**Authors' Response to Reviewer 2 of GMDD Manuscript "*Operational chemical weather forecasting with the ECCC online Regional Air Quality Deterministic Prediction System version 023 (RAQDPS023) – Part 2: Multi-year prospective and retrospective performance evaluation*" by Moran et al. (2026)**

**General comments:**

The RAQDPS is one of the main North-American chemical weather / air quality forecasting systems. A detailed documentation of both the modelling aspects as well as the scores from the comparison with observations at the surface are valuable for the large group of users of the forecasts, and is provided by the part 1 and 2 of this article. The paper part 2 by Moran et al. is providing a very comprehensive evaluation of the skill of the modelling system to predict the main air pollutants ozone, NO2, PM2.5, relevant for health. Apart from this the authors compare to available other trace gas observations as well as measurements of the chemical composition of the aerosols in air and in precipitation. This extra set of comparisons give important additional information on the skill of the model concerning the emissions and chemical components that contribute to the formation of the main pollutants. The scores are compared to benchmarks for chemistry modelling and to other AQ prediction efforts, which puts the scores in perspective. In particular section 4.4 is a highlight of the paper because it identifies processes that may be responsible for differences with the observations, and discusses potential improvements in the model. The authors are honest in documenting the comparisons, showing skill for ozone and NO2, but more mixed results for PM2.5 (Fig. 8, 11, 15).

The paper is a very complete reference to the skill of the RAQDPS model to simulate surface concentrations. It is well written, and has a good list of important references. At the same time it is also very long, and even interested / motivated readers may be hesitant to go through all the results. Nevertheless, the paper, together with part 1, serves as a complete reference for RAQDPS. As such I am in favour of publication after the authors have answered my more minor comments and questions.

We thank the reviewer for their thoughtful assessment of this manuscript, for their constructive comments and questions, and for their patience in reviewing a long manuscript. We respond below (in blue font) to their specific comments and questions (in black font). *Italics* have been used to denote selections of manuscript text, **boldface** is used to indicate changes to the manuscript text, and red font denotes new references.

**Specific comments and questions:**

l 63: "expanded rapidly over the last two decades". I found that the discussion of air quality forecasting systems worldwide, as provided in the introduction, was providing a valuable historic overview of relevant literature but the references are not fully up-to-date and may be expanded to other countries/continents (e.g. China, Japan, Europe, global-scale forecasting). For instance, in Europe the regional air quality forecasting is organised by the Copernicus Atmosphere Monitoring Service (CAMS) and is described in a recent paper by Colette et al, https://doi.org/10.5194/gmd-18-6835-2025, which I propose to add to the references.

Response: This is a good suggestion. We have added the Colette et al. (2025) reference to this sentence as well as references for East Asian, U.S., and global-scale forecasting (Kajino et al., 2018; Wang et al., 2022; Williams et al., 2022; Li et al., 2025).

Kajino, M., Deushi, M., Sekiyama, T. T., Oshima, N., Yumimoto, K., Tanaka, T. Y., Ching, J., Hashimoto, A., Yamamoto, T., Ikegami, M., Kamada, A., Miyashita, M., Inomata, Y., Shima, S., Adachi, K., Zaizen, Y., Igarashi, Y., Ueda, H., Maki, T., and Mikami, M.: NHM-Chem, the Japan Meteorological Agency's regional meteorology – chemistry model (v1.0): Model description and aerosol representations, https://doi.org/10.5194/gmd-2018-128, 21 June 2018.

Li et al. (2025) was already cited elsewhere in the manuscript.

Wang, H., Zhang, X. Y., Wang, P., Peng, Y., Zhang, W. J., Liu, Z. D., Han, C., Li, S. T., Wang, Y. Q., Che, H. Z., Huang, L. P., Liu, H. L., Zhang, L., Zhou, C. H., Ma, Z. S., Chen, F. F., Ma, X., Wu, X. J., Zhang, B. H., and Shen, X. S.: Chemistry-weather interacted model system GRAPES_Meso5.1/CUACE CW V1.0: Development, evaluation and application in better haze/fog prediction in China, J. Adv. Model. Earth Syst., 14, e2022MS003222, https://doi.org/10.1029/2022MS003222, 2022.

