of EXP_{2020} per province (Fig. 2) highlight some inter-regional variability of the performance, with poorer statistics in some provinces, at least for one type of station. At most stations, the bias remains below ±20% while nRMSE ranges between 15 and 45% (highest nRMSE around 50% in Teruel, Tenerife and Fuerteventura). Most provinces show PCC around 0.6-0.9, with only a few exceptions below 0.6 (urban background sites in Bizkaia, Fuerteventura, Huesca and traffic sites in Granada and Gran Canaria).” Note that we also added a Table A2 in the Appendix with detailed statistics on urban background and traffic stations.

- L159 Figure 1 shows that a number of stations have mean concentrations ~5ppvb. Thus these intervals are very huge, making the result not reliable. I suggest to present these intervals in a different way and not

averaging all that data. In this study, the uncertainties affecting our ML predictions are estimated using the most conservative approach, precisely in order to ensure the reliability of the NO2 reductions highlighted. These uncertainty intervals provided are indeed large but correspond to the uncertainties of the ML predictions at the daily scale (between January and April). Therefore, they cannot be compared to the (multi-) annual NO_2 averages shown for instance in Figure 1. As already explained in the manuscript, and as expected due to error compensations, the longer the time scale, the shorter these uncertainties. Therefore, the reviewer is here misleading his interpretation of the numbers provided in the text. We modified the sentence to avoid confusion : "*Averaged over all Spanish provinces, the uncertainty interval of ML predictions at the daily scale is [-5.1, +5.3] ppbv at urban background stations, and [-6.6, +6.7] ppbv at traffic stations.*" (Note that the uncertainty intervals are here slightly modified compared to the initial manuscript as they correspond to the results obtained with the extended set of features).

- L167-168 This argument is not clear. Please explain in detail. Here we simply mean that errors at the daily scale can at least partly compensate each other, which implies that averaging the ML-based predictions of daily NO2 mixing ratios to longer time scales (a week for instance) is expected to reduce the uncertainty. This is quite common, also for traditional chemistry transport models (reproducing the daily mean NO2 concentrations always goes with stronger uncertainties than the weekly, monthly or annual mean NO2 concentrations). We modified the sentence: "*These uncertainties are suited for our ML-based daily NO2 predictions. Because these daily uncertainties are likely at least partly uncorrelated, NO2 daily predictions averaged over periods longer than one day are expected to have smaller uncertainties due to error compensations.*"
- Table1 The test cases N seems very low, are these implying number of stations or total number of test days for all stations? Table 1 in the initial version of the manuscript gives the "*the statistics averaged over all Spanish provinces*", so the test cases N corresponds to neither the number of stations, nor the total number of test days, but the number of test days per station (on average over all stations). For each station in each Spanish province, training is performed over 2017-2019 (maximum N for training is therefore 3x365 = 1,095 points per station) and testing over 2020 before lockdown (maximum N for testing is therefore 31+28+14 = 73 points per station). In this Table, statistics were first computed for each station individually, and then averaged together to give the numbers provided in Table 1. Results at individual stations are still visible in Fig. 2.
  In the updated version of the manuscript, we greatly reshaped all this discussion, following a previous comment of the reviewer. Table 1 now gives the overall statistical results, computed over the entire data (i.e. combining all provinces together), which gives a broader view of the performance obtained by the ML-based predictive models.
- L255 In some cities, such zones, resulted in much higher traffic in peripheral road networks. Thus the stations at 3 and 9 km, might experiencing heavier traffic due to LEZ in the center. This should be answered locally by explaining the main routes and the traffic of each city.

Investigating in more detail the traffic pattern of Madrid is far beyond the scope of this paper. Although the reviewer is right in principle, to our opinion, the three reasons already mentioned here in the text – namely the very limited area of this LEZ zone (5 km^2), the rather large distance to the stations selected and last but not least, the expected progressive transition to a new traffic pattern, given the absence of fines before April 1st (and postponed to September 15th 2020 due to the COVID-19 situation (we added this new element of information in the revised manuscript : "*In our case, we expect a limited impact because the LEZ was still in its transition phase (strict enforcement through fines to offending motorists was not expected until April 1st and was finally postponed to September 15th 2020 due to the COVID-19 situation) and the two stations selected in Madrid province are located outside the LEZ (at 9 and 3 km from the city center).*") – combined together, reasonably justify our assumption that only a "*limited impact is expected*" in Madrid.

- L263 "Statistically significant" should not be used without proper definition and explanation. Explain which significance tests you used, what was the outcome and then provide such conclusions. Here we did not use any statistical test. Uncertainties of daily (weekly) NO_2 mixing ratios were computing empirically as the 5th and 95th percentiles of the daily (weekly) residuals obtained over past experiments. They are thus expected (by construction) to represent the 90% confidence interval. We modified the sentence : "*The uncertainty at weekly scale is here used as an estimate of the uncertainty at 90% confidence level (by construction, given that they are computed as the 5th and 95th percentiles of the weekly residuals, see Sect. 2.3.3) affecting the mean NO2 change*."
- 3.3 I think it is important to present some representative cases of other stations' time series in figures similar to 3 and 4. These provide a very clear picture of the conditions during the lockdown phases. Are there any periods of higher than business as usual concentration, probably in the stations with low mean values (Granada and Murcia probably)? Besides the time series for Madrid and Barcelona (Figs. 3 and 4), we are now providing the Supplement the time series obtained in all other Spanish provinces (Figs. S1-48), in order to allow the reader to check the results obtained in specific locations. Results obtained in the other provinces are generally consistent with those already discussed in Madrid and Barcelona. Thus, we do not think that it is particularly useful to present and discuss other cases in the manuscript.

  To answer the specific question of the reviewer, it is indeed possible to encounter observed NO2 concentrations higher to the ML-based business-as-usual concentrations on specific days, although it rarely happens. With the updated results obtained with the extended set of features, over all daily data available during the lockdown, only 4% (110 points over 2844) of the daily NO2 exceed the predicted business-as-usual NO2 estimates. Over these points, the observed NO2 mixing ratios are on average 1.3 ppbv higher than the business-as-usual (20% in relative). For information purpose, we included in the text: "*Results highlight that the reduction previously described in Madrid and Barcelona*

extends to most Spanish provinces, although with some inter-regional variability in the extent of the change and the degree of statistical significance. During the lockdown period, 96% (2734 points over 2844) of the observed daily NO2 mixing ratios are lower than the ML-based business-as-usual NO2 estimates.". Note that the corresponding observed NO2 mixing ratios are not particularly low since their average reaches 7.8 ppbv (compared to 5.4 ppbv for the entire NO2 observational dataset). Note also that additional information can already be found in Table 2 where we provided the maximum NO2 changes (among all provinces) during the three different phases and the entire lockdown period : in the revised version of the manuscript, you can see that the maximum NO2 changes (i.e. in our case, the changes closest to zero since values are negative) are all negative or close to zero (-14% during phases I+II+III for both urban background and traffic stations, -14 and -1% during phase I for urban background and traffic stations, respectively, etc.). This means that although observed NO2 can be higher than the business-as-usual NO2 on specific days, this is never the case along an entire phase (otherwise results would show some increases of NO2 during specific phases).

It is worth noting here that as explained in the manuscript, when selecting the stations, we required at least 10% of daily data during the entire lockdown period (41 days), which represents 4 days. However, we did not apply a similar criteria at the smaller scale of the individual lockdown phases. Although the data coverage in Madrid and Barcelona is very good, in some other provinces, the average NO2 reductions computed during specific lockdown phases can be based on very few data. This can now be seen in the Supplement. If we consider for instance the urban background station in Murcia, data are available during 7, 5 and 5 days in phases I, II and III, respectively (therefore quite well balanced). However, at the urban background station in Granada, data are available during 1, 1 and 9 days in phase I, II and III, respectively. More importantly, the only daily data available in phase I is on the first day of the phase (March 15th), i.e. at the very beginning of the lockdown, which likely explains the low increase of NO2 highlighted during phase I (see Fig. A1 in Appendix). The data coverage in these two provinces is almost complete for the traffic station. Over all Spanish provinces, largest data gaps during the lockdown period are found at background stations in Fuerteventura, Granada, Albacete, Alicante, Cuidad Real, Cádiz, Mallorca, Menorca, Murcia and Salamanca, and at traffic stations in Cádiz and Huelva.
We realize now that this can bring some confusion regarding the representativeness of the NO2 reductions highlighted in the paper. Therefore, in the revised version of the manuscript, we now require at least 3 days of available data during each lockdown phase. For computing the NO2 change during phases I+II+III, we required data available during at least 2 over 3 phases, to avoid cases where data is actually available only during one specific phase. As a consequence, some provinces during

specific lockdown phases have been removed in Figs. 5 and A1-A4. The overall discussion remains unchanged.

- 3.5 A figure showing all three time series (climatological, business as usual and measured) would be very useful, at least for some representative stations. Following the suggestion of the reviewer, we added the monthly climatological mean NO2 in the time series plots (Figs. 3, 4 and Figs. S1-48 in the Supplement), as well as the NO2 changes obtained with the climatological average approach in Figs. 5 and Figs. A1-A4 in the Appendix.
- L.384-387 This is a very important finding at should be highlighted more and included in the conclusions, because it is general for future application of climatological values. We added the following sentence in the conclusion : "*We also demonstrated the benefits of our meteorology-normalization approach compared to a simple climatological-based approach, especially at smaller temporal and spatial scales*."
- L445 It is not clear if all the flagged data were removed for the process or if different flags were treated differently. All the flagged data were indeed removed. We added a sentence at the end of this paragraph: "*All the corresponding measurements were removed from the dataset*."

Other modifications

Given the recent publication of a few new relevant studies on the topic (focusing on Spain), we updated some sentences in the manuscript :

- *"While such an extraordinary situation has obviously impacted the levels of air pollution in the country, as seen in both surface and satellite observations (Tobías et al., 2020; Bauwens et al., 2020), the extent of such reductions remains uncertain."*
- *"Actually, the lockdown offers unique opportunities for so-called dynamical CTM evaluations (Rao et al., 2011), i.e., testing the ability of CTMs to reproduce the observed changes of concentrations under unusually different emissions (Guevara et al., 2020a; Menut et al., 2020)."*
- *"A more detailed analysis of the activity data in these different emission sectors is required to better quantify how the emission forcing has been modified by the lockdown (Guevara et al., 2020a) and to understand the reductions of NO2 obtained in this study."*
- *"In a separate study, our meteorology-normalized estimates are used to quantify the circumstantial reduction in the mortality attributable to the short-term effects of NO2 during the lockdown (Achebak et al., 2020)."*

With the corresponding references :

- Achebak, H., Petetin, H., Quijal-Zamorano, M., Bowdalo, D., García-Pando, C. P., and Ballester, J.: Reduction in air pollution and attributable mortality due to COVID-19 lockdown, The Lancet Planetary Health, 4, e268, https://doi.org/10.1016/S2542-5196(20)30148-0, https://linkinghub.elsevier.com/retrieve/pii/S2542519620301480, 2020.

- Bauwens, M., Compernolle, S., Stavrakou, T., Müller, J., Gent, J., Eskes, H., Levelt, P. F., van der A, R., Veefkind, J. P., Vlietinck, J., Yu, H., and Zehner, C.: Impact of Coronavirus Outbreak on NO 2 Pollution Assessed Using TROPOMI and OMI Observations, Geophysical Research Letters, 47, https://doi.org/10.1029/2020GL087978, https://onlinelibrary.wiley.com/doi/abs/10.1029/2020GL087978, 2020.
- Guevara, M., Jorba, O., Soret, A., Petetin, H., Bowdalo, D., Serradell, K., Tena, C., Denier van der Gon, H., Kuenen, J., Peuch, V.-H., and Pérez García-Pando, C.: Time-resolved emission reductions for atmospheric chemistry modelling in Europe during the COVID-19 lockdowns (in review), Atmospheric Chemistry and Physics Discussions, https://doi.org/10.5194/acp-2020-686, 2020a
- Menut, L., Bessagnet, B., Siour, G., Mailler, S., Pennel, R., and Cholakian, A.: Impact of lockdown measures to combat Covid-19 on air quality over western Europe, Science of The Total Environment, 741, 140 426, https://doi.org/10.1016/j.scitotenv.2020.140426, https://linkinghub.elsevier.com/retrieve/pii/S0048969720339486, 2020.

Complete list of changes :

- Title : "*Meteorology-normalized impact of the COVID-19 lockdown upon NO_2 pollution in Spain*"
- Affiliations : "*ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Spain*"
- L1 : "*The spread of the new coronavirus SARS-COV-2 causing COVID-19 forced the Spanish Government […]*"
- L10 : "*The ML predictive models were found to perform remarkably well in most locations, with overall bias, root-mean-squared error and correlation of +4%, 29% and 0.86, respectively.*"
- L24 : "*The rapid spread of the new coronavirus SARS-COV-2 that causes COVID-19 […]*"
- L39 : "*While such an extraordinary situation has obviously impacted the levels of air pollution in the country, as seen in both surface and satellite observations (Tobias et al., 2020; Bauwens et al., 2020), the extent of such reductions remains uncertain*."
- L45 : "*[…] testing the ability of CTMs to reproduce the observed changes of concentrations under unusually different emissions (Guevara et al., 2020b; Menut et al., 2020)*."
- L75 : "*The fraction of E1a data is 0% in 2020, 99% in 2019 and 100% in 2013-2018.*"
- L76 : "*All NO_2 measurements taken into account here are operated using chemiluminescence with an internal Molybdenum converter. Although predominantly used over Europe for measuring NO_2, this measurement technique is well known to be have strong positive artifacts due to interferences of NO_z compounds (e.g. nitric acid, peroxyacetyl nitrates, organic nitrates), especially during daytime when these species are photo-chemically formed, up to a factor of 2-4 as observed during summertime in urban atmospheres (e.g. Dunlea et al., 2007; Villena et al., 2012). In our case, the positive artifacts at urban background stations are probably lower*

since the period of study (late winter and early spring) is less photo-chemically active than summertime. Even lower interferences are expected at traffic stations where the NO_z/NO_x ratio is typically lower due to the proximity to fresh NOx emissions. In any case, the present study focuses on the relative changes of NO_2 due to the lockdown, so biases in the NO_2 measurements are of lower importance."

