Reply to the reviewer's comments on the manuscript "Country and species-dependent parameters for the Heating Degree Day method to distribute NOx and PM emissions from residential heating in the EU-27: application to air quality modelling and multi-year emission projections".

Guion et al.

The authors would like to thank the reviewers for their careful reading of the paper and for their comments that improved the quality of the manuscript. All comments (in blue) have been addressed and a point-by-point answer is provided in the following (in black after the corresponding comment). The line numbers given in response to comments correspond to the latest version submitted. Finally, modifications made in the new manuscript version are highlighted in the track changes file provided by the authors.

Referee #1

This paper provides parameters to derive HDD-based daily temporal factors to distribute PM and NO_x emissions from residential heating for EU-27 countries. Various experiments with different configurations of HDD-based temporal factors were carried out to assess the sensitivity of simulated $PM_{2.5}$, PM_{10} and NO_2 surface concentration to the different experiments and to identify the best parameterization compared to in situ observations on the one hand and to use HDDs as a method to model national emission totals from GNFR_C and to compare them with persistence and uncertainty on the other hand. The paper is well structured, employs appropriate methodology, and presents reasonable results. It aligns with the scope of Atmospheric Chemistry and Physics. Considering the questions, comments and suggestions described below, I recommend the publication of the paper.

Specific questions and comments:

1. [Section 1] You describe here that the residential sector is responsible for 22.7% and 10.3% on average of the mortality in European cities attributed to $PM_{2.5}$ and NO_2 pollution and that both the urban centres and, due to air advection, the suburban areas are affected. In this context, it would be interesting to know how long the pollutions remain in the atmosphere at a high concentration that is dangerous for people.

The question of the duration of pollution episodes is an interesting but complex one. As each case is specific, it would not be very appropriate to answer by an average number. As described in Morawska et al. (2021), the duration of episodes depends on several complex processes related to accumulation and dispersion mechanisms (meteorological conditions, air transport, landscape geometry) but also to chemical composition (emissions of species with highly varied life spans in the troposphere) and transformations (reaction types and rates). It is a complete research topic in itself.

Morawska, L., Zhu, T., Liu, N., Amouei, Torkmahalleh M., Fatima, De, Andrade, M., Barratt, B., Broomandi, P., Buonanno, G., Carlos Belalcazar Ceron, L., Chen, J., Cheng, Y., Evans, G., Gavidia, M., Guo, H., Hanigan, I., Hu, M., Jeong, C.H., Kelly, F., Gallardo, L., Kumar, P., Lyu, X., Mullins, B.J., Nordstrøm, C., Pereira, G., Querol, X., Yezid Rojas Roa, N., Russell, A., Thompson, H., Wang, H., Wang, L., Wang, T., Wierzbicka, A., Xue, T., Ye, C., 2021. The state of science on severe air pollution episodes: quantitative and qualitative analysis. Environ. Int. 156, 106732 https:// doi.org/10.1016/j.envint.2021.106732.

2. [Section 3.1] Table 2 provides a very good overview for understanding the experiments carried out.

We thank the reviewer for this comment.

3. [Section 4.1.1] In Table 3 the validation scores for the CHIMERE reference simulation calculated from AQ-eReporting observations for $PM_{2.5}$, PM_{10}, NO_2 and O_3 species are shown. It is not entirely clear to me why the bias (model-observation) for annual average is mostly negative. This means that the model always underestimates it. What could be the reason for this?

With the aim to provide a general evaluation of the reference simulation (average monthly profiles in the residential sector), the validation scores presented in Table 3 are averaged over all European stations (only the "background" type and combining "rural", "suburban" and "urban" areas). Indeed, the average bias over the year 2018 is negative for NO_2, PM10 and PM2.5. However, this does not mean that the model "always" underestimates concentrations compared with observations. There are major temporal and spatial disparities.