Williams et al. (2022) was already cited elsewhere in the manuscript.

l 145: "dynamic; and probabilistic" evaluations are not discussed in this paragraph.

Response: As noted by the reviewer, this was a gap in the manuscript, especially as it was then stated on line 195 that the present study "constituted a dynamic evaluation of opportunity". We have added the following two sentences after line 145 and four new references to the manuscript:

"***Dynamic evaluations, the third of these evaluation types, assess model skill in quantifying the impact of changes in input emissions or meteorology (e.g., Gilliland et al., 2008; Godowitch et al., 2010; Foley et al., 2015). And fourth, probabilistic evaluations examine the uncertainty and level of confidence in model predictions (e.g., Hanna et al., 2005; Mallet and Sportisse, 2006; Galmarini et al., 2010; Kioutsioukis et al., 2025).***"

Galmarini, S., Bonnardot, F., Jones, A., Potempski, S., Robertson, L., and Martet, M.: Multi-model vs. EPS-based ensemble atmospheric dispersion simulations: A quantitative assessment on the ETEX-1 tracer experiment case, Atmos. Environ., 44, 3558–3567, https://doi.org/10.1016/j.atmosenv.2010.06.003, 2010.

Gilliland et al. (2008) was already cited elsewhere in the manuscript.

Godowitch et al. (20010) was already cited elsewhere in the manuscript.

Foley et al. (2015) was already cited elsewhere in the manuscript.

Hanna, S. R., Russell, A. G., Wilkinson, J. G., Vukovich, J., and Hansen, D. A.: Monte Carlo estimation of uncertainties in BEIS3 emission outputs and their effects on uncertainties in chemical transport model predictions, J. Geophys. Res., 110, D01302, 15 pp., https://doi.org/10.1029/2004JD004986, 2005.

Kioutsioukis, I., Hogrefe, C., Makar, P. A., Alyuz, U., Bash, J. O., Bellasio, R., Bianconi, R., Butler, T., Clifton, O. E., Cheung, P., Hodzic, A., Kranenburg, R., Lupascu, A., Momoh, K., Perez-Camaño, J. L., Pleim, J., Ryu, Y.-H., San Jose, R., Schwede, D., Sokhi, R., and Galmarini, S.: Operational, diagnostic, and probabilistic evaluation of AQMEII-4 regional-scale ozone dry deposition: time to harmonize our LULC masks, Atmospheric Chem. Phys., 25, 12923–12953, https://doi.org/10.5194/acp-25-12923-2025, 2025.

Mallet, V. and Sportisse, B.: Ensemble-based air quality forecasts: A multimodel approach applied to ozone, J. Geophys. Res. Atmospheres, 111, 2005JD006675, https://doi.org/10.1029/2005JD006675, 2006.

l 305: "Note also that neither system version considered some other types of natural emissions, .." This may be formulated more clearly.

Response: A related concern was raised by one of the reviewers of the Part 1 manuscript (Moran et al., 2026a). This concern was addressed in that manuscript by adding more detail about these other types of natural emissions and their relative magnitudes and importance. For this manuscript we have appended the following text to the sentence identified by the reviewer:

"*... and volcanic emissions,* ***but these sources are assumed to be less important for Canada (see Moran et al., 2026a)***."

l 356: "Two automated data filters .." Stations close to major roads or other big emitters will not be representative for a model grid cell of 10km diameter. Is a site classification available for the stations in the US and Canada? Did the authors consider to remove stations that will not give representative comparison results? (Urban vs rural statistics is discussed in line 475)

Response: Measurements from stations close to major roads were removed as a preliminary step prior to pairing model predictions with measurements. This step was noted in Sect. S2.4 (lines 171–173) as follows:

"*For this reason, all measurements from roadside monitors have been removed from the AQS and NAPS measurement data sets (based on station metadata) before the pairing step to avoid the obvious representativeness problem relative to a regional-scale model's larger grid cells.*"

However, as noted by the reviewer this is an important point. In order to raise it in the main body of the paper, two sentences (bold font) have been added in Sect. 2.3:

"*Values flagged as suspicious or invalid were not used for this study.* ***Measurements from stations located near roadways were discarded as well to ensure spatial representativeness****. A temporal completeness criterion was also imposed on this NRT measurement data set to ensure temporal representativeness of the evaluation data: individual station data sets were required to have at least 75% valid values out of the total possible values for a one-year evaluation period to be considered complete.* ***More details about these representativeness constraints are provided in Sect. S2.4.***"

Sec 2.3: As far as I understand, AirNow is employing NO2 analysers which use molybdenum converters. The resulting NO2 measured values are known to overestimate NO2, because the technique is also sensitive to other NOy components which are also converted in the instrument. Please discuss this aspect and how it influences the NO2 evaluation.