- L100 : " *[…] total cloud cover, surface net solar radiation, surface solar radiation downwards, downward UV radiation at the surface and boundary layer height.*"
- L114 : "*Choice of features and modeling strategy"*
- L118 : *"[…] total cloud cover, surface net solar radiation, surface solar radiation downwards, downward UV radiation at the surface, boundary layer height […]"*
- L124 : "*Including such a feature with unique values (going from 0 for 2013/01/01 to 2669 for 2020/04/23)"*
- L136 : "*This ML experiment is hereafter referred to as EXP_{2020}."*
- L155 : "*These ML experiments are hereafter referred to as EXP_{2016}, EXP_{2017}, EXP_{2018} and EXP_{2019}, respectively."*
- L159 : *"Averaged over all Spanish provinces, the uncertainty interval is [-5.1, +5.3] ppbv at urban background stations, and [-6.6, +6.7] ppbv at traffic stations."*
- L167 : *"Because these daily uncertainties are likely at least partly uncorrelated, NO2 daily predictions averaged over time periods longer than one day are expected to have smaller uncertainties due to error compensations."*
- L172 : *"On average over all provinces, the weekly uncertainty interval obtained are [-3.8, +3.6] ppbv at urban background stations, and [-4.9, +4.7] ppbv at traffic stations, which represents a reduction of 28% for both types of stations, with respect to the daily uncertainties."*
- L179 : *"Note that these ancillary ML experiments used here for quantifying the uncertainties also allow to evaluate the performance of our modeling strategy during the period of the year of the lockdown (as explained later in Sect. 3.1)."*
- L181 : *"Time series in the other 48 Spanish provinces can be found in the Supplement."*
- L186 : "*The performance of the ML predictions in each Spanish province and station type is shown in Fig. 2, and the statistics over all Spanish provinces reported in Table 1. Statistical results in Table 1 are given for both the reference ML experiment (EXP2020) and the other experiments combined together (EXP2016, EXP2017, EXP2018 and EXP2019, hereafter referred to as EXP2016–2019). Besides providing a broader view of the performance of our modeling strategy, considering these past experiments also allows assessing the performance of the ML predictions during the period of the year of the lockdown (14/03-30/04, for years 2016 to 2019), which may be important given the potential seasonality of prediction errors. Statistics obtained at urban background and traffic stations are given in Table A2 in Appendix."*
- L190 : "*For information purposes, we included the statistical results obtained over the training dataset (2017/01/01-2019/12/31 in*

EXP_{2020}). Checking results over the training data may be useful for highlighting obvious situations of overfitting, when the performance is almost perfect. At both urban background and traffic stations, results show no bias, low nRMSE (always below 35%, 19% when considering all provinces), and a high PCC of 0.96. Similar results are obtained when considering the ensemble of all past experiments ($EXP_{2016–2019}$)."

- L195 : "*On the test dataset of the EXP_{2020} reference experiment (2020/01/01-2020/03/13, before the lockdown), the performance remains reasonably good in most provinces. Over all Spanish provinces, the nMB increases to +4%, the nRMSE to 29% and the PCC is reduced to 0.86, in very close agreement with the performance obtained with $EXP_{2016–2020}$ (nMB of +1%, nRMSE of 28% and PCC of 0.86). In comparison, the performance obtained in $EXP_{2016–2019}$ during the period of the year of the lockdown (14/03-30/04) is a bit lower but remains reasonable, with a nMB of +4%, a nRMSE of 37% and a PCC of 0.80. Although moderate, such a deterioration of the performance after mid-March might reflect some seasonality in the ML model errors and/or could be related to the presence of trends in the NO2 concentrations. Concerning this last point, as previously discussed in Sect. 2.3.2, including the date index feature in the ML model aims at limiting this potential issue but likely cannot completely solve it. Generally, only minor differences of performance are found between urban background and traffic stations. Results of EXP_{2020} per province (Fig. 2) highlight some inter-regional variability of the performance, with poorer statistics in some provinces, at least for one type of station. At most stations, the bias remains below ±20% while nRMSE ranges between 15 and 45% (highest nRMSE around 50% in Teruel, Tenerife and Fuerteventura). Most provinces show PCC around 0.6-0.9, with only a few exceptions below 0.6 (urban background sites in Bizkaia, Fuerteventura, Huesca and traffic sites in Granada and Gran Canaria).*"
- L225 : *"like in the Canary Islands (e.g. Tenerife and Fuerteventura)."*
- L233 : "*89% (4240 points over 4788)*"
- L246 : *"(nMB of -3 and +6%, nRMSE of 19 and 22%, PCC of 0.87 and 0.85, respectively)."*
- L254 : *"(strict enforcement through fines to offending motorists was not expected until April 1st and was finally postponed to September 15th 2020 due to the COVID-19 situation)"*
- L265 : *"The uncertainty at weekly scale is here used as an estimate of the uncertainty at 90% confidence level (by construction, given that they are computed as the 5th and 95th percentiles of the weekly residuals, see Sect. 2.3.3) affecting the mean NO_2 change.*"
- L267 : "*-7[-13,-1] ppbv*"
- L268 : *"-39[-74,-4]%"*
- L269 : *"-10[-15,-5] ppbv, or -59[-87,-30]%"*
- L276 : *"(nRMSE of 25%) and correlations (PCC of 0.72)"*
- L276 : *"The positive bias in the traffic station started in early February and persisted during the following weeks"*
- L277 : *"(+0%), and reaches +8%"*

- L284 : *"start before April 1st (postponed to September 15th 2020 due to the COVID-19 situation)."*
- L304 : *"decreased by -7[-12,-2] ppbv (-47[-78,-16]%)"*
- L306 : *"-15[-20,-10] ppbv (-61[-80,-38]%)."*
- L317 : *"significance. During the lockdown period, 96% (2734 points over 2844) of the observed daily NO2 mixing ratios are lower than the ML-based business-as-usual NO2 estimates."*
- L318 : *"-4[-8,-0] ppbv (-49[-95,-0]% in relative terms)"*
- L320 : *"and -1 ppbv (-31%)."*
- L321 : *"22 out of 38 provinces,"*
- L322 : *"-7[-11,-2] ppbv (or -50[-91,-8]%), and 26 out of 37 stations"*
- L329 : *"about -42% at both station types, and further increased to about -54% during phases II and III."*
- L332 : *"between -20 and -40% depending on the type of station during phases II and III, compared to only -9 to -19% during phase I."*
- L337 : *"Barcelona Supercomputing Center (Guevara et al., 2020b)."*
- L353 : *"lockdown (Guevara et al., 2020a)"*
- L364 : *"-44 and -53% at the urban background and traffic stations, respectively"*
- L367 : *"-50 and -63% at urban background and traffic stations"*
- L368 : *"NO2 reductions of -43 and -60%"*
- L382 : *"The NO2 changes obtained with the climatological average approach are reported on Fig. 5 (and for the different phases in Figs. A1, A2, A3, A4 in Appendix)."*
- *L391 : "biased by +27%."*
- L395 : *"+12, +2.3 and +1.8%"*
- L396 : *"-21/+52, -34/+44 and -41/+36% during phases I, II and III, respectively. For the case of Barcelona province, these relative biases are +35, +19 and 22%."*
- L412 : *"fed by meteorological data and time variables (Julian date, day of week and date index)"*
- L417 : *"We also demonstrated the benefits of our meteorology-normalization approach compared to a simple climatological-based approach, especially at smaller temporal and spatial scales."*
- L440 : *"The results of the present study provide a valuable reference for validating similar assessments of the impact of the COVID-19 lockdown on air quality based on chemistry transport models and emission scenarios derived from activity data during the lockdown (e.g. Guevara et al., 2020a; Menut et al., 2020)."*
- L441 : *"during the lockdown (Achebak et al., 2020)."*
- L442 : *"EEA AQ e-Reporting,"*
- L458 : *"All the corresponding measurements were removed from the dataset."*

Figures and tables :

- We modified the color bar of Figs 1 and 6
- We reshaped Table 1 and its legend

- We added monthly climatological NO2 mixing ratios on Figs. 3 and 4, and modified the legend : "*The climatological monthly averages computed over the period 2017-2019 are also shown (in black). The vertical black line identifies the beginning of the lockdown, the next dotted lines separate the different lockdown phases (phase I : 2020/03/14-2020/03/29, phase II : 2020/03/30-2020/04/09, phase III : 2020/04/10-2020/04/23).*"
- NO2 changes in Table 2 have been slightly modified, according to the new results obtained with the extended set of features.
- We added the NO2 changes obtained with the climatological average approach in Fig. 5 and modified the legend : "*For comparison, the mean NO_2 changes obtained using the climatological average (over 2017-2019) rather than ML-based business-as-usual NO2 concentration are also shown (stars), as well as the relative difference between both approaches (circles).*"

Appendix :

- Figs A1-A4 have been modified (we added information regarding NO2 changes obtained with the climatological average approach)
- Table A2 added (with detailed information about the statistical results obtained at urban background and traffic stations)

Supplement : We included the time series (similar to Figs. 3 and 4) for 48 Spanish provinces.

Meteorology-normalized impact of the COVID-19 lockdown upon NO_2 pollution in Spain

Hervé Petetin[1], Dene Bowdalo[1], Albert Soret[1], Marc Guevara[1], Oriol Jorba[1], Kim Serradell[1], and Carlos Pérez García-Pando[1,2]

[1]Barcelona Supercomputing Center, Barcelona, Spain
[2]ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Spain

Correspondence: Hervé Petetin (herve.petetin@bsc.es)

Abstract. The spread of the new coronavirus SARS-COV-2 causing COVID-19 forced the Spanish Government to implement extensive lockdown measures to reduce the number of hospital admissions, starting on March 14th 2020. Over the following days and weeks, strong reductions of nitrogen dioxide (NO_2) pollution were reported in many regions of Spain. A substantial part of these reductions is obviously due to decreased local and regional anthropogenic emissions. Yet, the confounding effect of meteorological variability hinders a reliable quantification of the lockdown impact upon the observed pollution levels. Our study uses machine learning (ML) models fed by meteorological data along with other time features to estimate the "business-as-usual" NO_2 mixing ratios that would have been observed in the absence of the lockdown. We then quantify the so-called meteorology-normalized NO_2 reductions induced by the lockdown measures by comparing the business-as-usual with the actually observed NO_2 mixing ratios. We applied this analysis for a selection of urban background and traffic stations covering the more than 50 Spanish provinces and islands.

The ML predictive models were found to perform remarkably well in most locations, with overall bias, root-mean-squared error and correlation of +4%, 29% and 0.86, respectively. During the period of study, going from the enforcement of the state of alarm in Spain on March 14th to April 23rd, we found the lockdown measures to be responsible for a 50% reduction of NO_2 levels on average over all provinces and islands. The lockdown in Spain has gone through several phases with different levels of severity in the mobility restrictions. As expected the meteorology-normalized change of NO_2 was found to be stronger during the phases II (the most stringent one) and III than during phase I. In the largest agglomerations where both urban background and traffic stations were available, a stronger meteorology-normalized NO_2 change is highlighted at traffic stations compared to urban background ones. Our results are consistent with foreseen (although still uncertain) changes in anthropogenic emissions induced by the lockdown. We also show the importance of taking into account the meteorological variability for accurately assessing the impact of the lockdown on NO_2 levels, in particular at fine spatial and temporal scales.

Meteorology-normalized estimates such as the ones presented here are crucial to reliably quantify the health implications of the lockdown due to reduced air pollution.

1 Introduction

The rapid spread of the new coronavirus SARS-COV-2 that causes COVID-19 has led numerous countries worldwide to put their citizens on various forms of lockdown, with measures ranging from light social distancing to almost complete restrictions on mobility (Anderson et al., 2020). Spain has been among the countries most severely affected by COVID-19, and where proportional (and therefore drastic) containment measures have been implemented. Spanish authorities declared the constitutional state of alarm on March 13th 2020, to be enforced on the 14th. During this period (phase I) residents had to remain in their primary residences except for purchasing food and medicines, work or attend emergencies. Non-essential shops and businesses, including bars, restaurants, and commercial businesses had to close. Due to the persistent rise in hospital admissions, an even more severe second phase (phase II) of the lockdown was implemented between March 30th and April 9th, when only essential activities including food trade, pharmacy, and some industries were authorized. A third phase (phase III) started on April 10th, when some non-essential sectors, including construction and industry, were allowed to return to work.

The shutdown of both social and economic activities in Spain has reduced anthropogenic pollutant emissions. Among the sectors presumably most affected, road transport, which is a dominant source of air pollution in urban areas, and air traffic have fallen to unprecedentedly low levels. The impact on the industrial sector is presumably more contrasted, as some essential industries (e.g. fuel and energy related, petrochemical) were authorized to continue their production, while some others were forced to halt their activity.

While such an extraordinary situation has obviously impacted the levels of air pollution in the country, as seen in both surface and satellite observations (Tobías et al., 2020; Bauwens et al., 2020), the extent of such reductions remains uncertain. Besides emissions, air pollution is strongly influenced by meteorological conditions driving their dispersion and short- to long-range transport, and affecting their removal and chemical evolution. As highlighted by Tobías et al. (2020) in Barcelona, this makes the quantification of air pollution reductions during the lockdown unreliable when solely based on the analysis of in-situ observations. Chemistry-transport models (CTMs) are an essential tool for investigating both actual and alternative states of the atmosphere under different emission scenarios. Actually, the lockdown offers unique opportunities for so-called dynamical CTM evaluations (Rao et al., 2011), i.e., testing the ability of CTMs to reproduce the observed changes of concentrations under unusually different emissions (Guevara et al., 2020a; Menut et al., 2020). However, given the difficulty of accurately estimating the changes in emissions induced by the lockdown along with the inherent limitations of CTMs, particularly in urban areas, estimating the reductions with this method remains a complex task sullied by substantial uncertainties that are difficult to quantify.

The need for attributing changes in pollutant concentrations to changes in emissions recently motivated the development of so-called weather normalisation techniques based on machine learning (ML) algorithms (Grange et al., 2018; Grange and Carslaw, 2019). The idea consists in training ML models to predict pollutant concentrations at air quality (AQ) monitoring stations based a set of features including meteorological data and other time variables. This allows for the building of ML models that learn the influence of meteorology upon pollutant concentrations under a given average emission forcing. These ML models can then be used for predicting pollutant concentrations under a range of meteorological conditions, with the associated average

referred to as meteorology-normalized time series in Grange et al. (2018) and Grange and Carslaw (2019). In addition, such ML models can be used for predicting business-as-usual pollutant concentrations during periods with presumably different emissions, i.e., estimating the pollutant concentrations that would have been experienced without the change in emissions.

Following the ideas introduced in Grange et al. (2018) and Grange and Carslaw (2019), the present study uses ML models to investigate the reduction of nitrogen dioxide (NO_2) concentrations in Spain due to the COVID-19 lockdown. Since road transport and industry are major sources of NO_2 emissions, the impact of the lockdown on this primary pollutant is expected to be strong and thus easier to detect and quantify. Due to its short lifetime and relatively simple chemistry, NO_2 is likely more directly impacted by meteorological conditions than other pollutants like particulate matter that depend upon more numerous and complex processes.

2 Data and methods

2.1 NO_2 data

This study primarily relies on hourly NO_2 measurements performed routinely in Spanish AQ surface monitoring stations. We considered the time period going from 2013/01/01 to 2020/04/23. We used the NO_2 data available through the GHOST (Globally Harmonised Observational Surface Treatment) project developed at the Earth Sciences Department of the Barcelona Supercomputing Center. GHOST is a project dedicated to the harmonisation of global surface atmospheric observations and metadata, for the purpose of facilitating quality-assured comparisons between observations and models within the atmospheric chemistry community (Bowdalo, in preparation). GHOST ingests numerous publicly available AQ observational datasets. In this study, we used the NO_2 data from the European Environmental Agency (EEA) AQ e-Reporting (EEA, 2020). We prioritized the validated data (E1a) and used the near-real time data (E2a) only when necessary. The fraction of E1a data is 0% in 2020, 99% in 2019 and 100% in 2013-2018.