The table 3 shows that for winter (JFM) and autumn (OND) seasons, the average bias is positive for PM2.5. The bias to observations can also vary daily for each species, as shown in the time series of Figure S4 (in the supplementary material). In addition, these average values encompass a wide spatial diversity. The maps below (Figure A) show the mean annual bias for each European station (used to calculate the values in Table 3). The annual bias for PM10 is, for example, positive in the Netherlands (around +5µg/m^3) and negative in Poland (around -15 µg/m^3). It is important to note that the distribution of stations in Europe is heterogeneous, so that a greater number of stations in one region will have more weight in the average. To better present these values, these discussion elements are added to the manuscript (lines 317-319), as well as the bias maps at the stations (Figure S5 in the supplementary material).

> Lines 317-319: "*It should also be pointed out that these mean values show considerable spatial and temporal variability. The temporal distribution of the daily bias (see Fig. S4 in SM) and the spatial distribution of the stations with their corresponding annual bias (see Fig. S5 in SM) are presented in the supplementary material*."

Deviations from observations can have several causes (related to emissions, meteorology, chemistry) and vary according to the space-time domain and the species. An important point to emphasise is the comparison between the observation stations and the model grids. As detailed in lines 319-324 (see below) of the manuscript, a well-known representativeness bias may lead to a reduction in PM and NO_2 peaks at stations points when the regional CTM simulated average concentrations over 20km grids, especially in heterogeneous grids covering urban areas. The validation presented in Table 3 includes all background stations, whether rural, suburban or urban. By calculating the validation on "rural background" station only, the average bias is considerably better. It decreases from -3.09 µg/m^3 to -2.47 µg/m^3 for PM2.5, from -9.07 µg/m^3 to -7.00 µg/m^3 for PM10, from -7.45 µg/m^3 to -3.11 µg/m^3 for NO_2, from +8.27 µg/m^3 to +2.19 µg/m^3 for O_3.

Lines 319-324: “*The majority of background stations included in the validation calculation are of the urban or suburban type (between 69% and 77% depending on the species). A representativeness bias may lead to a reduction in PM and NO_2 peaks at station points when the regional CTM simulates average concentrations over 20km grids. Nevertheless, these scores show an overall agreement with those presented in the latest articles using the 2020 version of CHIMERE (e.g. Menut et al., 2021; Guion et al., 2023), as well as with other regional CTMs (e.g. Bessagnet et al., 2016).*”