Response: The reviewer raises an important point which should be mentioned to provide context to the interpretation of the $NO_2$ scores. The following text has been added in Sect. 3.2.1, where $NO_2$ statistical scores are discussed for the first time, and is then referred to elsewhere in the paper and Supplement in later discussions of $NO_2$ scores (e.g., lines 562, 578):

"*...* ***It was also noted in Sect. 2.3 that most network $NO_2$ monitors in North America suffer from positive biases due to interference from other oxidized nitrogen species (e.g., Dunlea et al., 2007; Dickerson et al., 2019). These biases, however, vary greatly with time and location: they***

***are expected to be smallest when fresh $NO_2$ emissions dominate, for example, in urban areas, in the early morning hours (4-9 a.m.) when the PBL is shallow, and in the winter, but largest relatively speaking for aged air, as in rural areas, in the afternoon hours, and in the summer (e.g., Godowitch et al., 2010; Lamsal et al., 2015; Jaeglé et al., 2018; He et al., 2019; Toro et al., 2021). Thus, the annual NMB value for $NO_2$ of -0.19 would be smaller if these measurement biases were accounted for."***

Dickerson, R. R., Anderson, D. C., and Ren, X.: On the use of data from commercial NOx analyzers for air pollution studies, Atmos. Environ., 214, 116873, https://doi.org/10.1016/j.atmosenv.2019.116873, 2019.

Dunlea et al. (2007) was already cited elsewhere in the manuscript.

Godowitch et al. (2010) was already cited elsewhere in the manuscript.

He, H., Vinnikov, K. Y., Krotkov, N. A., Edgerton, E. S., Schwab, J. J., and Dickerson, R. R.: Chemical climatology of atmospheric pollutants in the eastern United States: Seasonal/diurnal cycles and contrast under clear/cloudy conditions for remote sensing, Atmos. Environ., 206, 85–107, https://doi.org/10.1016/j.atmosenv.2019.03.003, 2019.

Jaeglé, L., Shah, V., Thornton, J. A., Lopez-Hilfiker, F. D., Lee, B. H., McDuffie, E. E., Fibiger, D., Brown, S. S., Veres, P., Sparks, T. L., Ebben, C. J., Wooldridge, P. J., Kenagy, H. S., Cohen, R. C., Weinheimer, A. J., Campos, T. L., Montzka, D. D., Digangi, J. P., Wolfe, G. M., Hanisco, T., Schroder, J. C., Campuzano-Jost, P., Day, D. A., Jimenez, J. L., Sullivan, A. P., Guo, H., and Weber, R. J.: Nitrogen oxides, emissions, chemistry, deposition, and export over the northeast United States during the WINTER aircraft campaign, J. Geophys. Res. Atmospheres, 123, https://doi.org/10.1029/2018JD029133, 2018.

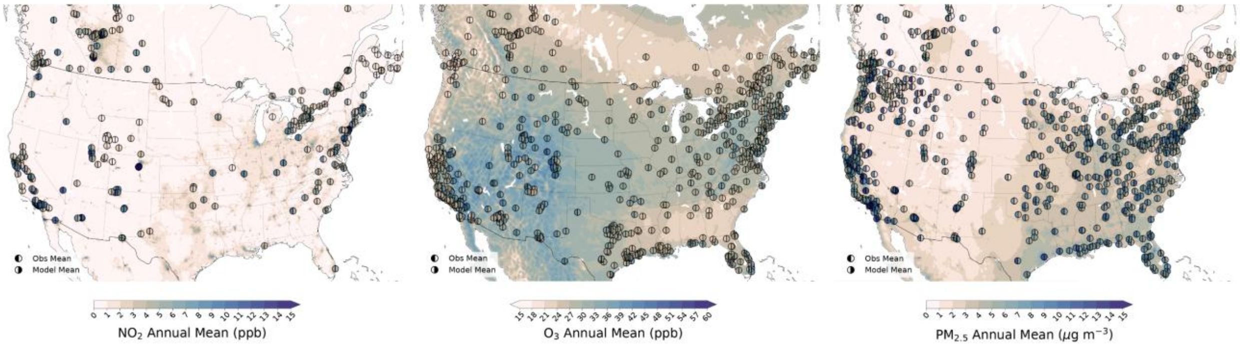
Lamsal et al. (2015) was already cited elsewhere in the manuscript.