All NO_2 measurements taken into account here are operated using chemiluminescence with an internal Molybdenum converter. Although predominantly used over Europe for measuring NO_2, this measurement technique is well known to have potentially strong positive artifacts due to interferences of NO_z compounds (e.g. nitric acid, peroxyacetyl nitrates, organic nitrates), especially during daytime when these species are photo-chemically formed, up to a factor of 2-4 as observed during summertime in urban atmospheres (e.g. Dunlea et al., 2007; Villena et al., 2012). In our case, the positive artifacts at urban background stations are probably lower since the period of study (late winter and early spring) is less photo-chemically active than summertime. Even lower interferences are expected at traffic stations where the NO_z/NO_x ratio is typically lower due to the proximity to fresh NO_x emissions. In any case, the present study focuses on the relative changes of NO_2 due to the lockdown, so biases in the NO_2 measurements are of lower importance.

GHOST provides a wide range of harmonized metadata and quality assurance (QA) flags for all pollutant measurements. In this study, we took benefit of these flags to apply an exhaustive QA screening. More details on the QA flags used can be found in Appendix A.

NO_2 measurements are available over the period 2013 to 2020 in 551 stations in Spain. This study aims at investigating the reduction of NO_2 over a variety of environments and geographical locations. We thus designed an algorithm for automatically selecting (when possible) one urban/suburban background station and one traffic station in each Nomenclature of Territorial Units for Statistics level 3 (NUTS-3) (Ceuta and Melilla excluded), which corresponds to Spanish provinces over mainland and individual islands over the Balearic and Canary Islands (hereafter referred to as provinces for convenience). After the QA screening of NO_2 data, we set different thresholds for minimum data availability over different periods of interest, namely 50% of daily data over the entire period of study, 50% over the period 2017/01/01-2019/01/01 (used for training the ML models, see below), 25% over the period 2020/01/01-2020/03/13 (used for testing the ML models) and 10% during the lockdown period. Stations in each province were then selected to maximize the surrounding population density (within a geodesic radius of 5 km) and the data availability (both before and during the lockdown). The population density at AQ monitoring stations was retrieved through GHOST, which ingests the Gridded Population of the World (GPW) version 4 dataset (Center for International Earth Science Information Network - CIESIN - Columbia University, 2018). Stations fulfilling the different criteria were identified in 50 provinces of Spain and are considered in this study (38 provinces with urban background stations and 37 provinces with traffic stations). No appropriate stations were found in Palencia, Ávila and some islands (La Palma, La Gomera, El Hierro, Lanzarote, Eivissa and Formentera). A map of the entire NO_2 monitoring network is shown in Fig. 1 together with the stations selected in each Spanish province. Names and geographical locations of the stations are reported in Table A1 in Appendix.

2.2 Meteorological data

Meteorological data are taken from the ERA5 reanalysis dataset (Copernicus Climate Change Service (C3S), 2017). ERA5 data have a spatial resolution of about 31 km. At all AQ monitoring surface stations, we extracted the following variables at the daily scale : daily mean 2-m temperature, minimum and maximum 2-m temperature, surface wind speed, normalized 10-m zonal and meridian wind speed components, surface pressure, total cloud cover, surface net solar radiation, surface solar radiation downwards, downward UV radiation at the surface and boundary layer height.

2.3 Methodology

We implement and train ML models to estimate the daily NO_2 mixing ratios that would have been observed without the implementation of the lockdown in each selected station, i.e. under business-as-usual emission forcing. Hereafter, we will refer to these mixing ratios as business-as-usual NO_2.

2.3.1 Machine learning model

In this study, we retain the Gradient Boosting Machine (GBM), a popular decision tree-based ensemble method belonging to the boosting family (Friedman, 2001). More information on this model is given in Appendix B. ML models based on decision trees offer several interesting attributes. First, they internally handle the process of feature selection, which allows including potentially useless features without strong deterioration of the prediction skills. Second, they provide useful information about

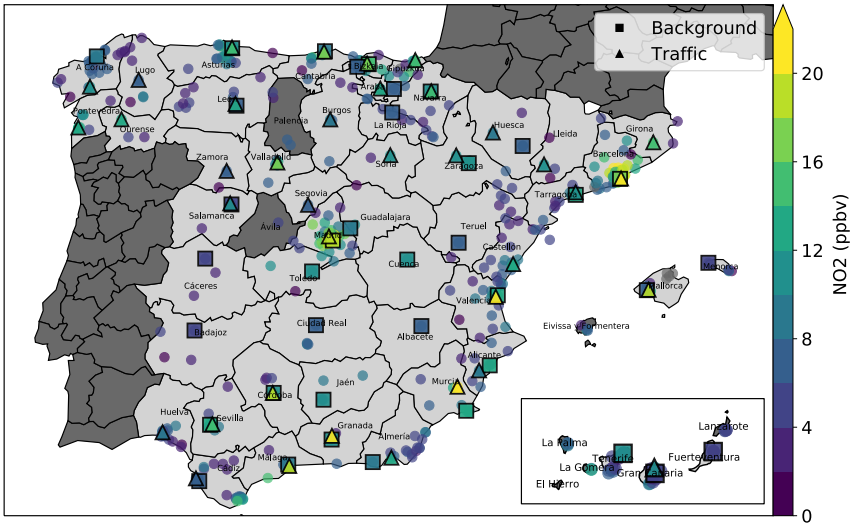

Figure 1. Mean NO_2 mixing ratios [ppvb] (2013-2020) at all (circles) and selected (squares and triangles) stations. Administrative borders show the NUTS-3 administrative units, which correspond to the Spanish provinces over mainland and to individual islands. Dark gray areas indicate provinces and islands with a lack of stations that fulfill the selection criteria.

the importance of the different features. Third, in contrast to most parametric methods that derive a unique (more or less sophisticated) function supposedly valid over the whole features' space, non-parametric methods based on decision trees internally rely on successive splitting operations (a mother branch being divided into two daughter branches), which may be convenient for designing one single model able to work efficiently under different seasons and weather regimes.

2.3.2 Choice of features and modeling strategy

Following the work of Grange and Carslaw (2019), the idea here is to use past recent data to train a ML model able to reproduce the NO_2 mixing ratios based on a combination of meteorological features and other time features. The features used in this study are : daily mean 2-m temperature, minimum and maximum 2-m temperature, surface wind speed, normalized 10-m zonal and meridian wind speed components, surface pressure, total cloud cover, surface net solar radiation, surface solar radiation downwards, downward UV radiation at the surface, boundary layer height, date index (days since 2013/01/01), Julian date and weekday. All the data used in this study are daily. Some pollutant concentrations are known to strongly vary depending on the season, day of week and hour of the day, notably due to the variability of emissions and chemistry. The two last time features act as proxies for these processes and aim at representing their climatological variations. Over longer (multi-annual) time scales, typically air pollutant concentrations cannot be considered as stationary due to substantial trends (especially in emissions), which is intrinsically problematic for training ML models. Following Grange et al. (2018) and Grange and Carslaw (2019), we introduced the date index as a proxy for this potential trend. Including such a feature with unique values (going

from 0 for 2013/01/01 to 2669 for 2020/04/23) is not expected to directly help the ML model to learn about NO_2 variability. However, it allows us to train one single ML model over a relatively long and thus potentially non-stationary time series. In contrast to linear regression, GBM does not learn equations relating the target variable to the different features, but rather builds non-parametric relationships between target and features. As a consequence, such a model will always make NO_2 predictions within the range of NO_2 values used in the training, regardless of the inclusion of the aforementioned date index feature or the feature values it takes for making the predictions. However, if NO_2 strongly increases (decreases) with time in the training dataset, the GBM model is able to split the data using the trend feature and therefore predict NO_2 in the range of the higher (lower) mixing ratios reached by the end of the training period. We note that even with a trend feature, such a model is not expected to stay valid very far in time relative to the training data when the training data is following a too strong trend. Our sensitivity tests have clearly shown that the behaviour of the ML models substantially improves when including the trend feature.

In our study, the GBM models are trained over the 3 last full years, namely 2017-2019 and then used for predicting business-as-usual NO_2 mixing ratios over the 4 following months, from January to April 2020. This ML experiment is hereafter referred to as EXP_{2020}. Such a duration for training is expected to allow capturing a substantial part of the inter-annual variability of NO_2 mixing ratios and meteorological conditions and ensures some past data is available for quantifying the uncertainties of our ML modeling strategy (as explained later in Sect. 2.3.3). Note that no improvement was found with extended training periods of 4 or 5 years. Although our interest is to predict NO_2 during the lockdown period, the two and half preceding months were kept to test the validity of our predictions and uncertainty estimates.

The machine learning modeling in this study is performed using the *scikit-learn* Python package (Pedregosa et al., 2011). The GBM model comprises a number of hyperparameters to be tuned. Since features are temporal variables, instances cannot be considered as independent due to autocorrelation. We thus tuned our ML models using the so-called time series cross-validation with five splits, which corresponds to a rolling-origin cross-validation in which data used for the validation is always posterior to the data used for the training (*TimeSeriesSplit* in *scikit-learn*). Over a selection of the most important hyperparameters, we applied a so-called *randomized search* over a range of possible hyperparameter values. Compared to the so-called *grid search* in which all combinations of hyperparameters are tested, the randomized approach tests only a certain number (20 in our case) of tuning configurations chosen randomly. This allows to explore a large part of the hyperparameters space at a greatly reduced computational cost, and tends to be less prone to overfitting. More details on the tuning of the GBM model can be found in Appendix C.

2.3.3 Uncertainty estimation

In order to quantify our prediction uncertainty, we replicated four similar experiments over the past years since 2013, i.e., training ML models over 2013-2015, 2014-2016, 2015-2017 and 2016-2018, and testing them over the 4 first months of 2016, 2017, 2018 and 2019, respectively. These ML experiments are hereafter referred to as EXP_{2016}, EXP_{2017}, EXP_{2018} and EXP_{2019}, respectively. We obtained on average 538 daily residuals (predicted minus observed NO_2 daily mixing ratios) for each station and we took the associated 5^{th} and 95^{th} percentiles as the uncertainty interval for our ML-based predictions of daily NO_2

mixing ratios. Therefore, for each station, we obtained a fixed asymmetric 90% confidence interval used to characterize the uncertainty of our predictions during the first 4 months of 2020. Averaged over all Spanish provinces, the uncertainty interval is [-5.1, +5.3] ppbv at urban background stations, and [-6.6, +6.7] ppbv at traffic stations.

In 2020, the period before the lockdown, namely January 1st to March 13th, is used to check the performance of the ML models trained over 2017-2019 against the observed NO_2 mixing ratios, given the aforementioned uncertainty. Ideally, we expect the differences between observed and predicted NO_2 mixing ratios to remain within the estimated uncertainty during that period. Conversely, after April 14th, due to the reduction of NO_2 emissions caused by the lockdown, we expect the observed NO_2 mixing ratios to quickly decrease compared to the business-as-usual NO_2 mixing ratios predicted by the ML model, eventually down to a level at which the differences are statistically significant.

These uncertainties are suited for our ML-based daily NO_2 predictions. Because these daily uncertainties are likely at least partly uncorrelated, NO_2 daily predictions averaged over time periods longer than one day are expected to have smaller uncertainties due to error compensations. We estimated the uncertainty affecting our ML predictions at the weekly scale. We used a similar approach than previously described for the daily uncertainty, but based on the 7-day running average of the daily residuals (by requiring a minimum of 5 over 7 days with available data). The 5th and 95th percentiles were computed based on the entire set of residuals (514 residuals on average at each station over 2016-2019). On average over all provinces, the weekly uncertainty interval obtained are [-3.8, +3.6] ppbv at urban background stations, and [-4.9, +4.7] ppbv at traffic stations, which represents a reduction of 28% for both types of stations, with respect to the daily uncertainties.

Our main interest in this study is to quantify the mean NO_2 changes during the lockdown period. We decided to keep the weekly scale uncertainties for the predictions of business-as-usual NO_2 mixing ratios averaged over its different phases (10-13 days each) and over the entire lockdown period (41 days). The use of weekly uncertainties is likely conservative when used for the entire lockdown average, but accounts for potential data gaps, particularly when estimating the shorter phases therein.

Note that these ancillary ML experiments used here for quantifying the uncertainties also allow to evaluate the performance of our modeling strategy during the period of the year of the lockdown (as explained later in Sect. 3.1).

3 Results and Discussion

In this section, we first evaluate the ML-based predictions of business-as-usual NO_2 mixing ratios (Sect. 3.1). We then illustrate our methodology in the two provinces with largest population density, namely Madrid and Barcelona (Sect. 3.2). Time series in the other 48 Spanish provinces can be found in the Supplement. We then analyze the meteorology-normalized changes of NO_2 obtained in all Spanish provinces (Sect. 3.3). We discuss in Sect. 3.4 the potential relationships with emission reductions. Finally, we discuss in Sect. 3.5 the advantages of our ML-based approach for estimating the baseline NO_2 pollution compared to the climatological approach.

3.1 Evaluation of ML predictions

The performance of the ML predictions in each Spanish province and station type is shown in Fig. 2, and the statistics over all Spanish provinces are reported in Table 1. The statistical results in Table 1 are given for both the reference ML experiment (EXP_{2020}) and the other experiments combined together (EXP_{2016}, EXP_{2017}, EXP_{2018} and EXP_{2019}, hereafter referred to as $EXP_{2016-2019}$). Besides providing a broader view of the performance of our modeling strategy, considering these past experiments also allows assessing the performance of the ML predictions during the period of the year of the lockdown (14/03-30/04 for years 2016 to 2019), which may be important given the potential seasonality of prediction errors. The statistics obtained at urban background and traffic stations are given in Table A2 in Appendix. Results are evaluated using the following metrics, calculated based on daily NO_2 mixing ratios : mean bias (MB), normalized mean bias (nMB), root mean square error (RMSE), normalized root mean square error (nRMSE) and Pearson correlation coefficient (PCC).

For information purposes, we included the statistical results obtained over the training dataset (2017/01/01-2019/12/31 in EXP_{2020}). Checking results over the training data may be useful for highlighting obvious situations of overfitting, when the performance is almost perfect. At both urban background and traffic stations, results show no bias, low nRMSE (always below 35%, 19% when considering all provinces), and a high PCC of 0.96. Similar results are obtained when considering the ensemble of all past experiments ($EXP_{2016-2019}$). Although such a performance obtained is very good, there are no clear signs of too prejudicial overfitting at this stage.

On the test dataset of the EXP_{2020} reference experiment (2020/01/01-2020/03/13, before the lockdown), the performance remains reasonably good in most provinces. Over all Spanish provinces, the nMB increases to +4%, the nRMSE to 29% and the PCC is reduced to 0.86, in very close agreement with the performance obtained with $EXP_{2016-2020}$ (nMB of +1%, nRMSE of 28% and PCC of 0.86). In comparison, the performance obtained in $EXP_{2016-2019}$ during the period of the year of the lockdown (14/03-23/04) is a bit lower but remains reasonable, with a nMB of +4%, a nRMSE of 37% and a PCC of 0.80. Although moderate, such a deterioration of the performance after mid-March might reflect some seasonality in the ML model errors and/or could be related to the presence of trends in the NO_2 concentrations. Concerning this last point, as previously discussed in Sect. 2.3.2, including the date index feature in the ML model aims at limiting this potential issue but likely cannot completely solve it. Generally, only minor differences of performance are found between urban background and traffic stations (Table A2).