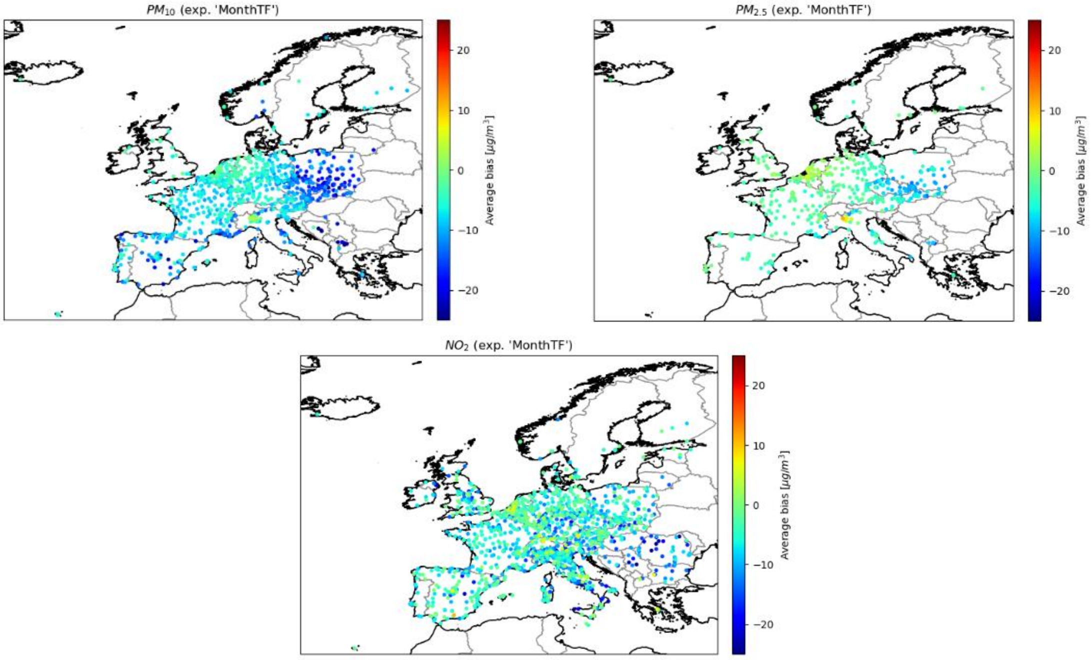

Figure A - Average bias (model - observations) [µg/m3] for 2018 for PM10 (top left), PM2.5 (top right) and NO2 (bottom).

4. [Section 4.1.2] It is not clear to me where the values for the thresholds (25 µg/m^3 for $PM_{2.5}$, 50 µg/m^3 for PM_{10} and 40 µg/m^3 for NO_2) come from (line 394). Maybe I missed it, but I can't understand it at this point in the text.

These values are based on the Directive 2008/50/EC of the European Parliament and of the Council on ambient air quality and cleaner air for Europe (https://eur-lex.europa.eu/eli/dir/2008/50/oj). They represent concentration limit values at 24-hour averaging period. This will be specified in the manuscript (lines 405-407) and the caption of Table 4.

Lines 405-407: “*The number of exceedance of daily concentration thresholds of 25µg/m^3 for PM2.5, 50µg/m^3 for PM10 and 40µg/m^3 for NO_2 (based on the Directive 2008/50/EC of the European Parliament and of the Council on ambient air quality and cleaner air for Europe) have been calculated at each monitoring station over Europe for JFM months of 2018.*”

Technical suggestions:

1. line 41: … need to **be** distributed …

2. line 67: This parameter should mainly **depend** on …
3. line 78: … are based **on** …
4. line 86: delete the space before the colon
5. line 87: “by comparing” or “in comparison”
6. line 104: … assuming that there **are** no …
7. line 129: … and their proportion**s** …
8. line 130: … come from very different energy type**s** …
9. line 154: …, transmission systems, …
10. line 183: … ambi**e**nt temperature.
11. line 239: … PM are **split** into …
12. line 256: … calculates …
13. line 260: … from the operational …
14. line 287: … projections using the HDD method **are** …
15. line 326: Using a**n** HDD …
16. line 341: … the HDD parameterization **does** not appear …
17. line 415: … no reason **that** …
18. line 430: As expected **a** higher deviation **is** obtained …
19. line 435: delete “can”

Thank you for reading carefully. Spelling and typographical errors corresponding to comments 1 to 19 have been corrected. The changes can be followed in the track change file.

20. Figure 8, S4, S5, S6, S7 and S8: the unit on the y-axes in the figures is µg m^{-2}, but µg m^{-3} in the text

As the reviewer rightly pointed out, there was a typographical error on the units in Figures 8, S4, S6, S7, S8 and S9, which has been now corrected.

Referee #2

The manuscript by Guion et al. presents a method to temporally redistribute historical estimates of emissions of air pollutants from the residential sector tied to space heating based on the day-to-day variability in the weather. The contribution of biomass burning is a particularly important source of particulate matter (PM) from residential heating and is well known to significantly contribute to PM in certain urban areas during cold weather, so it could be expected that accounting for the shorter-term variability in weather would improve air quality simulations.

The authors derive a set of country-specific factors for the EU-27 group of countries to account for variations across the region necessary to split apart the general emissions category that includes space heating available for the EU-27 and assign temporal variability. With these factors, the authors proceed to temporally re-allocate annual emissions from space heating to account for the observed variability in weather. The authors find some improvement in the model simulation of PM2.5 in the winter months for 2018, particularly during several cold spells with below average temperatures across large parts of the EU. Effects of the parameterization on NO2 and PM10 were generally more modest, but did show some improvements in certain countries. The authors also investigate the use of their parameterization to improve hindcasts of air quality while using emission inventories from two or three years before the simulation year. The idea here is to investigate whether the method could be used to improve real-time air quality forecasts, which use the most recent available inventory that is generally two or three years old.