Toro et al. (2021) was already cited elsewhere in the manuscript.

Fig.1. Looking at the figure it seems that the model values at the stations does not match the continental modelled distribution, especially for NO2 and PM where model-at-station values seem to be higher than the background? This figure was confusing to me.

Response: The reviewer's comment is understandable. There were two contributing factors, one simple and one more significant. The first factor was that the resolution of the original figure was degraded when it was inserted into the manuscript. The size of the original figure is 2557 x 1308 pixels, but the size of the manuscript version is 691 x 195 pixels. The second factor is that for the model horizontal grid spacing of 10 km, large concentrated emissions sources such as large point sources and urban centres result in localized surface concentration maxima with strong horizontal gradients for primary pollutants such as $NO_2$ and primary-secondary pollutants such as $PM_{2.5}$.

The value of this novel "coffee bean" presentation is that these strong local gradients are more visible when the image is viewed at a higher resolution and the coffee beans provide an even higher resolution as they show 10 km *grid-cell* values paired with the observed value for the measurement station located in that grid cell. The image below shows the same figure but inserted into this response while preserving its full resolution. If the image is expanded by a factor of five or more, many sub-regional features become more visible and the agreement between model predictions and observations looks more reasonable at many locations. The size of the original figure is within the size limits specified by the journal and it will be provided to the journal instead of the degraded review version, which should address the reviewer's concern.




To help the reader the sentence referring to this figure in Sect. 3.2.1 has been expanded as follows:

"*Generally good agreement is evident in Fig. 1 between the observed and predicted annual mean values of both pollutants* ***(although viewing the $NO_2$ panel with higher magnification is helpful);*** ***note that*** *the higher $NO_2$ VMRs associated with urban centers are smaller-scale features caused by high $NO_x$ emissions over urban centers (cf. Fig. S1a).*"

l 637: "NME scores"  Should this be NMAE?  Please use consistent acronyms.

Response:  The reviewer is correct that NME and NMAE have been used interchangeably in the manuscript since they were understood by the authors to be synonymous and different from NMB. However, it appears from an internet search that some publications treat NME and NMB as being synonymous and different from NMAE (e.g., Janssen and Heuberger, 1995, https://doi.org/10.1016/0304-3800(95)00084-9; Tan et al., 2016, https://doi.org/10.1175/JHM-D-16-0079.1).  The manuscript has been revised to use NMAE consistently in place of NME.

l 656: "U.S. NOx emissions used for these forecasts may have been too low ". Are there other indications, e.g. from literature, that this is the case?

Response:  Thank you for this question.  Given the time that has passed since the RAQDPS-OP023 became operational in 2021, retrospective anthropogenic annual national emissions inventories for 2021 are now available.  Based on a review of these inventories, it appears that the $NO_x$ emissions considered by the RAQDPS-OP023 in 2021 and 2022 for the U.S. (and for Canada) were actually too high rather than too low.  The manuscript text has now been revised to reflect this additional information as follows:

"*A seventh finding was that annual, seasonal, and monthly negative NMB values were considerably larger for $NO_2$ in the 2021/22 forecasts for the two U.S. regions (-0.31, -0.25) vs. the two Canadian regions (-0.06, -0.03).  The same pattern was not seen in the 2013 2016 hindcasts, which used retrospective rather than projected emissions (Table S7).  This is one example of the value of computing and then comparing statistics for the individual networks as well as the combined networks (e.g., Tables S3A, S3S).  The differences in biases between countries found for the 2021/22 forecasts might suggest that the U.S. $NO_x$ emissions used for these forecasts were too low, whereas the Canadian $NO_x$ emissions that were used were more representative of 2021/22 conditions.* ***Since some time has passed since 2021 when the RAQDPS-OP023 became operational, retrospective national emissions inventories are now available for 2021 for both Canada and***

