Response to Referee#2 for manuscript : Preprint egusphere-2025-708

Referees' comments are in **black.**

Authors'answers are in **blue**, *text from the original manuscript in black, italic* and *modified or added portions in blue italic.*

2.1 - This manuscript investigates from different perspectives how the current usage of near-surface temperature observations, usually considered measured at 2-metre, impacts model evaluation and data assimilation in mountainous regions in an operational NWP model. The authors consider three different sources of uncertainties: 1) altitude difference between model grid-point and station height; 2) difference in the height above the surface of the sensor and the one of the model, usually assumed constant at 2-metre above the ground; 3) inhomogeneities in the observation's distributions.

This is an important topic as usually point 2) and point 3) are not considered in-depth during both model evaluation and assimilation of surface observations. A better usage of these observations can potentially improve forecasting of near-surface variables and give more meaningful indications on model developments pathways.

I find some of the conclusions worth publishing as they can be a good contribution to the present literature. In particular, the impact of the difference in the height above ground between the sensor and the model in case of snow cover is of particular importance, as this is usually neglected in operational data assimilation systems.

We thank the reviewer for this assessment.

However, there are major concerns (see "General comments" below) that should be addressed before publication.

General comments

2.2 The manuscript should be deeply revised before resubmission. There are typos, missing parentheses and many sentences that are very difficult to read in English, which I strongly recommend rephrasing. A few examples are given in the specific comments below, but I encourage the authors to edit the text throughout. Also, in some places the tone of writing is too "colloquial", for instance see lines 464-466. While posing questions can be a good way to engage the reader, an excessive use does not align with the scientific style of a research manuscript.

We thank the reviewer for pinpointing some of the language errors, edits and typos, and we are thoroughly revising the manuscript in view of the submission of a revised version.

2.3 From a scientific perspective, I also strongly suggest revisiting the way the scientific questions are introduced and some of the results presented. In the introduction, it is hard to clearly get the scientific questions the authors would like to answer and what is the methodology used to address them. One way could be to clearly states what hypotheses and questions relate to model evaluation, and which ones to data assimilation. The last paragraph (lines 107 – 111) has elements of it, but it should be improved. For instance, the hypothesis that "In the present paper we examine the assimilation of mountain near-surface temperatures as a possible cause for the cold bias observed in Arome forecasts. ", in Sect. 2.3, should be introduced earlier in the text, possibly in the Introduction. Sect. 3.2 should be better introduced to guide the reader that the manuscript is now moving to the "assimilation" part and analysis of the sensitivity experiments.

We thoroughly revised the structure of the paper and especially the research questions, which are now much better framed in the Introduction, and formulated so that they provide the overarching structure of the Results sections. We invite the Referee to read the **answer to point 1.4 from Referee 1** for a complete answer to this comment.

2.4 Regarding the logical order, some of the "Discussion" sections seems more "Results" and should be rearranged accordingly. For instance, Sect. 4.2 seems better placed together with Sect.3.1.1. Figure 8 is discussed in Sect. 3.1.3 so it should be introduced accordingly in the text.

The Results and Discussion sections have been reorganised for greater clarity. We especially followed your advice to put section 4.2 into the Results section 3.1; it now constitutes a dedicated subsection 3.1.4. Similarly, the former Figure 8 has been relocated in the Result section 3.1.3 (see point 1.4 from Referee 1).

2.5 Some parts of the Discussion and Conclusions are too tightly related to AROME. This could be ok, as this is the numerical tool the authors are using, however I strongly encourage the authors to draw some generality out of some of these statements for the wider community, for instance talking about the physical reasons to use or replace a particular scheme rather than referring to a specific scheme's name.

This has been improved and we invite the referee to read the new version of the Discussion in the revised manuscript, as it would be too long to reproduce it here.

The Conclusion was entirely rewritten, accounting for this comment and also as demanded by Referee#1. It now follows the overarching research questions explicit in the Introduction, and is reproduced here:

"This study investigated the impact of inhomogeneities of the observational network specific to mountain regions, on the evaluation of the NWP system Arome and on the effects of surface data assimilation within this system.

We first questioned whether the differences in height above the surface between sensors should be cared for when evaluating models in terms of near-surface air temperature. These differences are correlated with altitude and induced by the need to prevent the sensors from being buried in thick snowpacks in high-altitude terrain over the winter. We showed that T5m and T2m should not be considered equivalent when performing model evaluations: despite a limited mean difference over winter at our mid- and high-altitude research sites, both temperatures can differ significantly in specific situations, especially low-winds and clear skies. Therefore, taking one for another introduces errors. Furthermore, at the instance of the Arome model, atmospheric models may present very different biases at these different heights, so that the confusion between both temperatures leads to erroneous interpretations of model biases. We therefore recommend a distinct evaluation of modeled T5m and T2m against the relevant observations in mountain terrain. Only such kind of altitudinally differenciated evaluation can foster a better understanding of the model limitations and promote efficient model improvements over mountain regions.