The article is well written and clearly presents the effects of the day-specific emissions allocation as compared to the monthly-average allocation. Given that models always include a variety of errors that vary spatially and temporally, I am not surprised that the sensible allocation of day-specific emissions improves the simulation in certain regions while producing little improvement in other areas. Figure 6 nicely shows this, comparing the model simulated PM2.5 against observations and finding that the parameterization produces fairly widespread improvements in the correlation and root mean square error but not for all stations.

Comments :

1. The only significant comment I have on the manuscript is on the presentation of the multi-year emissions projections in Section 4.2. I understand that the motivation for this work is to explore whether the day-specific allocation could improve real-time air quality forecasts, but because the parameterization requires the meteorology for the full year to be known it is not clear how the parameterization in its current form could be used for real-time air quality forecasting. The analysis of the effects of the day-specific emission allocation is also confounded with the effects of revisions of the emission datasets from one year to the next, that includes a representation of the real change in emissions with time as well as revisions in the methodology to estimate emissions. The authors do demonstrate some improvement in the simulation applying the weather parameterization to previous year emissions, but it is difficult to clearly see how the projected emissions estimates are necessarily better than persistence and how any improvement is related to the use of the day-specific emission allocation. I am not suggesting the authors remove this part, because I am sure the application is of interest for air quality forecasting, but I would suggest the authors consider clearer ways to demonstrate the day-specific emissions allocations have improved the emissions estimates and to demonstrate how.

Thank you to the reviewer for this comment. Section 4.2 “Multi-year emission projections” could indeed be improved with a few more explanations.

As introduced at the beginning of section 4.2 (lines 422-431), the main risk of using HDDs to project emissions from a previous year (taken as the base year for which HDDs are summed) to the current year (in forecast setup) is to deviate from the official total emissions that will be reported (once the year has passed). This section has therefore focused on this issue, using the 2009-2018 period to replay projections. However, no air quality simulation has been carried out with CHIMERE using projected emissions as part of this work.

> Lines 422-431: “*Based on temperature, HDDs can be used to project emissions from the GNFR_C sector for a given year as an annual total (see Eq. 6), but also in near-real time on a daily time step (by normalising the HDDs of the current day with the daily average HDDs of the base year). The use of HDDs to simulate the day-to-day variability of residential emissions does not carry a specific risk in the context of simulations for a past year (reanalyses), as the annual heatsum of the corresponding year is known and can be used to normalised total emissions. On the contrary, when used in a forecast setup, using HDDs can induce a deviation from the input emissions, as the heatsum of the running year is unknown. This deviation is legitimate as in a colder (resp. milder) than expected winter, emissions should rightfully be larger (resp. lower) than originally prescribed. There is no reason that using emission from a past year (such an approach, referred to as persistence, is routinely used for the operational forecast) would be more legitimate. It is however important to document that risk of deviation. This second part is therefore devoted to the use of HDDs to estimate the total annual emissions of the GNFR_C sector.*”

The aim was to assess the deviation from the official reporting using the HDD method compared with the persistence and an estimated uncertainty. As the emission inventories (CAMS-REG-AP, CEIP) provide annual totals, the emissions were summed over the year to enable deviations to be calculated. However, the parameterisation does not necessarily require meteorology for the whole year. The HDD method can be used to project emissions from the residential sector in near-real time on a daily time step. The HDDs of the current day can be normalised by the daily average HDDs of the base year, providing therefore a temporal factor ($HDD(d, n_{ref+x})$, / $\overline{HDD}(n_{ref})$). Figure B illustrates this with the projection of a daily temporal factor for 2017 based on 2016, calculated under conditions of near-real time conditions of temperature availability over the 90 first days (full curve) and over the full year (dashed curve).