***the U.S. (ECCC, 2025a; U.S. EPA, 2025). Interestingly, a comparison of these inventories with Table 1 showed that the projected annual $NO_x$ inventory emissions considered for both countries were in fact higher than the actual inventory values for 2021: 8% higher for the U.S. and 13% higher for Canada. Other explanations must thus be sought.***"

ECCC: Canada's Air Pollutant Emissions Inventory Report 1990–2023, Environment and Climate Change Canada, Gatineau, Quebec, 72 pp., https://publications.gc.ca/collections/collection_2025/eccc/En81-30-2023-eng.pdf, 2025a.

U.S. EPA: Air Pollutant Emissions Trends Data: Criteria pollutants National Tier 1 for 1970 - 2024, Data set, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, https://www.epa.gov/system/files/other-files/2025-04/national_tier1_caps_21feb2025.xlsx, 2025.

l 776: "The fact that annual NMB scores for gravimetric PM2.5 mass measurements are more positive than those for continuous PM2.5 measurements is thus surprising."

After reading the paper these differences and comparisons remain unclear to me. Maybe a few sentences in the concluding section could be added to highlight this aspect once more, and how this impacts the interpretation of the evaluation.

Response: We believe that the use of multiple $PM_{2.5}$ total mass data sets is a strength of this paper, but we acknowledge that it introduces additional complexity. We have followed the reviewer's suggestion by modifying lines 1588–1594 of the Summary and Conclusions section as follows:

"*The evaluation of $PM_{2.5}$ predictions for the 2013–2016 annual hindcasts was expanded by considering two additional types of $PM_{2.5}$ measurements available from multiple networks, namely (i) daily gravimetric $PM_{2.5}$ total mass measurements* ***from the U.S. $PM_{2.5}$ mass monitoring network*** *and (ii) daily $PM_{2.5}$ chemical composition measurements* ***from the CSN, IMPROVE, and NAPS networks****. These additional $PM_{2.5}$ data sets* ***included measurements from nearly as many stations as the hourly $PM_{2.5}$ total mass data sets but employed very different measurement technologies with different error characteristics. By comparing and contrasting evaluation results for these multiple networks,*** *insights* ***could be drawn*** *into causes of the hourly $PM_{2.5}$ mass underpredictions. One finding was that RAQDPS-OP023 predictions of daily gravimetric $PM_{2.5}$ total mass were less negatively biased than those for hourly $PM_{2.5}$ mass, which* ***immediately*** *draws attention to instrument and analysis differences between the two types of measurements.* ***Unlike hourly continuous $PM_{2.5}$ total mass measurements, the daily gravimetric $PM_{2.5}$ total mass measurements are known to have negative artefacts due to the volatilization of semi-volatile species such as ammonium, nitrate, and particle water. This difference suggests that hourly continuous $PM_{2.5}$ measurements might be higher than gravimetric measurements if those measurements were taken at the same time and location, thus resulting in a larger negative bias for the former. Monthly evaluation scores for the daily gravimetric $PM_{2.5}$ total mass measurements were also closer to those for the CSN and NAPS networks than the IMPROVE network, which is consistent with the first three networks being urban-focused whereas the IMPROVE network is rural-focused***."

l 838: "but these scores were confounded by the model's inclusion of isoprene oxidation products in this lumped VOC species (Moran et al., 2025), suggesting that overpredictions should be expected."

Is there an estimate of the ratio of pure isoprene to lumped "isoprene", e.g. is a factor of more than 3 for the model/obs ratio to be expected?