Results of EXP_{2020} per province (Fig. 2) highlight some inter-regional variability of the performance, with poorer statistics in some provinces, at least for one type of station. At most stations, the bias remains below $\pm$20% while nRMSE ranges between 15 and 45% (highest nRMSE around 50% in Teruel, Tenerife and Fuerteventura). Most provinces show PCC around 0.6-0.9, with only a few exceptions below 0.6 (urban background sites in Bizkaia, Fuerteventura, Huesca and traffic sites in Granada and Gran Canaria).

Several factors may explain the poorer statistical results obtained at some stations. First and foremost, it may be due to deficiencies in the ML modeling, and in particular to some overfitting. This seems to be the case of Fuerteventura and Huesca,

Table 1. Performance of the ML predictions of NO_2 mixing ratios. Results are shown for both the reference experiment EXP_{2020} and the ensemble of past experiments combined together ($EXP_{2016-2019}$).

Experiments	Dataset	Period of the year (day/month)	MB [ppbv] (nMB [%])	RMSE [ppbv] (nRMSE [%])	PCC	N
EXP_{2020}	Training	01/01-31/12	-0.0 (-0%)	2.2 (19%)	0.96	72983
	Test	01/01-13/03	0.6 (4%)	3.8 (29%)	0.86	4788
$EXP_{2016-2019}$	Training	01/01-31/12	0.0 (0%)	2.2 (18%)	0.96	297609
	Test	01/01-13/03	0.1 (1%)	4.0 (28%)	0.86	19178
		14/03-23/04	0.5 (4%)	4.0 (37%)	0.80	11097
		01/01-23/04	0.2 (2%)	4.0 (31%)	0.85	30275

given the good performances obtained with the training data (note also that the data availability of test data in Fuerteventura is among the poorest). Since we are considering numerous stations in this study, we need a fixed procedure applied similarly to all ML models to be trained. As described in Sect. 2.3.2, we designed our training and tuning procedure in order to limit as much as possible this common issue, through rolling-origin cross-validation and randomized search in the hyperparameters space. Overall results are satisfactory but some overfitting can still persist in some cases.

Second, although moderately, some of the biases and errors may be partly due to trends and/or inter-annual variability of NO_2. As previously explained (Sect. 2.3.2), by model design, if NO_2 levels in the first months of 2020 are outside of the NO_2 range in the 2017-2019 training dataset, our predictions over the lockdown period could be equally biased. The different NO_2 time series indeed show some cases where NO_2 mixing ratios are lower than in the past years (since 2013). In the frame of our study, it is important to mention that, although the lockdown was officially implemented on March 14th, the COVID-19 started to perturb the business-as-usual situation in the days/weeks before, first through the cancellation of numerous events and, later, through unusual movements of a part of the population (e.g. to second homes). Although complicated to assess more precisely in each of the Spanish provinces, this likely explains part of the biases noticed in the second half of the test period.

Third, poor performances at some stations may be due to weaker relationships between meteorological input data and NO_2 mixing ratios. This points to uncertainties in the ERA5 meteorology data. For example, the relatively coarse spatial resolution (31 km) of ERA5 data may only capture part of the meteorological variability existing at a given station. This is especially true when considering stations located in urban areas where the complex urban morphology (e.g. presence of buildings, canyon streets) is known to locally distort the mesoscale circulation. Decision-tree based ML methods like GBM offer some interpretability by providing a measure of the importance of the different features included as input data. In our case, on average over all ML models, the most important feature is the boundary layer height (18$\pm$6%) followed by the surface wind speed (12$\pm$5%). These two parameters drive the ventilation and dispersion of the pollutants emitted at the surface, and their variability at some stations may be only partly captured by the ERA5 data at some urban stations. Also, the ERA5 data may poorly capture the meteorological conditions in some stations located on small islands with complex orography, like in the Canary

Islands (e.g. Tenerife and Fuerteventura).

The chosen training and tuning procedures applied in this study were designed to limit some of these different sources of uncertainty, but persistent errors cannot be excluded. This is why we added another layer of analysis in which we estimated the uncertainties of our ML predictions by replicating exactly the same procedure over the past years since 2013 (as explained in Sect. 2.3.3). Computed as the 5th and 95th percentiles of the daily residuals obtained over a large test period extending over several years (2016-2019), the uncertainty intervals are expected to cover most (90%) of the errors caused by these different sources of uncertainties. Indeed, considering all stations, our results indicate that 89% (4240 points over 4788) of the daily NO_2 observations in 2020 before the lockdown fall within the corresponding prediction uncertainty interval at each station, thus very close to 90%. This demonstrates that the daily uncertainty estimated in this study is well quantified.

All in all, we have shown that our ML predictions and associated uncertainties are qualified for estimating the business-as-usual NO_2 mixing ratios during the lockdown.

3.2 Illustration of the results in specific provinces

3.2.1 Madrid

The daily NO_2 mixing ratios observed and predicted in the province of Madrid are shown in Fig. 3 for both the urban background station and the traffic station, with station codes-*names* ES1941A-*Ensanche de Vallecas* and ES1938A-*Castellana*, respectively. The NO_2 mixing ratios observed over the past years since 2013 are also included. Since days of week are not consistent from one year to the other, we also show the NO_2 7-day running mean time series where a minimum of 5 over 7 days is required to compute the average.

In Madrid, the ML reproduces remarkably well the variability of NO_2 mixing ratios at the urban background and traffic stations before the lockdown (nMB of -3 and +6%, nRMSE of 19 and 22%, PCC of 0.87 and 0.85, respectively). Importantly, prediction errors remain within the uncertainty interval. The two sub-periods with lower NO_2 mixing ratios, during the second half of January and early March occur concomitantly with strong wind speeds in Madrid, above 6 m s^{-1} on a daily average (above the 95th percentile of the ERA5 daily wind speed over 2013-2020 during this season), and relatively high boundary layer heights (up to 1000-1500 m on a daily average). It is worth mentioning that a low emission zone (LEZ) with relatively strict vehicle restrictions applied for entering a limited area of about 5 km^2 corresponding to the heart of the city center was implemented in early January 2020. Such a change in emissions may in principle directly impact the performance of the ML predictions by inducing a positive bias (since the ML models are designed precisely for highlighting such events). In our case, we expect a limited impact because the LEZ was still in its transition phase (strict enforcement through fines to offending motorists was not expected until April 1st and was finally postponed to September 15th 2020 due to the COVID-19 situation) and the two stations selected in Madrid province are located outside the LEZ (at 9 and 3 km from the city center).

After the implementation of the lockdown, the observed NO_2 mixing ratios decreased down to about 11 and 7 ppbv on average, and reached daily minimum mixing ratios of 6 and 3 ppbv, respectively, over the entire period. Compared to the previous

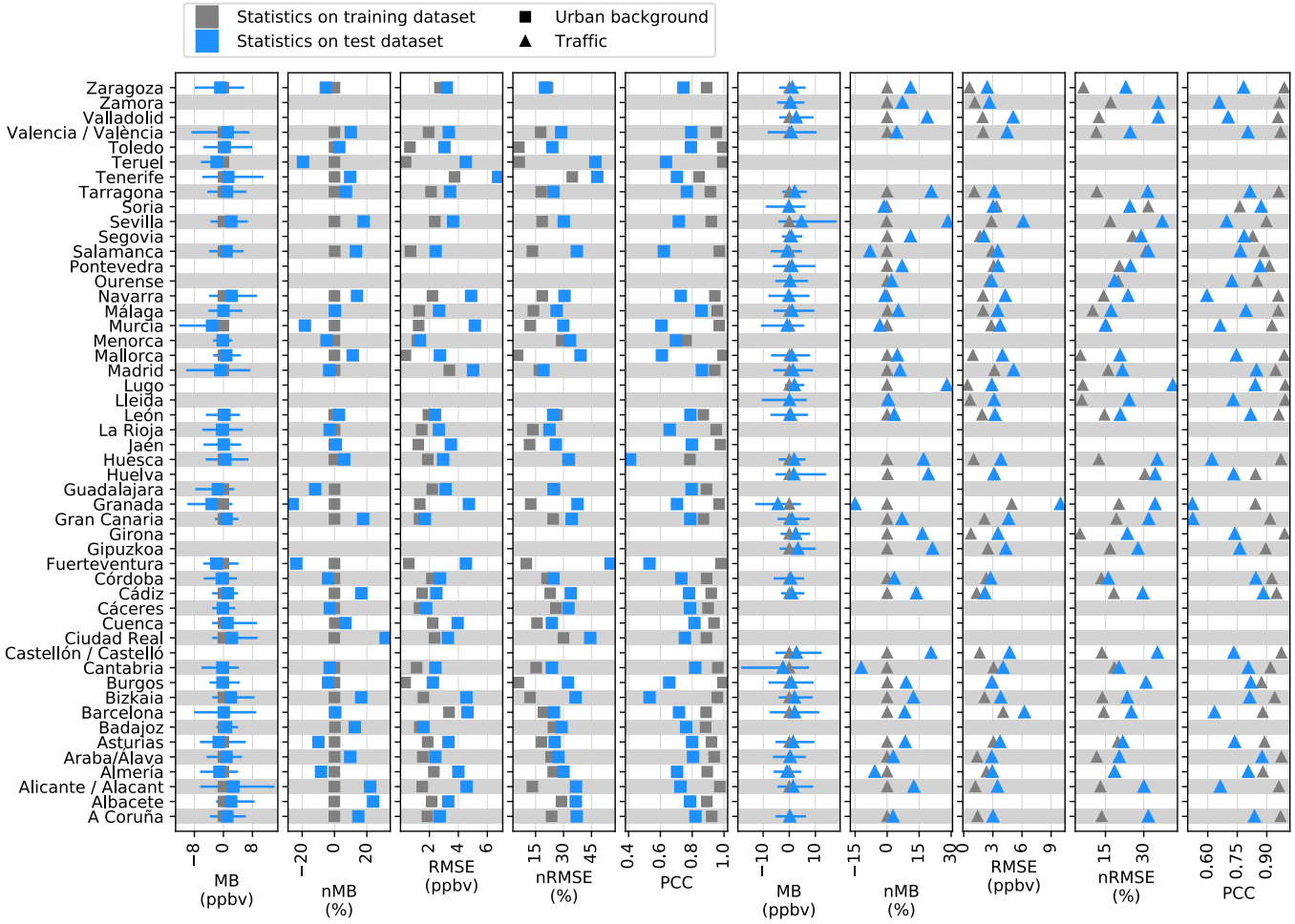

Figure 2. Statistical results of the ML-predicted business-as-usual NO_2 mixing ratios (EXP_{2020} reference experiment) over training dataset (2017-2019, in grey) and test dataset before lockdown (2020/01/01-2020/03/13, in blue). Metrics are mean bias (MB), normalized mean bias (nMB), root mean square error (RMSE), normalized root mean square error (nRMSE) and Pearson correlation coefficient (PCC). For information purposes, the uncertainties (90% confidence interval) at the daily scale are added to MB (horizontal blue bars).

years, the NO_2 mixing ratios at the urban background site are clearly in the lower tail of the distribution. In the traffic site, never NO_2 levels had been so low for such an extended period of time at least since 2013. In comparison, business-as-usual NO_2 mixing ratios at these two sites would have remained around 17-18 ppbv on average. After the lockdown, the differences between the observed and business-as-usual NO_2 are found to progressively increase, becoming more and more statistically significant. This demonstrates unambiguously that the lockdown considerably reduced the NO_2 pollution in Madrid, regardless of the meteorological conditions, which points to a drastic decrease of the business-as-usual emission forcing.

We computed the meteorology-normalized change of NO_2 during the lockdown period covered by this study (from March

14th to April 23th) as the mean difference between ML-based business-as-usual and observed NO_2 daily mixing ratios. The uncertainty at weekly scale is here used as an estimate of the uncertainty at 90% confidence level (by construction, given that they are computed as the 5th and 95th percentiles of the weekly residuals, see Sect. 2.3.3) affecting the mean NO_2 change. On average over the entire lockdown period, NO_2 levels have decreased by -7[-13,-1] ppbv at the urban background station, which corresponds to -39[-74,-4]% in relative terms. The impact is faster, stronger and more statistically significant at the traffic station than in the urban background one, with a mean NO_2 reduction of -10[-15,-5] ppbv, or -59[-87,-30]% in relative terms. This result is consistent with a lockdown affecting most strongly the sector of traffic emissions. At the daily scale, the reduction of NO_2 in Madrid reached its maximum at the end of the second and more stringent lockdown phase, while a strong reduction persisted during the third phase.

3.2.2 Barcelona

Figure 4 presents the results in Barcelona for both the urban background and traffic stations, with station codes-*names* ES1396A-*Sants* and ES1480A-*L'Eixample*, respectively. Compared to Madrid, the ML predictions in Barcelona have relatively similar errors (nRMSE of 25%) and correlations (PCC of 0.72). The bias is very low at the urban background station (+0%), and reaches +8% at the traffic station, which largely remains within the uncertainty interval. The positive bias in the traffic station started in early February and persisted during the following weeks, particularly after the second week of February. The ML model failed at reproducing these low NO_2 mixing ratios notably because some of the observed NO_2 mixing ratios during that period were lower than during the previous years. As in Madrid, a LEZ was implemented in Barcelona, starting in early January 2020, with less stringent vehicle restrictions but over a larger area (95 km^2). Both the urban background and traffic stations selected in Barcelona are included in this LEZ. The potentially stronger effect of the LEZ at traffic stations could explain at least partly this positive bias. As in the case of Madrid, fines for non-compliance with the LEZ restrictions were not planned to start before April 1st (postponed to September 15th 2020 due to the COVID-19 situation). Therefore the effect of the LEZ is expected to be progressive, which is consistent with the absence of bias in the beginning of the period. In addition, the 2020th edition of the World Mobile Congress (the largest annual event in Barcelona, with 109,000 visitors in 2019) that takes place every year by the end of February was officially canceled by the organizers due to the risks posed by the emerging COVID-19 pandemic. We therefore hypothesize this cancellation contributed to the reduction of NO_2 levels in the city and to the slight positive bias of the ML prediction before the lockdown.