We then questioned whether this difference in height plays a detrimental role in assimilation, as observations at 2 or 5 m are not discriminated within the assimilation system of Arome, an approximation that we estimate may be common among NWP systems. We showed that indeed, this confusion between heights in the assimilation process, leads in the case of Arome to an overestimation of the analysis increment in high-altitude regions, inducing an overestimation of T5m analysis at night and a degradation of performances with respect to the model without assimilation (background or forecast).

Finally, we questioned the effect of station vs model relief mismatch, and higher density in valley stations, onto the assimilation. The differences in altitude between stations and model grid-points, does not affect significantly the performance of assimilation. This may be due to the limited number of stations with an important (higher than 150 m) relief mismatch with respect to the Arome model, that runs with a high spatial resolution coming with a better representation of the topography that models with coarser spatial resolutions. With respect the imbalance between observation stations across altitude, we find that it also weekly affects the assimilation: as a matter of fact, the effect of low-altitude stations at high-altitude locations, is of the same order of magnitude that the effect of high-altitude stations assimilation onto the temperature of low-altitude areas. However, this effect is quite strong, changing the analysis temperature by about +/- 0.3°C. This means that data from a different altitude, bring a noticeable correction to the model at another altitude, where the biases can be different. This result illustrates the limitations of the current 3DVar assimilation system disregarding the effect of topography in the spatial structure of assimilation increments. Our analysis further confirms the strong analysis increment, at high-altitudes, induced by the assimilation of Nivose stations as if their measurements were at 2 m above the surface. We showed that this effect is probably the reason why the assimilation of surface observations degrades the performances of Arome in this altitude range, while relying on upper-air data (satellite, radar..) assimilation only would produce a better analysis.

To summarize, this study helped define guidelines for the improvement of high-resolution NWP systems in mountain terrains: In particular, sensors'height should be considered both in model evaluation and assimilation; topography should be accounted for in the spatial structure functions involved in assimilation; model biases at 2 m height and lower could possibly be reduced by the use of diagnostics more appropriate to mountain terrain, a higher number of vertical levels in the models and enhanced work on the surface scheme to improve the representation of soil-snow-atmosphere energy transfers."

2.6 From the point of view of the methodology, I have a few issues with the sensitivity experiments performed and the analysis presented in Sect. 3.2 and Figure 7. Firstly, it is hard to understand what is plotted in Figure 7: it should be better clarified when the difference between the sensitivity experiments and the control are plotted; I do not understand the reason to use a figure legend that is different from the sensitivity experiment names, which adds confusion to the reader. Secondly, it is not clear to me what the "mountain" line represents in Figure 7: the NO_NIGHT experiment already removes the assimilation of all surface observations during nighttime (at least this is the understanding from Sect. 2.3), so what is the reason to combine it with the NO_VALLEY experiment increments? Furthermore, if the aim of the authors was to check the impact of all mountain surface observations, an additional sensitivity experiment, in which all T2m observations were removed, would have been useful.

The linearity assumption used to combine the increments can be hard to justify, given the high non-linearities present in NWP models. At least the authors should justify why this experiment was not performed.

To clarify what is plotted on Figure 7, we added a new section in the Material and Methods section, entitled "2.4.2 Analysis of the experiment". This section defines the analysis increments that we use to analyze the results of the experiments, and describes precisely how we combine the increments of the different experiments to estimate the effects of valley, mountains, altitude.. observations. We also distinguish between *analysis increments* (Δ) and *virtual analysis increments* (Δ^v), the latter being diagnosed from complementary, denial

experiments and therefore incorporating compound effects which we cannot disentangle.
For instance, for nighttime, the NO_NIGHT experiment that suppresses the assimilation of surface temperature and humidity observations over night, enables to retrieve the contribution of upper-air (altitude) observations only. For the nighttime period, we define the virtual analysis increment from surface observations only, Δ^vobs_surface, as:

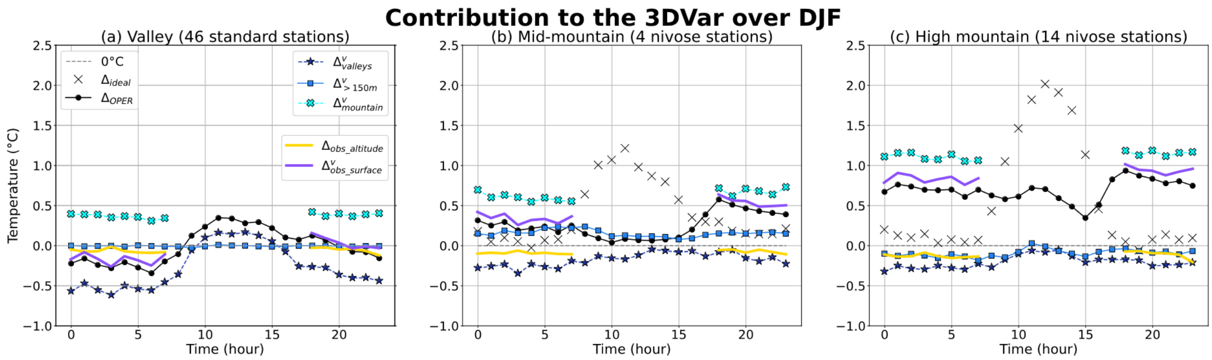
$$\Delta OPER = \Delta^v obs\_surface + \Delta NO\_NIGHT$$

This virtual analysis increment for surface observations only, likely differs from the one that would have been calculated by disabling the altitude analysis, due to compound effects between altitude and surface observations. In the decomposition we propose, these compound effects are integrated in the surface observation *virtual* analysis increment, hence distinguished as *virtual*. We now clearly acknowledge that the experiments performed do not enable to quantify the compounds effects.