Response: The model VOC species ISOP is in fact not lumped and corresponds to pure $C_5H_8$ whereas ETHE is the lumped VOC species that includes isoprene oxidation products. However, as the reviewer notes, the ratios of annual mean predicted ISOP to annual mean observed isoprene listed in Table 4 for 2013-2016 are all greater than a factor of 3, that is, very large overpredictions. This behaviour is not expected and it strongly suggests the presence of a systematic model error (while noting that these ratios are based on a small sample of 6 to 13 measurement stations: Table S2b). This is an important result, and as discussed in Sects. 4.4 and S3.3.1 the likeliest explanation is too high emissions of isoprene.

The ratios of annual mean predicted ETHE VMR to annual mean observed ethene VMR listed in Table 4 for 2013-2016 range from 1.68 to 2.19, again corresponding to a large overprediction. However, as shown in Table S1, annual anthropogenic ETHE emissions on the model grid were only 249 ktonnes versus annual biogenic ISOP emissions of 35,343 ktonnes, so it is very plausible that ETHE levels were enhanced by isoprene oxidation products. In addition, Table S1 shows that the seasonal split of anthropogenic ETHE emissions is 0.33 in winter, 0.24 in spring, 0.21 in summer, and 0.23 in autumn, i.e., a winter emissions maximum, versus the seasonal split of biogenic ISOP emissions of 0.02, 0.17, 0.65, and 0.16, i.e., a summer emissions maximum. If we then look at Table S4S, we see that seasonal NMB for ETHE VMR for 2013-2016 ranges from -0.06 to 0.13 in the winter but from 1.71 to 3.21 in the summer. That is, seasonal bias for ETHE VMR is small in the winter when ISOP emissions are smallest but very large in the summer when ISOP emissions are largest.

The need for special consideration of ETHE and HCHO scores due to these model species being lumped VOC species was discussed in Sect. S2.4 but only obliquely in the main manuscript in Sect. 3.3.1. The following sentences have now been added at the end of Sect. 2.4 to make this issue more visible to the reader before the evaluation sections:

"***For example, one important consideration related to pairing with VOC measurements is that while measurements of both ethene ($C_2H_4$) and formaldehyde (HCHO) are available, the ADOM-2 model VOC species named ETHE and HCHO are both lumped species that include additional VOC species in addition to ethene and formaldehyde. This misalignment must be considered when interpreting evaluation results for these two model species in Sect. 3.3.1 since the measurements are in effect lower bounds on the model predictions rather than direct counterparts***."

One related sentence was also added to Sect. 3.3.1:

"***However,*** *winter ETHE scores, when isoprene emissions were low* ***(Table S1)****, were better than those for the other three seasons****, when isoprene emissions were higher*** *(Table S4S).* ***For example, winter NMB values ranged from -6% to 13%, suggesting that predictions of pure ethene VMR driven by pure ethene emissions demonstrated skill for that season.***"

l 1226, Fig 19: This figure is a key result to document modelling improvements over time, showing improvements in O3 and NO2 scores, but not for PM2.5.

l1234: "The 2010–2019 seasonal R scores for PM2.5, on the other hand, are lower than those for NO2 and O3, are less cyclical seasonally, and do not show any improvement with time."

l 1256, wildfire: Important to improve the scores, especially PM2.5 in Summer. Is it considered to use RAQDPS-FW023 to replace RAQDPS023 as operational system?

Response: Some background information is required to answer this question. Early versions of the RAQDPS with wildfire emissions were experimental and their predictions of $NO_2$ and $O_3$ were worse than the standard operational version of the RAQDPS without wildfire emissions. As a consequence, for the first operational version of the RAQDPS with wildfire emissions, RAQDPS-FW015, which was released in April 2016 just before the major Fort McMurray fire in northern Alberta in May 2016, only its $PM_{2.5}$ forecast fields were considered in order to provide wildfire smoke plume forecasts, and $NO_2$ and $O_3$ forecasts continued to be provided by the RAQDPS-OP015. The first RAQDPS-FW version with competitive $NO_2$ and $O_3$ forecasts was RAQDPS-FW020.2, which was released in April 2019. This was the first time that the possibility of replacing the RAQDPS-OP system with the RAQDPS-FW system could even be considered, and both versions still needed to be run in order to provide wildfire smoke plume forecasts, which had become of greater interest due to increases in wildfire frequency and magnitude over the past decade. It was then decided to wait to replace the RAQDPS-OP forecasts with RAQDPS-FW forecasts until such time as only one of these systems needed to be run while still providing wildfire smoke plume forecasts, that is, a merger of the two systems. As this requirement demanded further research and development, adoption by ECCC of the wildfire version as the standard operational AQ forecast system had to wait until June 2024 when the RAQDPS-OP025 was released (discussed in the companion paper to this paper). Thus, the answer is "yes" to the reviewer's question. An additional response is provided below to the reviewer's final comment.

l 1442: Could the boundary conditions for ozone be partly responsible for ozone biases and east-west gradients? Please discuss.