After the lockdown, NO_2 mixing ratios decreased down to 8 and 11 ppbv on average at the urban background and traffic stations, respectively, both reaching minimum daily mixing ratios of 4 ppbv. Results highlight strong and statistically significant differences with the business-as-usual situation in which NO_2 levels would have remained around 15-21 ppbv during that period. As in Madrid, the strongest differences are found in April, during the phases II and III of the lockdown. Note that these differences exceed by large the aforementioned positive bias encountered after February. Interestingly, besides the strong reduction, observed NO_2 mixing ratios followed a very similar variability than business-as-usual NO_2, which highlights the major influence of meteorological conditions on the levels of pollution, as previously mentioned by Tobías et al. (2020). For instance, the increase of NO_2 mixing ratios between April 6th and April 9th appears linked to unusually low wind speeds over

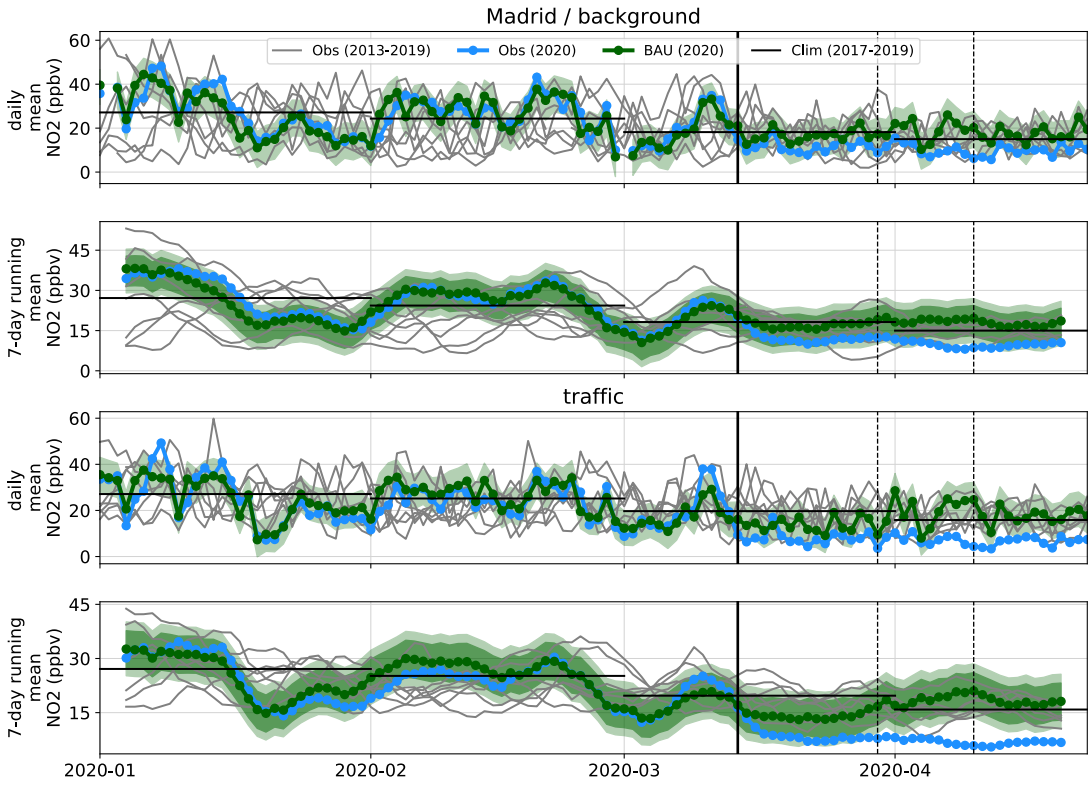

Figure 3. NO_2 mixing ratios in Madrid province. The two top panels show the daily mean and 7-day running mean at the urban background station, respectively. The two bottom panels show the time series at the traffic station. Each panel displays the NO_2 mixing ratios observed in 2020 (in blue) and during the past years (2013-2019, in grey), and predicted in 2020 by the ML model (business-as-usual (BAU), in green). The uncertainties of the ML predictions are given at a 90% confidence level at the daily (light green) and weekly scales (medium green). The climatological monthly averages computed over the period 2017-2019 are also shown (in black). The vertical black line identifies the beginning of the lockdown, the next dotted lines separate the different lockdown phases (phase I : 2020/03/14-2020/03/29, phase II : 2020/03/30-2020/04/09, phase III : 2020/04/10-2020/04/23).

Barcelona, 1.7 $m.s^{-1}$ on average over these days, which is slightly below the climatological (2013-2020) 5^{th} percentile of wind speed in April (1.8 $m.s^{-1}$). Without the lockdown, this stagnant situation associated with the business-as-usual emission forcing would have increased NO_2 by about 5-10 ppbv, according to the ML predictions. Observed NO_2 also slightly increased during the episode of stagnant meteorological conditions, but due to the lockdown, NO_2 remained at very low levels. This event illustrates the usefulness of considering a ML model fed by meteorological data for quantifying the baseline air pollution during the lockdown.

Over the entire lockdown period, NO_2 in Barcelona decreased by -7[-12,-2] ppbv (-47[-78,-16]%) at the urban background station, regardless of the meteorological conditions. As in Madrid, a stronger reduction is found at the traffic station, with -15[-

20,-10] ppbv (-61[-80,-38]%). Therefore, in relative terms, the lockdown has induced a relatively similar decrease of NO_2 in both Madrid and Barcelona.

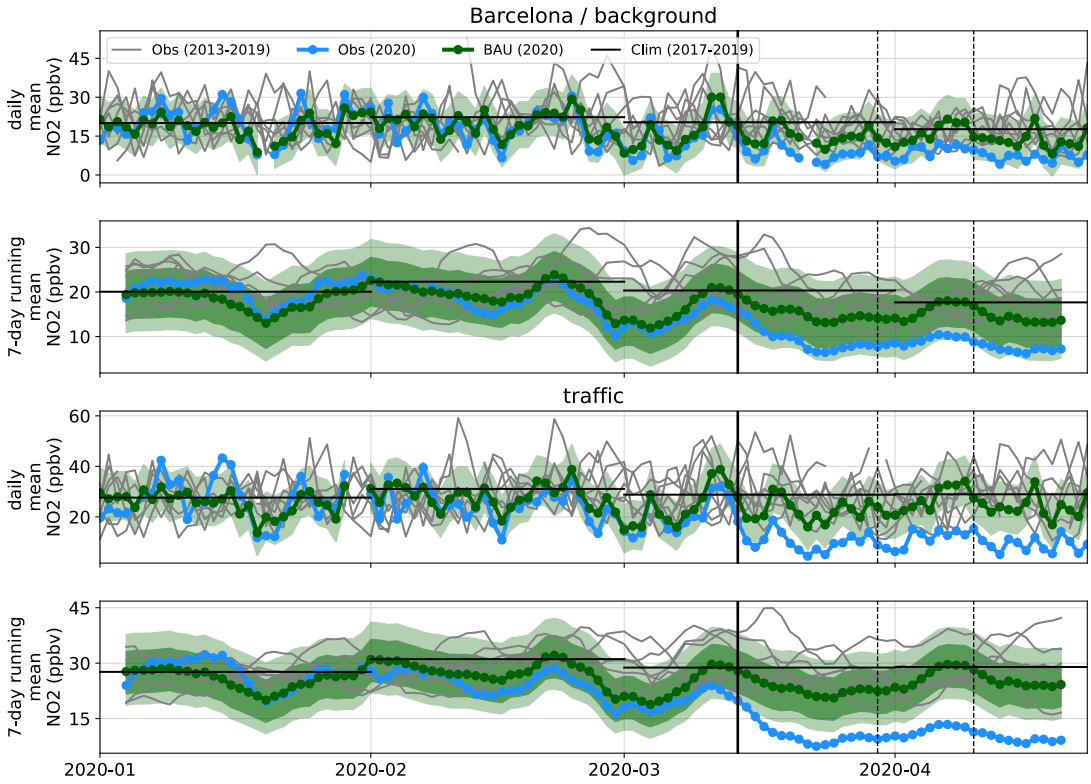

Figure 4. Similar to Fig. 3 in Barcelona province.

3.3 Meteorology-normalized changes of NO_2 mixing ratios over Spain

We computed the meteorology-normalized changes of NO_2 for all the selected stations. Results are presented in Fig. 5, together with the weekly uncertainty of our ML predictions (colored lines). For information purposes, we also display the daily uncertainty (black lines). Results are colored as a function of their degree of significance, here computed as the distance between the NO_2 change best estimate and the upper limit of the weekly uncertainty interval, normalized by the distance between the best estimate and zero. A degree of significance of 1 thus indicates a NO_2 change significant at a 90% confidence level. Statistics over the changes of NO_2 obtained in all provinces are reported in Table 2. A map of best estimates of NO_2 changes at each station is also given in Fig. 6.

Results highlight that the reduction previously described in Madrid and Barcelona extends to most Spanish provinces, although with some inter-regional variability in the extent of the change and the degree of statistical significance. During the lockdown

period, 96% (2734 points over 2844) of the observed daily NO_2 mixing ratios are lower than the ML-based business-as-usual NO_2 estimates. On average over all urban background stations during the entire lockdown period, NO_2 has decreased by -4[-8,-0] ppbv (-49[-95,-0]% in relative terms), independently from the meteorological conditions. The 5th and 95th percentiles (computed based on the mean NO_2 changes obtained in all provinces) are -7 ppbv (-65%) and -1 ppbv (-31%). The NO_2 change is significant with more than 90% confidence in 22 out of 38 provinces, with many of the remaining ones being relatively close to that confidence level. A similar, yet more statistically significant reduction is found at traffic stations, with a mean NO_2 decrease of -7[-11,-2] ppbv (or -50[-91,-8]%), and 26 out of 37 stations exceeding the 90% confidence level. The spread of NO_2 change between the different provinces is also quite similar between the two types of stations, with 5th and 95th percentiles of -69 and -29%, respectively. Generally, the meteorology-normalized NO_2 reductions in the provinces of the southern half of the country appear stronger and in more cases statistically significant.

As previously observed in Madrid and Barcelona, results in Table 2 highlight noticeable differences between the different phases of the lockdown. The corresponding figures (with both absolute and relative changes) can be found in Appendix (Figs. A1, A2, A3 and A4). The mean reduction of NO_2 during phase I was about -42% at both station types, and further increased to about -54% during phases II and III. The lower reduction during the first phase is partly explained by the fact that NO_2 concentrations started at their business-as-usual levels and took a few days to reach their minimum. During the two last phases, NO_2 was found to be reduced in many more provinces, as shown by the 95th percentile that ranges between -20 and -40% depending on the type of station during phases II and III, compared to only -9 to -19% during phase I.

3.4 Relationship to emission reductions

We contrasted our results with a detailed NO_x anthropogenic emission inventory at 4km x 4km resolution over Spain available through the bottom-up module of the HERMESv3 emission model, developed at the Earth Sciences Department of the Barcelona Supercomputing Center (Guevara et al., 2020b). Averaged over the different stations considered in this study, road transport emissions are the dominant source, with 66 and 69% of the total NO_x emissions in the vicinity of urban background and traffic stations, respectively. The other emission sources are the residential/commercial combustion sector (14 and 15%), industrial point sources (8 and 13%) and shipping and port activities (11 and 3%). In Spain, the public agency in charge of monitoring traffic (*Dirección General del Tráfico*) reported progressive reductions in total traffic down to levels about -60 to -90% lower than usual, with substantial day-to-day variability and strongest reductions during weekends. Assuming to first order a linear relationship between NO_2 urban background mixing ratios and local surrounding NO_x emissions (within a 4km x 4km cell) and applying a 70% (80%) reduction of road transport would lead to a NO_2 reduction of about 47% (54%), which is consistent with our findings. Our knowledge about the impact of the lockdown on the other emission sectors remains at this stage quite limited. NO_x emissions from industry likely also decreased but quantifying this reduction, even roughly, is more complex as some industries were considered as essential and thus not affected by the lockdown. Although 9-13% of the surrounding emissions (in the 4km x 4km cell of the inventory) are associated to this sector, the impact of idling industrial activities on the pollution levels observed at the selected stations may be relatively small considering that none of these stations are classified as "industrial". The residential/commercial emission sector represents another unknown since the expected emission increment

Table 2. Meteorology-normalized changes of NO_2 mixing ratios in Spain during the lockdown (phase I : 2020/03/14-2020/03/29, phase II : 2020/03/30-2020/04/09, phase III : 2020/04/10-2020/04/23). Statistics are computed based on the mean NO_2 changes in the different Spanish provinces.

Change	Metric	Phases I+II+III		Phase I		Phase II		Phase III	
		Background	Traffic	Background	Traffic	Background	Traffic	Background	Traffic
absolute (ppbv)	mean	-4.1	-6.5	-3.4	-5.6	-5.2	-7.4	-4.3	-6.8
		[-7.8,-0.3]	[-11.1,-1.6]	[-7.1,0.4]	[-10.2,-0.7]	[-8.9,-1.4]	[-11.9,-2.4]	[-7.9,-0.4]	[-11.3,-2.0]
	std	2.0	3.4	1.8	3.2	2.4	3.6	2.2	3.7
	min	-10.0	-15.5	-8.4	-13.3	-10.8	-16.1	-10.9	-16.8
	p05	-7.1	-12.8	-6.3	-11.5	-9.2	-14.2	-7.7	-13.5
	p10	-6.8	-11.4	-5.5	-10.9	-8.3	-12.8	-7.0	-12.3
	p25	-5.3	-7.4	-4.8	-6.9	-6.8	-8.2	-5.3	-9.5
	p50	-3.9	-6.1	-3.2	-5.0	-4.7	-7.0	-3.8	-5.9
	p75	-2.6	-4.5	-2.0	-3.9	-3.2	-5.0	-2.5	-4.3
	p90	-2.1	-2.6	-1.5	-1.7	-2.9	-3.3	-1.9	-2.6
	p95	-1.4	-2.0	-1.2	-0.6	-2.5	-2.4	-1.2	-2.3
	max	-0.8	-0.8	-0.5	-0.0	-1.1	-1.6	-0.7	-0.7
relative (%)	mean	-49	-50	-41	-42	-55	-53	-53	-55
		[-95,-0]	[-91,-8]	[-89,8]	[-82,-0]	[-95,-11]	[-90,-13]	[-100,-1]	[-97,-11]
	std	13	12	14	17	9	11	15	13
	min	-72	-71	-65	-67	-69	-73	-76	-73
	p05	-65	-69	-62	-63	-68	-71	-73	-72
	p10	-64	-63	-59	-60	-67	-68	-70	-70
	p25	-58	-58	-53	-55	-65	-60	-65	-65
	p50	-51	-52	-41	-46	-54	-54	-55	-56
	p75	-39	-43	-29	-38	-47	-46	-42	-51
	p90	-34	-33	-24	-14	-43	-35	-36	-39
	p95	-31	-29	-19	-9	-40	-34	-20	-31
	max	-14	-14	-14	-1	-39	-27	-12	-12