Using France as an example (2016-2018), Figure C illustrates the benefits of using HDDs to project annual emissions from the residential sector, especially compared to the persistence method. Compared with 2016, 2017 and 2018 were warmer (+0.3°C and 0.8°C respectively on annual average) and the number of HDDs therefore decreased. Based on this method, a decrease in emissions has been projected, in line with what has been officially reported. A deviation of only 1.49% from the official report has been calculated, which is still less than the estimated uncertainty of 3.0% (based on recalculations of emissions up to two years later). In the absence of inter-annual variability, the persistence method misses this decrease, and its deviation from the official report reaches 11.61%.

Considering this comment, clarifications have been made in the manuscript (lines 422-431, 451-458) and Figure C (Figure 10 in the manuscript) has been added.

> Lines 451-458: “*The average relative deviations induced by the "DayTF_Tbfit_fspec." projection and the persistence method are compared to the reporting uncertainty estimated from the variability of emission reporting (see Sect. 3.3.2) for each EU-27 country. An example is given*

with France in Figure 10 over the 2016-2018 period. Compared with 2016, 2017 and 2018 were warmer (+0.3°C and 0.8°C respectively on annual average) and the number of HDDs therefore decreased. Based on this method, a decrease in PM2.5 emissions has been projected, in line with what has been officially reported. A deviation of only 1.49% from the official reporting has been calculated, which is still less than the estimated uncertainty of 3.0% (based on recalculations of emissions up to two years later). In the absence of inter-annual variability, the persistence method misses this decrease, and its deviation from the official report reaches 11.61%."

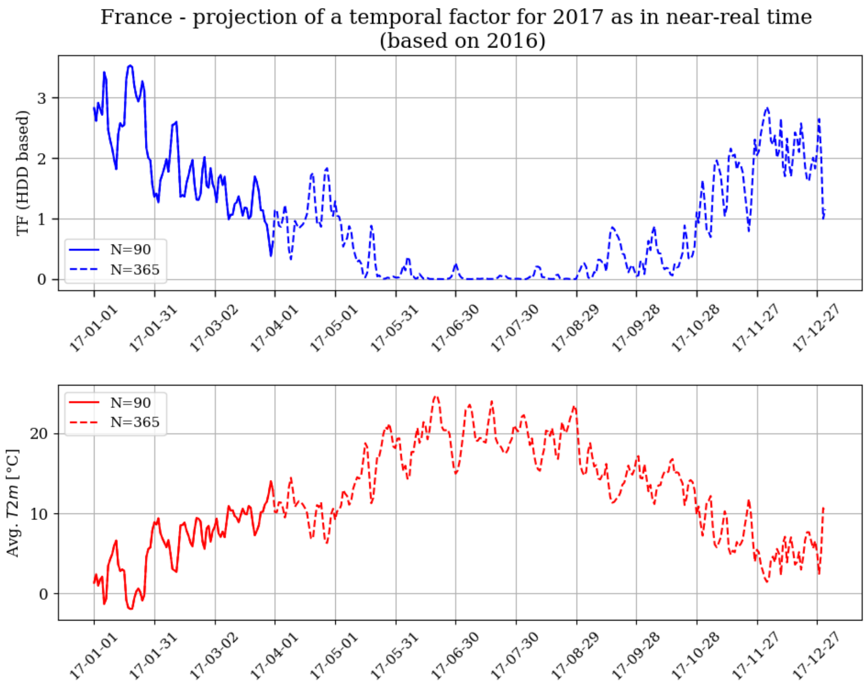

Figure B – Projection of the daily temporal factor TF (blue curve) for the year 2017 (based on the HDD sum of 2016) calculated under conditions of near-real time conditions of temperature availability over the 90 first days (full curve) and over the full year (dashed curve). Daily average temperature at 2m (red curve).