Response: Yes, the ozone boundary conditions are very likely implicated, and as discussed in the companion paper to this paper, some improvements were made to the ozone boundary conditions in the RAQDPS-OP025 that resulted in reduced $O_3$ biases. Lines 1447–1449 in this paper have been modified as follows to make this explanation more explicit:

"*Different explanations for these systematic springtime and western underpredictions are possible, but one obvious candidate for investigation is the model's treatment of $O_3$ chemical lateral boundary conditions (e.g., Pendlebury et al., 2018; Moran et al., 2026a),* ***where $O_3$ levels at the western boundary in the spring and summer may be too low***."

l 1493: Wet deposition modelling is mentioned as cause for low NH3. What about dry deposition, which is a dominant removal mechanism? Please discuss.

Response: Gaseous $NH_3$ has three possible removal pathways: dry deposition; wet deposition; and gas-particle partitioning. The dry deposition of $NH_3$ was discussed on lines 1469–1470 after a discussion of the $SO_2$ dry deposition parameterization, where it was noted that $SO_2$ dry deposition would increase if two removal pathways for $SO_2$ dry deposition that were incorrectly neglected in the RAQDPS-OP023 were added back. Due to $SO_2$ being the archetypal species for both $NH_3$ and $HNO_3$ dry deposition in the RAQDPS-OP023 treatment of dry deposition, it was then suggested that $NH_3$ dry deposition was likely *too low* in the RAQDPS-OP023, although increasing $NH_3$ dry deposition would of course decrease $NH_3$ levels even more. By contrast, the overpredictions of $PM_{2.5}$-$NH_4$ air concentration, $NH_4^+$ concentration in precipitation, and $NH_4^+$ wet deposition that were noted in lines 1483–1505 suggest that the other two removal pathways for $NH_3$ might have contributed to the $NH_3$ VMR underpredictions. For the gas-particle partitioning removal pathway for $NH_3$, it was noted that $HNO_3$ levels were also overpredicted. Reductions in $HNO_3$ levels, which would occur with the two modifications to $SO_2$ dry deposition by increasing $HNO_3$ dry deposition, would thus reduce the amount of $NH_3$ that is partitioned to the particle phase. And the wet deposition pathway will remove both gaseous $NH_3$ and $PM_{2.5}$-$NH_4$, but since $PM_{2.5}$-$NH_4$ is overpredicted, the other potential cause of too low $NH_3$ levels is thus too high removal of $NH_3$ by washout and rainout. This discussion is admittedly complicated, but we believe the current text does discuss all three removal pathways for $NH_3$.

In section 4.4 there could be more attention for transport aspects, vertical mixing, stability. The evaluation focusses on the surface but vertical profiles may have issues as well. A few remarks on this would be welcome.

Response: The goal for Sect. 4.4 was to discuss possible model improvements that were suggested by an assessment of the full set of evaluation results from Sects. 3 and S3. Only one meteorological evaluation was considered, that for precipitation which was discussed in Sect. 3.1. Other meteorological factors such as horizontal transport and vertical mixing are definitely important in AQ modelling and forecasting, but they were not considered in this study. The same comment applies to the vertical distribution of pollutants, for which relatively few studies have been published given the scarcity of available above-surface measurements. These limitations have now been acknowledged by adding the following new paragraph at the end of Sect. 4.4:

"***This section has discussed possible improvements to the RAQDPS-OP023 as suggested by an assessment and analysis of the full set of evaluation results from Sects. 3 and S3. Note that it has not been possible, however, to consider all aspects of the model. For example, due to limitations on the availability of measurement data, performance scores were examined for only one meteorological quantity, precipitation, even though, as noted in Sect. 3.1, other meteorological factors such as horizontal transport and vertical mixing are also very important in AQ modelling and forecasting. The same comment applies to the vertical distributions of pollutants, which are also of interest when assessing model performance but for which relatively few studies have been published (e.g., Solazzo et al., 2013; Astitha et al., 2018) given the scarcity of available measurements in the vertical. While acknowledging these limitations, it is still possible to extend the model performance assessment through inference. For example, it was noted in the discussion of Fig. 17 in Sect. 3.3.2 that the times of day when the maxima of primary pollutants occur is determined by the balance between surface emissions and vertical stability and mixing. Figures 9, S165, S168, S228 S238, and S241 show all-station annual-mean, seasonal-mean, and annual-mean regional diurnal time series for $NO_2$ VMR. In all of these figures, the timing of the twice-daily maxima in $NO_2$ VMR is well represented by the model. This suggests that both the diurnal variations in emissions and in vertical stability and mixing have been captured well by the RAQDPS-OP023. As a second example, it was noted in Sect. 2.1 that the injection of surface emissions into the lowest layer of the atmosphere is handled differently by the RAQDPS-OP023 forecasts for 2021/22 versus the hindcasts for 2013–2016, leaving open the possibility that unrealistically high surface pollutant abundances might have been predicted in the hindcasts for primary pollutants under conditions of strong near-surface vertical stability and low wind speeds. However, examination of the five years of mid-season monthly density scatterplots for $NO_2$ in Fig. S169 show little difference in the January scatterplots for the highest $NO_2$ monthly values between January 2022 and the other four Januarys. This comparison thus provides at least an indication that unrealistically high $NO_2$ surface VMR values were very infrequent even if they did occur.***"

Astitha, M., Kioutsioukis, I., Fisseha, G. A., Bianconi, R., Bieser, J., Christensen, J. H., Cooper, O. R., Galmarini, S., Hogrefe, C., Im, U., Johnson, B., Liu, P., Nopmongcol, U., Petropavlovskikh, I., Solazzo, E., Tarasick, D. W., and Yarwood, G.: Seasonal ozone vertical profiles over North America using the AQMEII3 group of air quality models: model inter-comparison and stratospheric intrusions, Atmospheric Chem. Phys., 18, 13925–13945, https://doi.org/10.5194/acp-18-13925-2018, 2018.

Solazzo, E., Bianconi, R., Pirovano, G., Moran, M. D., Vautard, R., Hogrefe, C., Appel, K. W., Matthias, V., Grossi, P., Bessagnet, B., Brandt, J., Chemel, C., Christensen, J. H., Forkel, R., Francis, X. V., Hansen, A. B., McKeen, S., Nopmongcol, U., Prank, M., Sartelet, K. N., Segers, A., Silver, J. D., Yarwood, G., Werhahn, J., Zhang, J., Rao, S. T., and Galmarini, S.: Evaluating the capability of regional-scale air quality models to capture the vertical distribution of pollutants, Geosci. Model Dev., 6, 791–818, https://doi.org/10.5194/gmd-6-791-2013, 2013.

Sec 4.4: May be good to add a remark on the injection height of emissions in this section. First layer vs. first two layers was discussed earlier.

Response: Agreed. See response to previous comment.

Will RAQDPS-FW023 with biomass burning emissions replace the RAQDPS forecasting system in the future? Please discuss pros and cons.

Response: Yes, the RAQDPS025, which was went into operation in June 2024, included biomass burning emissions and merged the RAQDPS-OP025 and RAQDPS-FW025 systems into one system. This merger is discussed in the companion paper by Moran et al. (2026a), but in addition the following short paragraph has been added to the end of the Summary and Conclusions section to better connect the two papers:

"***As a final comment it should be noted that the RAQDPS025, which was introduced operationally at ECCC in June 2024, did implement some of the changes recommended above (see companion paper by Moran et al., 2026a). One very significant change was to merge the two operational system versions, the RAQDPS-OP024 and RAQDPS-FW024, into a single system version that included NRT biomass burning emissions. This particular change had the multiple advantages of improving forecast scores, simplifying maintenance of the operational AQ forecast system, and reducing computer usage.***"