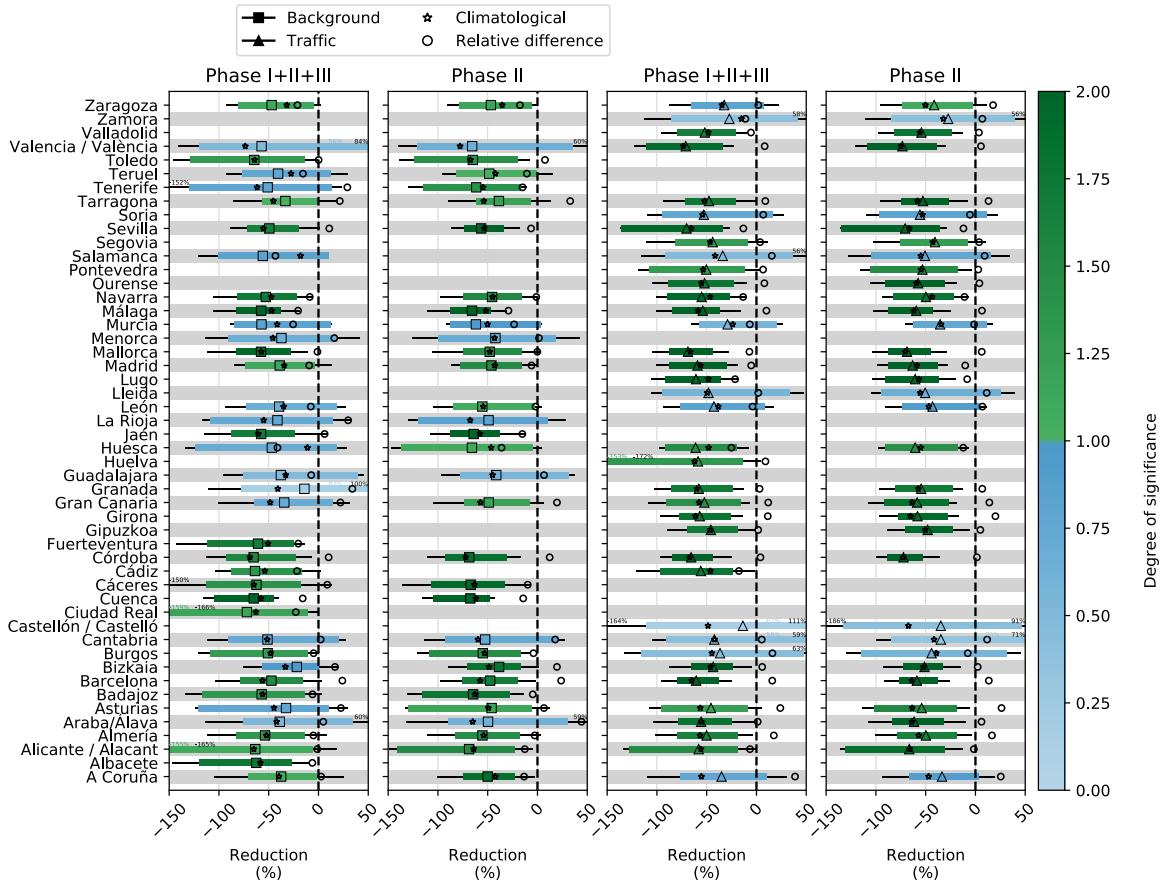

Figure 5. Meteorology-normalized mean NO_2 changes at urban background (squares) and traffic (triangles) stations during the COVID-19 lockdown. Changes are shown during the entire lockdown period and during the second and most stringent phase. Best estimates and weekly uncertainties are colored according to the degree of significance (a value of 1 indicates a change statistically significant at a 90% confidence level, see text for more details). For information purposes, daily uncertainties are also indicated (black lines). For comparison, the mean NO_2 changes obtained using the climatological average (over 2017-2019) rather than ML-based business-as-usual NO_2 concentration are also shown (stars), as well as the relative difference between both approaches (circles).

caused by a population spending more time at home may be compensated by the closure of most shops, schools and offices. A more detailed analysis of the activity data in these different emission sectors is required to better quantify how the emission forcing has been modified by the lockdown (Guevara et al., 2020a) and to understand the reductions of NO_2 obtained in this study.

Concerning traffic stations, although HERMESv3 gives a quite similar contribution of the different emission sectors compared to urban background stations, a larger contribution of road transport emissions is evidently expected since measurement instruments are deployed under the direct influence of vehicles. As a consequence, assuming that road transport is the emission sector

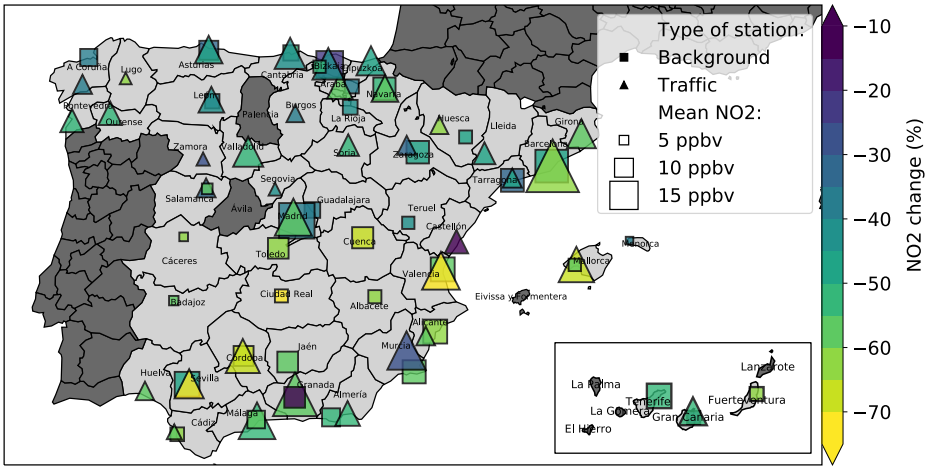

Figure 6. Meteorology-normalized mean NO_2 changes at selected urban background and traffic stations during the COVID-19 lockdown in Spain. The size of symbols is proportional to the annual average NO_2 mixing ratio (over 2013-2020).

most impacted by the lockdown (together with air traffic, but this last sector does not emit strong amounts of NO_x around our set of stations), we could expect a stronger relative reduction of NO_2 at traffic stations, compared to urban background stations. At first glance, Table 2 does not highlight such a difference between the two types of stations. This seems to be due to the fact that we here gather urban background and traffic stations not always collocated in the same cities, and/or located in cities of very different sizes. In both Madrid and Barcelona provinces, the two selected stations are located in the same agglomeration, and results do highlight substantial differences of NO_2 reductions (Sect. 3.2). In total, urban background and traffic stations are collocated in the same agglomeration in 16 provinces. On average over this set of provinces, the NO_2 reduction is -44 and -53% at the urban background and traffic stations, respectively, thus showing a noticeable but still relatively small difference. Focusing on the 6 largest cities within this group of provinces (Madrid, Barcelona, Valencia, Sevilla, Málaga and Mallorca), the difference of NO_2 reductions increases, with -50 and -63% at urban background and traffic stations, respectively. Focusing on the 2 largest cities, namely Madrid and Barcelona, the discrepancy further increases, with the NO_2 reductions of -43 and -60%, respectively. Therefore, results suggest that the lockdown has impacted more strongly the business-as-usual NO_2 levels at traffic stations than at urban background ones, and that this difference tends to be stronger in the largest cities.

3.5 ML-based business-as-usual NO_2 versus climatological average NO_2

We developed the ML-based approach arguing that it allows avoiding a potentially erroneous assessment of the lockdown-related NO_2 changes caused by the variability of meteorological conditions. In this section, we illustrate quantitatively the benefits of our method. Besides the business-as-usual NO_2 daily concentrations obtained with our ML-based approach, we consider here the mean NO_2 concentrations observed in 2017-2019 at this period of the year (this approach being hereafter referred to as the climatological average approach). We compared the mean NO_2 concentrations obtained in each province with

both approaches during the different phases of the lockdown. Taking the ML-based approach as the reference, we computed the bias of the climatological average approach. In this frame, in a given province, a small bias between the two approaches should indicate that the meteorological conditions prevailing during a given phase of the lockdown are relatively close to their climatological values at this time of the year. For convenience, both urban background and traffic stations are gathered in this analysis.

The NO_2 changes obtained with the climatological average approach are reported on Fig. 5 (and for the different phases in Figs. A1, A2, A3, A4 in Appendix). Considering the entire lockdown period, the mean business-as-usual NO_2 mixing ratios predicted by the ML models averaged over all provinces is 10.3 ppbv, in close agreement with the corresponding climatological mean NO_2 that is 10.6 ppbv. This corresponds to a mean bias (of the climatological average approach) of only +0.3 ppbv (or +2% in relative terms). This shows that under a business-as-usual scenario, the NO_2 concentrations during the lockdown period should have been close to the values typically observed at this time of the year. However, this holds at a relatively large temporal (the entire lockdown period in this case, i.e. 41 days) and spatial (all Spanish provinces) scale. These relative biases between both approaches are shown for all stations in Fig. 5 (black circles). Among the different provinces, they range between -41 and +33%, with 5^{th} and 95^{th} percentiles of -22 and +27%, thus greatly larger than its average of +2%. This highlights the presence of substantial departures from the climatology at the province scale. For instance, in Barcelona province, the ML-based business-as-usual and climatological mean urban background NO_2 mixing ratios during the lockdown period are 15 and 19 ppbv, respectively, which corresponds to a climatological approach positively biased by +27%. Such a result is not surprising since encountering climatological conditions simultaneously in all Spanish provinces is very unlikely.

Higher when considered at the province scale, the bias of the climatological average approach can also further increase when computed over shorter time periods. Indeed, during the 3 phases of the lockdown, it gets to +12, +2.3 and +1.8%, respectively, when averaged over all provinces. Among the different provinces, the corresponding 5^{th}/95^{th} percentiles reach -21/+52, -34/+44 and -41/+36% during phases I, II and III, respectively. For the case of Barcelona province, these relative biases are +35, +19 and 22%.

This analysis demonstrates the need to take into account (with ML or other techniques) the meteorological variability to accurately estimate the baseline pollution and assess the changes of pollution induced by an altered emission forcing, which appears all the more crucial when pollution changes are investigated at a fine temporal and/or spatial scale.

4 Conclusions

The fast spread of the COVID-19 coronavirus disease pushed Spanish authorities to implement a severe lockdown of the population, with drastic restrictions of social and economic activities starting on March 14^{th} 2020. Such a situation had an impact on the anthropogenic emissions from numerous activity sectors, some of them unambiguously (road transport and air traffic, and to a lesser extent the industrial sector), others with still unclear response (residential/commercial sector). Concomitantly, a reduction of NO_2 mixing ratios was reported in many locations, based on in-situ NO_2 measurements operated by air quality monitoring stations or space-based remote sensing (e.g. TROPOMI). Part of the reduction of NO_2 pollution is likely explained

by the modified emission forcing caused by the lockdown. However, the potential confounding impact of the meteorological variability (a major driver of the NO_2 variability) prevents to directly relate the reduction of NO_2 mixing ratios to the lockdown-related reduction of emissions.

To tackle this issue, we used ML models fed by meteorological data and time variables (Julian date, day of week and date index) to estimate the NO_2 mixing ratios that would have been normally observed during the COVID-19 lockdown period under a business-as-usual emission forcing and meteorological conditions prevailing during that period. We also estimated (conservative) uncertainties affecting our ML predictions. This allowed us to quantify the changes of NO_2 during the lockdown that are not directly related to the variability of meteorological conditions. On average over Spain, NO_2 mixing ratios at urban background and traffic stations were found to decrease by about -50% due to the lockdown, with stronger reductions in phases II and III (about -55%) than in phase I (about -40%). We also demonstrated the benefits of our meteorology-normalization approach compared to a simple climatological-based approach, especially at smaller temporal and spatial scales.

Due to the peculiarities of NO_2 (e.g. primary pollutant, short chemical lifetime, simple chemistry), we expect these changes to be mainly driven by the reduction of NO_x anthropogenic emissions. Considering that the lockdown also impacted the emissions of numerous other chemical compounds, an alteration of the business-as-usual chemical fate of NO_2 (through a modification of its oxidation into nitric acid) cannot be excluded. However, we are considering here urban stations located close to the NO_x emission sources, where this effect is likely small compared to the reduction of direct emissions.

Regarding our methodology, we note that the COVID-19 lockdown and the associated changes of pollutants like particulate matter should have also altered the meteorological conditions by perturbing the radiative fluxes and clouds. Indeed, this methodology precludes the remote and local influences of lockdown-related air pollution changes upon local weather. In any case, given the chaotic nature of the atmosphere and the long duration of the lockdown, it would be indeed impossible to know the weather conditions that would had been observed during the lockdown in a business-as-usual scenario.

It is also worth noting that the quality of the ERA5 meteorological data may have deteriorated due to the lockdown through the strong reduction of air traffic. Indeed, although satellites remain the dominant provider of meteorological observations, commercial aircraft provide valuable amounts of in-situ meteorological observations in the troposphere and lower stratosphere, especially for wind speed. However, some meteorological services are currently operating additional atmospheric soundings to compensate this loss of data. In any case, the impact on the meteorological conditions close to the surface is probably limited.

In this work, we analyzed the NO_2 data available in Spain over the first 41 days of lockdown, which includes the phase of most stringent lockdown in early April. At the date of submission of this study, the lockdown was still on-going in Spain, with restrictions planned to be progressively relaxed until late June at least. Indeed, the impact of the lockdown upon air pollution levels will likely extend way beyond the period considered in this study. Besides the direct effects of the lockdown-related restrictions, the foreseen economic downturn whose size, length and characteristics are still uncertain may also substantially affect the levels of NO_2 pollution, as already observed following the 2008-2009 economic recession, with one-year recession-driven NO_2 reductions of 10-30% across Spain and Europe (Castellanos and Boersma, 2012).

The results of the present study provide a valuable reference for validating similar assessments of the impact of the COVID-19 lockdown on air quality based on chemistry transport models and emission scenarios derived from activity data during the

lockdown (e.g. Guevara et al., 2020a; Menut et al., 2020).

In a separate study, our meteorology-normalized estimates are used to quantify the circumstantial reduction in the mortality attributable to the short-term effects of NO_2 during the lockdown (Achebak et al., 2020).

Code and data availability. The EEA AQ e-Reporting, ERA5 and Gridded Population of the World (GPW) version 5 datasets used in this study are publicly available. The HERMESv3_BU (Bottom-Up) code package with its documentation is publicly available at the following gitlab repository: https://earth.bsc.es/gitlab/es/hermesv3_bu (https://doi.org/10.5281/zenodo.3521897, Guevara et al., 2019).

Appendix A: Quality Assurance (QA) applied to NO_2 dataset

Using the information provided by GHOST (Globally Harmonised Observational Surface Treatment; Bowdalo, in preparation), we applied numerous QA screening to the NO_2 dataset, in order to remove : missing measurements (flag 0), infinite values (flag 1), negative measurements (flag 2), zero measurements (flag 4), measurements associated with data quality flags given by the data provider which have been decreed by the GHOST project architects to suggest the measurements are associated with substantial uncertainty or bias (flag 6), measurements for which no valid data remains to average in temporal window after screening by key QA flags (flag 8), measurements showing persistently recurring values (rolling 7 out of 9 data points; flag 10), concentrations greater than a scientifically feasible limit (above 5000 ppbv) (flag 12), measurements detected as distributional outliers using adjusted boxplot analysis (flag 13), measurements manually flagged as too extreme (flag 14), data with too coarse reported measurement resolution (above 1.0 ppbv) (flag 17), data with too coarse empirically derived measurement resolution (above 1.0 ppbv) (flag 18), measurements below the reported lower limit of detection (flag 22), measurements above the reported upper limit of detection (flag 25), measurements with inappropriate primary sampling for preparing NO_2 for subsequent measurement (flag 40), measurements with inappropriate sample preparation for preparing NO_2 for subsequent measurement (flag 41) and measurements with erroneous measurement methodology (flag 42). All the corresponding measurements were removed from the dataset.