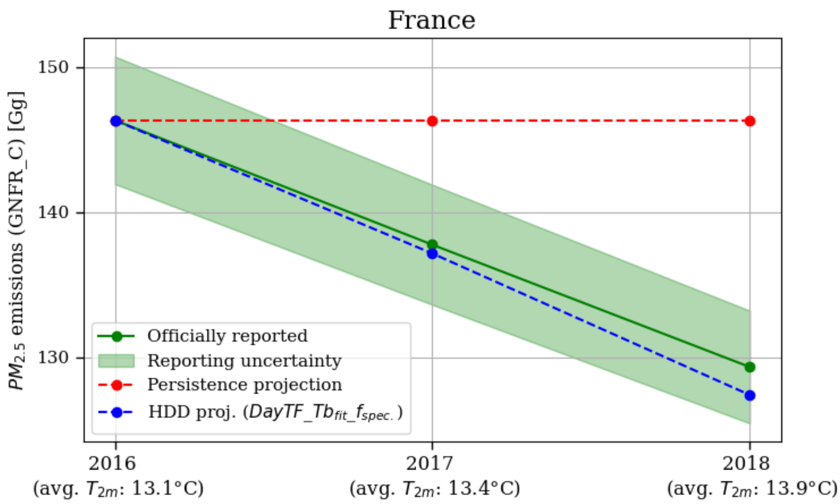

Figure C – Example of a projection of total national emissions (GNFR_C sector) for France up to years *n+2* (2017 and 2018) from year *n* 2016. The blue curve represents the projection using the HDD method ("DayTF_TBfit_fspec."), the red curve the persistence projection, the green curve the official reported emissions (CEIP) and the green shaded area the reporting uncertainty estimated up to *n+2* years.

2. Figure 5 presents the difference in Spearman correlation coefficient and RMSE for the different experiments broken down for four different countries. Do the authors have any explanation for why the full parameterization (DayTF_Tbfit_fspec, shown in red) seems not to perform noticeably better, and in some countries worse, than the partial parameterization DayTF_Tbfit? And while there are differences in results between the parameterizations, the differences are much smaller than the difference between countries. Italy has a much lower correlation coefficient than the other countries and France has a RMSE at least one half the size of the RMSE for the other three countries. For the RMSE in particular, do the authors have any explanation for the country-to-country differences? From Figure 8 it does not appear that PM2.5 concentrations are much lower in France than the rest of Europe, which may be expected to result in a smaller absolute RMSE. Does the small RMSE for France suggest the emissions inventory for France is better than for other countries?

Indeed, Figure 5 shows that for some countries, mainly Belgium and the Netherlands, the RMSE can be higher with the "DayTF_Tbfit_fspec." parameterisation than with "DayTF_Tbfit". This can be explained by the "f_spec." value (relative contribution of space heating to total residential combustion emissions) which is lower in "DayTF_Tbfit_fspec" based on Eurostat dataset (0.01 for Belgium and Netherlands compared to 0.20 as reference value in the other experiments). As the reference simulation already shows a positive bias (i.e. an overestimation of the model compared to observations) for these two countries (Figure S5), the inclusion of this parameter leads to an increase of the simulated concentrations during cold periods. As a result, the variability may increase (with the coefficient R) for some countries, but also the RMSE. Finally, the quality of this experience inevitably depends on the accuracy of the Eurostat dataset. These elements of discussion have been added to the manuscript (lines 344-347).

> Lines 344-347: "*Indeed, the f_PM parameter is lower in "DayTF_Tbfit_fspec." (0.01 for Belgium and Netherlands) than f_ref. (0.20). As the simulations show a positive bias for these two countries (already the case with "MonthTF", see Fig. S5), the RMSE scores are higher for "DayTF_Tbfit_fspec." than for "DayTF_Tbfit". The validation scores obtained also depend on the accuracy of the values provided by the Eurostat dataset.*"

However, it should be noted that Figure 5 focuses on daily average scores over the whole season and only for certain countries (those for which the Tb parameter has been derived from national gas data). The added value of the "DayTF_Tbfit_fspec" experiment is also shown in Table 4, which considers all the stations in Europe and focuses on days of exceedance concentration values (therefore on peak concentrations). In addition to having the most complete parameterisation, "DayTF_Tbfit_fspec" presents the best scores for good detection and missed alarms among the experiments for NO_2, PM10 and PM2.5 (although the number of false alarms may increase for NO_2).