Appendix B: Decision tree-based ensemble methods

Among the myriad of ML models available nowadays, we opted for decision tree-based ensemble methods. The general idea of ensemble methods is to combine an ensemble of independent base learners (or weak learners). Base learners here designate simple models that perform only slightly better than a random guessing. Decision trees are currently the base learner most commonly used in ML ensemble methods (but other types of learners could be possible). Given a training dataset and a regression problem, one characteristic of decision trees lies in the fact that it is always possible to reach a high accuracy (by growing a large enough tree) but at the cost of very poor generalization skills. In ML terminology, such large trees are said to have a small bias but a large variance. To be appropriate base learners, decision trees used in ensemble methods are thus constrained to have a low number of branches (sometimes referred to as trunks), which increases the bias but reduces the

variance. The strength of ensemble methods then stems out from the fact that combining a sufficiently large number of base learners (of quite poor performance individually) allows to reach an enhanced performance in addition to better generalization skills, the corresponding ensemble being less unstable to the addition of new data.

Once the form of the base learner is chosen, a strategy is required for building this ensemble of *independent* base learners. Three main approaches have been proposed over the past: (i) bagging, (ii) boosting, (iii) random forests (RF). Bagging consists in aggregating base learners trained on a bootstrap sample of the training dataset. Boosting consists in aggregating base learners trained on different labels: the first base learner is trained on the dataset, the second on the errors left by the previous one, the third on the errors left by the two previous ones, and so on. RF (used by Grange et al. (2018) and Grange and Carslaw (2019)) consists in aggregating base learners trained on random subsets of the training dataset based on a random subset of features.

Appendix C: Tuning of the GBM model

The training of the model is conducted together with a search of the optimal hyperparameter tuning. We retained a so-called *randomized search* in which a range of values is given for each hyperparameter of interest and a total number of hyperparameters combinations to test (20 in our case). Compared to the so-called *grid search* in which all combinations of hyperparameters are tested, this choice allows to explore a large part of the hyperparameters space for a greatly reduced computational cost, and is less prone to overfitting.

We used the *scikit-learn* Python package. The learning rate was fixed to 0.05 and the number of features to consider when looking for the best split is fixed to the square root of the number of features (*max_features* in *scikit-learn*, set to "sqrt"). Besides that, the tuning of the GBM model was done over the following set of hyperparameters: the tree maximum depth (*max_depth* in the *scikit-learn* Python package: values from 1 to 5 by 1), the subsample (*subsample* : from 0.3 to 1.0 by 0.1), the number of trees (*n_estimators*: from 50 to 1000 by 50) and the minimum sample in terminal leaves (*min_samples_leaf* : from 1 to 30). The maximum depth (or the maximum number of subsequent splits in the individual decision trees) controls how much interaction between the features can be taken into account. The subsample hyperparameter represents the fraction of samples to be used for fitting an individual base learner. Values below unity correspond to the so-called *stochastic gradient boosting* and usually allow to decrease the variance at the cost of an increased bias (low values also allow to speed up the training phase). The minimum sample leaf hyperparameter controls the minimum number of samples to allow in a terminal node (larger values limiting the risk of overfitting).

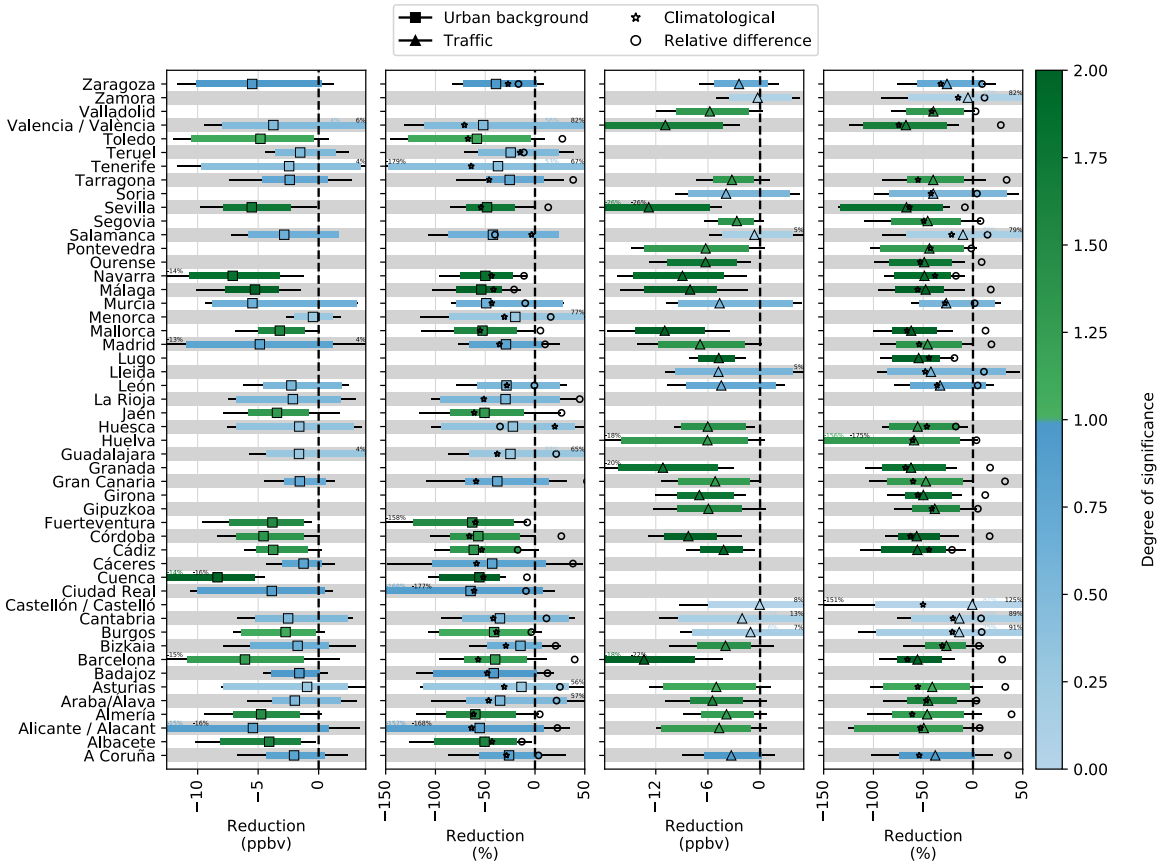

Figure A1. Absolute and relative meteorology-normalized NO_2 changes during phase I of the lockdown (2020/03/14-2020/03/29), at urban background (left panels) and traffic stations (right panels). The uncertainties shown with colored bars correspond here to the 90% confidence level interval computed at the weekly scale. For information purposes, the uncertainties affecting the ML-based daily predictions are also shown (black bars).

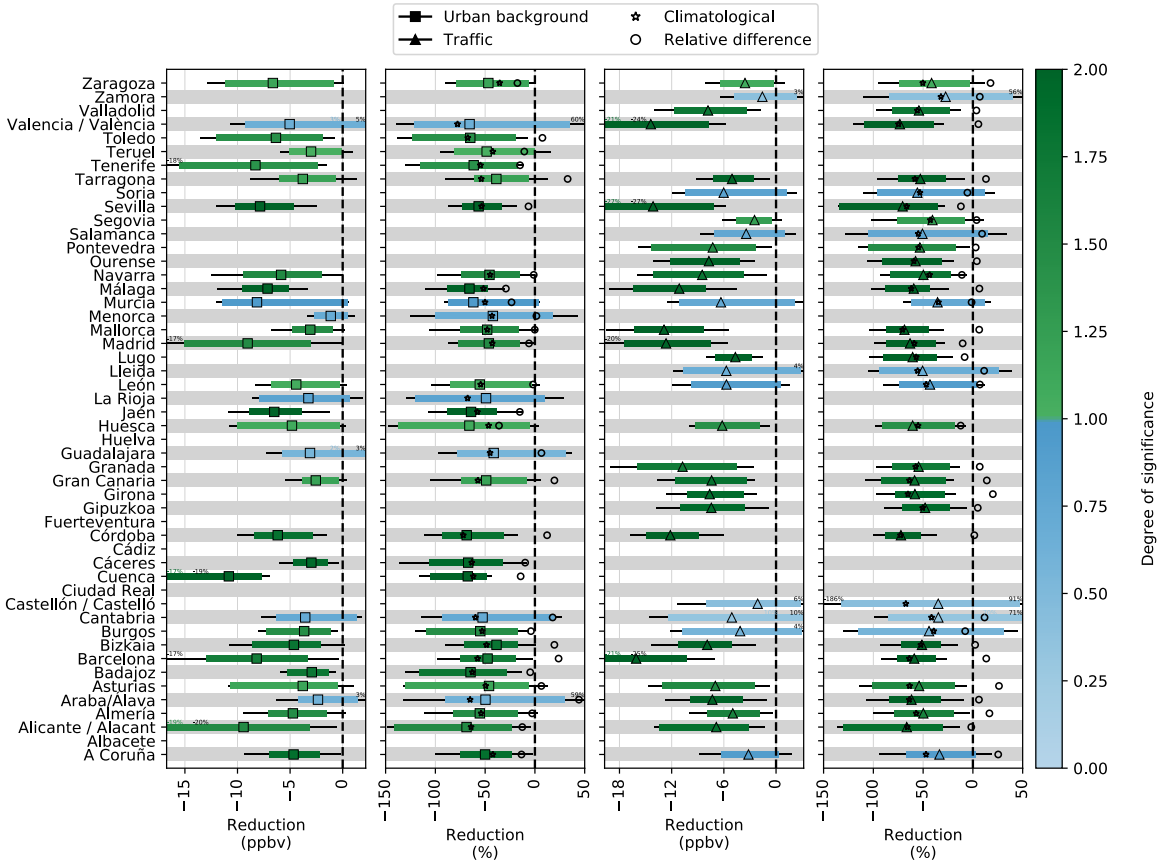

Figure A2. Similar to Fig. A1 for the phase II of the lockdown (2020/03/30-2020/04/09).

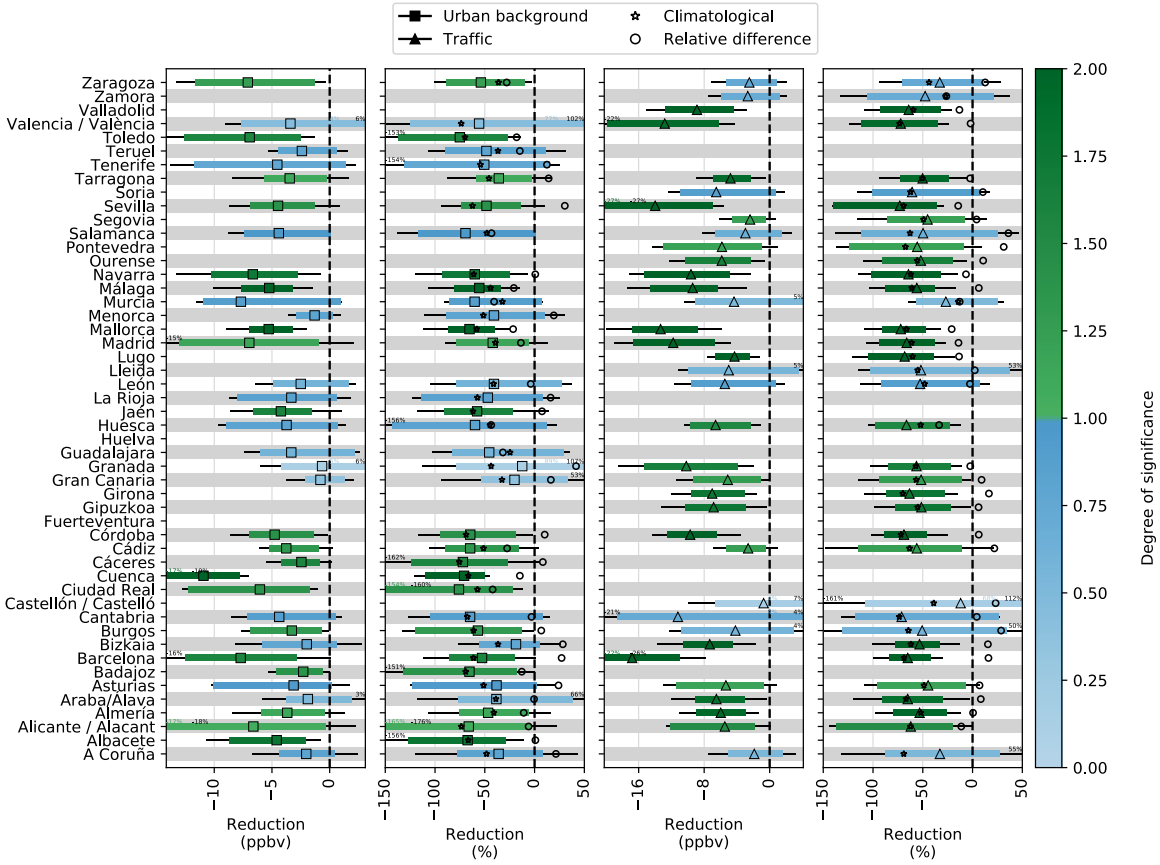

Figure A3. Similar to Fig. A1 for the phase III of the lockdown (2020/04/10-2020/04/23).

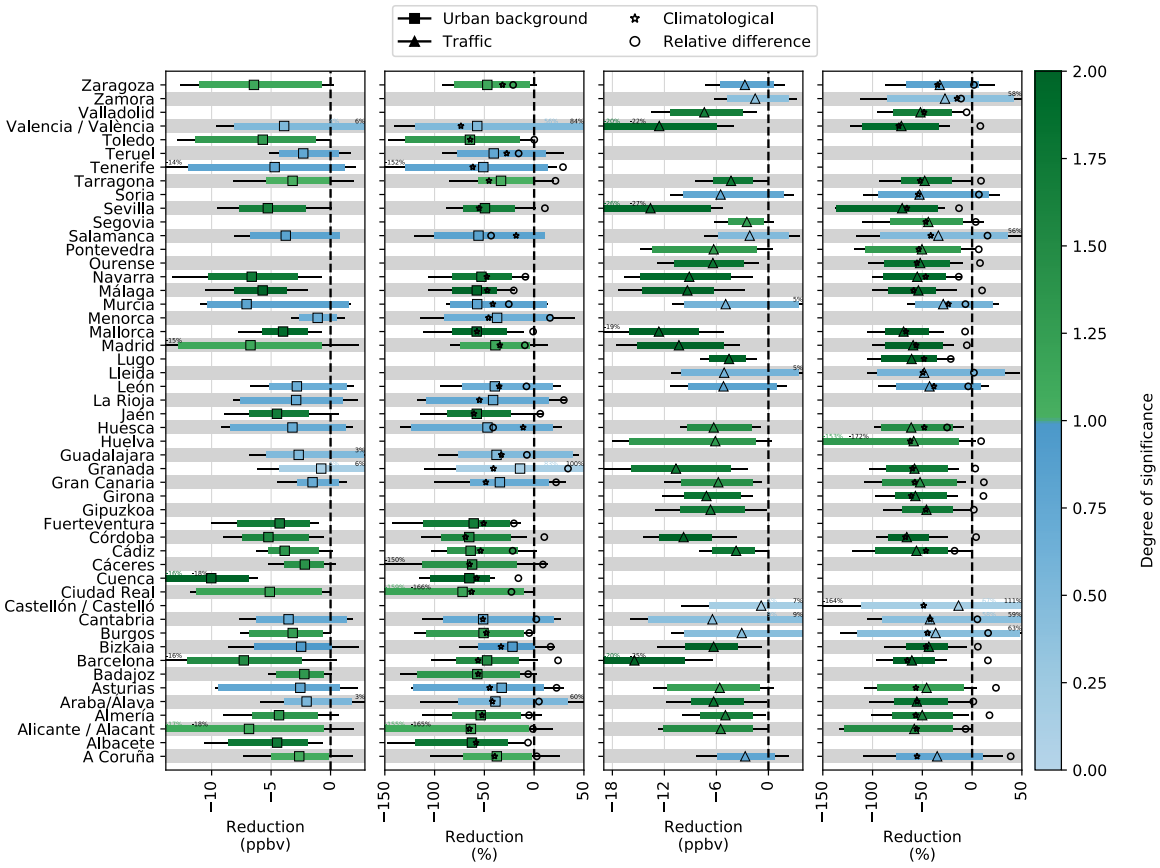

Figure A4. Similar to Fig. A1 for the entire lockdown period (2020/04/14-2020/04/23).