Furthermore, it is true that performance scores can vary considerably between countries (and even more than between experiments which differ only in the parameters of the temporal distribution of heating emissions). For example, the RMSE is significantly lower in France for PM2.5 and in Estonia for NO_2. The various CHIMERE simulations presented here all use the 'CAMS-REG-AP' emissions inventory over all the European domain (with no specific treatment for France or Estonia). The 'CAMS-REG-AP' inventory spatialises the national totals officially reported by the various EMEP member countries. However, there may be significant differences, even between countries with ‘similar characteristics”, depending on the calculation method used. These national differences in the calculation of anthropogenic emissions impact the concentration fields simulated by the CTMS. For example, the share of residential heating emissions in total emissions is not the same in all countries. Finally, the differences in RMSE between countries are not only due to emissions but also to the representation of other processes such as long-distance transport. These elements of discussion have been added to the manuscript (lines 353-356).

> Lines 353-356: “*Performance scores can vary considerably from one country to another (and even more so than between experiments which, as a reminder, differ only in the parameters of the temporal distribution of heating emissions). This is discussed in more detail in the discussion (see Sect. 5), but national differences in the calculation of emissions and the representation of certain processes such as long-distance transport have an impact on the simulation scores for each country.*”

Minor comments:

1. Line 78: The phrase ‘regional air quality forecasting’ seems to be missing the word ‘system’

The sentence has been corrected (lines 77-79).

> Line 77-79: “*Within the Copernicus Atmosphere Monitoring Service (CAMS, Peuch et al. (2022)) regional air quality forecasting system, forecasts are based by running several CTMs with emission inventories from a past year.*”

2. Lines 102 – 103: What measure of the outdoor temperature is used in the calculation of HDD? Is it the HDD calculated from hourly T2m or calculated from daily average temperature? If it is the daily average, how is the daily average calculated?

The HDD is calculated from the daily average temperature, using hourly T2m data from ERA5 reanalyses. This point is now clarified at the beginning of section 2 “Heating degree day (HDD) method to distribute emissions” (lines 101-104, 171-172).

> Lines 101-104: ”*The HDDs of the year n and the day d are computed for each grid cell of latitude i and longitude j by calculating the temperature difference between the daily average of the outdoor temperature at 2 meters T2m and the ambient temperature Tb above which a building is no longer heated (fixed threshold) (Eq. 1).*”
>
> Lines 171-172: ”*TF_HDD(c) is calculated from the daily average temperature, using hourly T2m data from the ERA5 reanalyses (Muñoz-Sabater et al., 2021).*”

3. Lines 171 – 172: Here it is stated that the surface temperature used in the calculation of the HDD threshold is taken from ERA5. Is this the daily average temperature or some other quantity?

This response supplements that of comment 2 (from referee 2 - Minor comments). This is the average daily temperature calculated from the hourly T2m. This is now specified (lines 101-104, 171-172).

4. Lines 323 – 325: The authors state that the comparison of the different experiments against observations will be made 'for each country for which the Tb parameter has been calculated based on the national gas data.' But Figure 5 shows a comparison for only four countries, while in Section 2.2.2 it is shown that it was possible to calculate country-specific Tb_fit for eight countries. Why the difference?

The comparison with the observations in Figure 5 is only made for 4 countries as there are no PM measurements available in 2018 in Hungary, Romania, Estonia and Latvia (unlike NO2 measurements, Figure 7). It was possible to calculate the Tb_fit parameter for these countries because national gas data were available. We have now specified that the comparison with observations is made in countries for which the Tb parameter has been calculated, only if measurements are available (lines 326-329).

> Lines 326-329: "*The effect of the different HDD parameterizations integrated in the emissions used in CHIMERE is analysed by comparing the Spearman correlation and RMSE scores in concentration with the available observations (stations averaged by country) for each experiment and more specifically for each country for which the Tb parameter has been calculated based on the national gas data.*"

5. Line 435: Typo in 'it can can provide'

The sentence has been corrected (line 449).