Table A1. Stations selected in each Spanish province.

Province	Urban background station	Traffic station
A Coruña	ES1957A Torre De Hércules (43.382800, -8.409200)	ES1901A San Caetano (42.887800, -8.531100)
Albacete	ES1535A Albacete (38.979300, -1.852100)	-
Alicante / Alacant	ES1915A Alacant-Florida-Babel (38.340278, -0.506667)	ES1849A Elx-Parc De Bombers (38.259167, -0.717500)
Almería	ES1549A El Ejido (36.769720, -2.810970)	ES1393A Mediterráneo (36.841330, -2.446720)
Araba/Álava	ES1544A Agurain (42.849000, -2.393700)	ES1492A Tres Marzo (42.856070, -2.667790)
Asturias	ES1974A Montevil (43.516600, -5.670700)	ES1272A Constitución (43.529900, -5.673500)
Badajoz	ES1819A Merida (38.907500, -6.338060)	-
Barcelona	ES1396A Barcelona (Sants) (41.378803, 2.133098)	ES1438A Barcelona (L'Eixample) (41.385343, 2.153822)
Bizkaia	ES1713A Parque Europa (43.254900, -2.902300)	ES1244A Mazarredo (43.267500, -2.935200)
Burgos	ES1598A Zalla (43.212910, -3.134400)	ES1160A Burgos 1 (42.350830, -3.675560)
Cantabria	ES1529A Tetuán (43.467780, -3.790280)	ES1580A Santander Centro (43.460560, -3.808610)
Castellón / Castelló	-	ES1834A Castelló-Patronat D'Esports (39.988889, -0.026111)
Ciudad Real	ES1857A Ciudad Real (38.993900, -3.937800)	-
Cuenca	ES1858A Cuenca (40.061900, -2.129700)	-
Cáceres	ES1997A Plasencia (40.077780, -6.147220)	-
Cádiz	ES1593A San Fernando (36.460590, -6.203070)	ES1479A Avda. Marconi (36.506020, -6.268570)
Córdoba	ES1799A Lepanto (37.892610, -4.762340)	ES2047A Avda. Al-Nasir (37.892600, -4.780100)
Fuerteventura	ES1978A Casa Palacio-Puerto Del Rosario (28.498380, -13.860830)	-
Gipuzkoa	-	ES1494A Ategorrieta (43.322000, -1.960700)
Girona	-	ES1999A Girona (Escola De Música) (41.976386, 2.816547)
Gran Canaria	ES1919A Parque De San Juan-Telde (28.003645, -15.411851)	ES1573A Mercado Central (28.133732, -15.432823)
Granada	ES1973A Ciudad Deportiva (37.135560, -3.619250)	ES1560A Granada - Norte (37.196100, -3.612660)
Guadalajara	ES1536A Azuqueca De Henares (40.571000, -3.264600)	-
Huelva	-	ES1340A Pozo Dulce (37.253360, -6.935140)
Huesca	ES2041A Monzón Centro (41.916140, 0.191101)	ES1417A Huesca (42.136110, -0.403890)
Jaén	ES1656A Ronda Del Valle (37.782550, -3.781570)	-
La Rioja	ES1602A La Cigüeña (42.464000, -2.428000)	-
León	ES1988A León 4 (42.575278, -5.566389)	ES1161A Barrio Pinilla (42.603889, -5.587222)
Lleida	-	ES1225A Lleida (Irurita - Pius Xii) (41.615795, 0.615726)
Lugo	-	ES1905A Lugo-Fingoy (42.997900, -7.550900)
Madrid	ES1941A Ensanche De Vallecas (40.372778, -3.611944)	ES1938A Castellana (40.439722, -3.690278)
Mallorca	ES1604A Bellver (39.563320, 2.620550)	ES1610A Foners (39.570080, 2.655830)
Menorca	ES1828A Ciutadella De Menorca (40.009440, 3.856480)	-
Murcia	ES1921A Mompean (37.603056, -0.975278)	ES1633A San Basilio (37.993611, -1.144722)
Málaga	ES1751A El Atabal (36.729560, -4.465530)	ES2031A Avenida Juan Xxiii (36.707300, -4.446000)
Navarra	ES1472A Iturrama (42.807220, -1.651390)	ES1740A Plaza De La Cruz (42.812220, -1.640000)
Ourense	-	ES1096A Gomez Franqueira (42.353000, -7.877900)
Pontevedra	-	ES1137A Arenal (42.219000, -8.742100)
Salamanca	ES1889A Salamanca 6 (40.960833, -5.639722)	ES1618A Salamanca 5 (40.979167, -5.665278)
Segovia	-	ES1967A Segovia 2 (40.955556, -4.110556)
Sevilla	ES1425A Principes (37.375250, -6.005580)	ES0817A La Ranilla (37.384250, -5.959620)
Soria	-	ES1643A Soria (41.766667, -2.466667)
Tarragona	ES1666A Tarragona (Parc De La Ciutat) (41.117388, 1.241650)	ES1124A Tarragona (Sant Salvador) (41.159450, 1.239704)
Tenerife	ES1975A Depósito Tristán-Sta Cruz De Tf (28.458160, -16.278776)	-
Teruel	ES1421A Teruel (40.336390, -1.106670)	-
Toledo	ES1818A Toledo2 (39.868100, -4.020800)	-
Valencia / València	ES1885A València-Politècnic (39.480300, -0.336400)	ES1239A València-Pista De Silla (39.456111, -0.375833)
Valladolid	-	ES1631A Arco De Ladrillo Ii (41.645556, -4.730278)
Zamora	-	ES1927A Zamora 2 (41.509722, -5.746389)
Zaragoza	ES1641A Renovales (41.635280, -0.893610)	ES1418A Alagón (41.762780, -1.143330)

Table A2. Performance of the ML predictions of NO_2 mixing ratios. Results are shown for both the reference experiment EXP_{2020} and the ensemble of past experiments combined together ($EXP_{2016-2019}$).

Experiments	Dataset	Period of the year (day/month)	Type of station	MB [ppbv] (nMB [%])	RMSE [ppbv] (nRMSE [%])	PCC	N
EXP_{2020}	Training	01/01-31/12	Urban background	0.0 (0%)	1.8 (19%)	0.96	36371
			Traffic	-0.0 (-0%)	2.5 (19%)	0.95	36612
			Any	-0.0 (-0%)	2.2 (19%)	0.96	72983
	Test	01/01-13/03	Urban background	0.3 (2%)	3.5 (31%)	0.85	2343
			Traffic	0.9 (6%)	4.0 (27%)	0.85	2445
			Any	0.6 (4%)	3.8 (29%)	0.86	4788
$EXP_{2016-2019}$	Training	01/01-31/12	Urban background	0.0 (0%)	1.9 (20%)	0.95	146237
			Traffic	0.0 (0%)	2.5 (17%)	0.95	151372
			Any	0.0 (0%)	2.2 (18%)	0.96	297609
	Test	01/01-13/03	Urban background	0.2 (2%)	3.7 (32%)	0.84	9437
		14/03-23/04	Urban background	0.5 (6%)	3.6 (41%)	0.75	5408
		01/01-23/04	Urban background	0.3 (3%)	3.6 (35%)	0.83	14845
		01/01-13/03	Traffic	0.1 (0%)	4.3 (25%)	0.85	9741
		14/03-23/04	Traffic	0.4 (3%)	4.4 (33%)	0.78	5689
		01/01-23/04	Traffic	0.2 (1%)	4.3 (28%)	0.83	15430
		01/01-13/03	Any	0.1 (1%)	4.0 (28%)	0.86	19178
		14/03-23/04	Any	0.5 (4%)	4.0 (37%)	0.80	11097
		01/01-23/04	Any	0.2 (2%)	4.0 (31%)	0.85	30275

Author contributions. Contributed to conception and design: HP, CPG-P. Contributed to acquisition of data: DB, KS. Contributed to analysis and interpretation of data: HP, DB, CPG-P, MG, AS, OJ. Drafted the article: HP, CPG-P.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433. We also acknowledge support by the European Research Council (grant no. 773051, FRAGMENT), the AXA Research Fund, the Spanish Ministry of Science, Innovation and Universities (RYC-2015-18690, CGL2017-88911-R, RTI2018-099894-B-I00 and Red Temática ACTRIS España CGL2017-90884-REDT), the BSC-CNS "Centro de Excelencia Severo Ochoa 2015-2019" Program (SEV-2015-0493), PRACE for awarding us access to Marenostrum Supercomputer in the Barcelona Supercomputing Center, and H2020 ACTRIS IMP (871115).

References

Achebak, H., Petetin, H., Quijal-Zamorano, M., Bowdalo, D., García-Pando, C. P., and Ballester, J.: Reduction in air pollution and attributable mortality due to COVID-19 lockdown, The Lancet Planetary Health, 4, e268, https://doi.org/10.1016/S2542-5196(20)30148-0, https://linkinghub.elsevier.com/retrieve/pii/S2542519620301480, 2020.

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D.: How will country-based mitigation measures influence the course of the COVID-19 epidemic?, The Lancet, 395, 931–934, https://doi.org/10.1016/S0140-6736(20)30567-5, https://linkinghub.elsevier.com/retrieve/pii/S0140673620305675, 2020.

Bauwens, M., Compernolle, S., Stavrakou, T., Müller, J., Gent, J., Eskes, H., Levelt, P. F., van der A, R., Veefkind, J. P., Vlietinck, J., Yu, H., and Zehner, C.: Impact of Coronavirus Outbreak on NO 2 Pollution Assessed Using TROPOMI and OMI Observations, Geophysical Research Letters, 47, https://doi.org/10.1029/2020GL087978, https://onlinelibrary.wiley.com/doi/abs/10.1029/2020GL087978, 2020.

Bowdalo, D.: Globally Harmonised Observational Surface Treatment: Database of global surface gas observations, in preparation.

Castellanos, P. and Boersma, K. F.: Reductions in nitrogen oxides over Europe driven by environmental policy and economic recession, Scientific Reports, 2, 265, https://doi.org/10.1038/srep00265, http://www.nature.com/articles/srep00265, 2012.

Center for International Earth Science Information Network - CIESIN - Columbia University: Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11 [data set], Tech. rep., Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)., https://doi.org/10.7927/H49C6VHW, 2018.

Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, 2017.

Dunlea, E. J., Herndon, S. C., Nelson, D. D., Volkamer, R. M., San Martini, F., Sheehy, P. M., Zahniser, M. S., Shorter, J. H., Wormhoudt, J. C., Lamb, B. K., Allwine, E. J., Gaffney, J. S., Marley, N. A., Grutter, M., Marquez, C., Blanco, S., Cardenas, B., Retama, A., Ramos Villegas, C. R., Kolb, C. E., Molina, L. T., and Molina, M. J.: Evaluation of nitrogen dioxide chemiluminescence monitors in a polluted urban environment, Atmospheric Chemistry and Physics, 7, 2691–2704, https://doi.org/10.5194/acp-7-2691-2007, http://www.atmos-chem-phys.net/7/2691/2007/, 2007.

EEA: Air Quality e-Reporting Database, European Environment Agency (http://www.eea.europa.eu/data- and-maps/data/aqereporting-8) (accessed 1 May 2020), 2020.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine., The Annals of Statistics, 29, 1189–1232, https://doi.org/10.1214/aos/1013203451, http://projecteuclid.org/euclid.aos/1013203451, 2001.

Grange, S. K. and Carslaw, D. C.: Using meteorological normalisation to detect interventions in air quality time series, Science of The Total Environment, 653, 578–588, https://doi.org/10.1016/j.scitotenv.2018.10.344, https://linkinghub.elsevier.com/retrieve/pii/S004896971834244X, 2019.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss PM 10 trend analysis, Atmospheric Chemistry and Physics, 18, 6223–6239, https://doi.org/10.5194/acp-18-6223-2018, https://www.atmos-chem-phys.net/18/6223/2018/, 2018.

Guevara, M., Tena, C., Jorba, O., and García-Pando, C. P.: HERMESv3_BU model (Version v0.1.1), Zenodo, https://doi.org/10.5281/zenodo.3521897, 2019.

Guevara, M., Jorba, O., Soret, A., Petetin, H., Bowdalo, D., Serradell, K., Tena, C., Denier van der Gon, H., Kuenen, J., Peuch, V.-H., and Pérez García-Pando, C.: Time-resolved emission reductions for atmospheric chemistry modelling in Europe during the COVID-19 lockdowns (in review), Atmospheric Chemistry and Physics Discussions, https://doi.org/10.5194/acp-2020-686, 2020a.

Guevara, M., Tena, C., Porquet, M., Jorba, O., and Pérez García-Pando, C.: HERMESv3, a stand-alone multi-scale atmospheric emission modelling framework - Part 2: The bottom-up module, Geoscientific Model Development, 13, 873–903, https://doi.org/10.5194/gmd-13-873-2020, https://www.geosci-model-dev.net/13/873/2020/, 2020b.

Menut, L., Bessagnet, B., Siour, G., Mailler, S., Pennel, R., and Cholakian, A.: Impact of lockdown measures to combat Covid-19 on air quality over western Europe, Science of The Total Environment, 741, 140 426, https://doi.org/10.1016/j.scitotenv.2020.140426, https://linkinghub.elsevier.com/retrieve/pii/S0048969720339486, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Rao, S. T., Galmarini, S., and Puckett, K.: Air Quality Model Evaluation International Initiative (AQMEII): Advancing the State of the Science in Regional Photochemical Modeling and Its Applications, Bulletin of the American Meteorological Society, 92, 23–30, https://doi.org/10.1175/2010BAMS3069.1, http://journals.ametsoc.org/doi/abs/10.1175/2010BAMS3069.1, 2011.

Tobías, A., Carnerero, C., Reche, C., Massagué, J., Via, M., Minguillón, M. C., Alastuey, A., and Querol, X.: Changes in air quality during the lockdown in Barcelona (Spain) one month into the SARS-CoV-2 epidemic, Science of The Total Environment, 726, 138 540, https://doi.org/10.1016/j.scitotenv.2020.138540, https://linkinghub.elsevier.com/retrieve/pii/S0048969720320532, 2020.

Villena, G., Bejan, I., Kurtenbach, R., Wiesen, P., and Kleffmann, J.: Interferences of commercial NO 2 instruments in the urban atmosphere and in a smog chamber, Atmospheric Measurement Techniques, 5, 149–159, https://doi.org/10.5194/amt-5-149-2012, https://www.atmos-meas-tech.net/5/149/2012/, 2012.