The increments defined in this new section 2.4.2, are incorporated in the legend of the Revised Paper Figure 7 for consistency and enhanced clarity.

The fact that the NO_NIGHT enables the analysis of the surface observation contributions *for the nighttime only*, is now stated more clearly.

Please find below the **Revised Paper Figure 7 (now Figure 9)** and the **new section 2.4.2 "Analysis of the experiments"**:



Revised Paper Figure 7 (now Figure 9): *Analysis increments (denoted Δ) obtained in different configurations of the pool of assimilated observations, as described in subsection 2.4.2. These increments are retrieved at stations'locations in valleys (a), mid-altitude mountains (b) and high-altitude, taking into account only Nivose stations in mountains (b,c). the difference between the observation and the background of Arome-OPER represents an idealised increment (black crosses). There is no measure at 5 m in valleys, so no idealised increment is calculated*

New section 2.4.2 "Analysis of the experiments" (see next page)**:**

2.4.2 Analysis of the experiments

In the Results section 3.3, the above-mentioned assimilation experiments will be analyzed to quantify the effect of varying observational network characteristics onto the assimilation result (i.e. the analysis). These characteristics include the exclusion of valley and flatland stations, of all surface stations at night, and of stations for which the altitude difference with respect to the model grid-cell exceeds 150 m. To highlight the effect of these variations in observational networks, we make use of the analysis increment Δ, whereby :

$$\Delta = x_a - x_b \tag{3}$$

with x_a the analyzed model state and x_b the background model state prior to assimilation.

At observation stations, an ideal analysis increment would enable the analysis to fully coïncide with the observation. We therefore define the ideal analysis increment at stations as:

$$\Delta_{ideal} = x_{obs} - x_b \tag{4}$$

where x_{obs} denotes the observation.

The NO_NIGHT experiment, disabling the assimilation of surface observations at night, enables to highlight the effect of the altitude observations only for the nighttime period. We hence call for the nighttime period :

$$\Delta_{obs\_altitude} = \Delta_{NO\_NIGHT} \tag{5}$$

For the nighttime period, we hence can define a virtual analysis increment coming from the analysis of surface observation only, $\Delta^v_{obs\_surface}$, by considering the following relationship between the analysis increment of the Arome-OPER experiment (Δ_{OPER}), and the ones that respectively result from the assimilation of altitude ($\Delta_{obs\_altitude}$) and surface observations ($\Delta^v_{obs\_surface}$) only:

$$\Delta_{OPER} = \Delta^v_{obs\_surface} + \Delta_{obs\_altitude} \tag{6}$$

In practice, this virtual analysis increment for surface observations only, likely differs from the one that would have been calculated by disabling the altitude analysis, due to compound effects between altitude and surface observations. In the decomposition proposed in relation (6), these compound effects are integrated in the surface observation analysis increment $\Delta^v_{obs\_surface}$, hence distinguished as a *virtual* increment analysis, and we do not have the possibility to quantify them.

Similarly, the analysis increment of Arome-OPER can also be decomposed on into the virtual contribution from the flatland and valleys $\Delta^v_{valleys}$, and what comes from the upper-air and mountain stations only included in the NO_VALLEY experiment. According to this decomposition:

$$\Delta_{OPER} = \Delta^v_{valleys} + \Delta_{NO\_VALLEY} \tag{7}$$

and also:

$$\Delta_{OPER} = \Delta^{v}_{valleys} + \Delta^{v}_{mountain} + \Delta_{obs\_altitude} \quad (8)$$

where relation (7) enables to retrieve $\Delta^{v}_{valleys}$ while relation (8) enables to retrieve the contribution from mountain stations only among surface observations, $\Delta^{v}_{mountain}$.

Another possible decomposition of Δ_{OPER} reads:

$$\Delta_{OPER} = \Delta_{150M} + \Delta^{v}_{>150m} = \Delta_{obs\_altitude} + \Delta^{v}_{<150m} + \Delta^{v}_{>150m} \quad (9)$$

where $\Delta^{v}_{>150m}$ (resp. $\Delta^{v}_{<150m}$) is the virtual analysis increment for surface stations with more (resp. less) than 150 m altitude departure with respect to model relief, while Δ_{150M} refers to the 150M experiment.

In these latter relations, similarly to the $\Delta^{v}_{obs\_surface}$ increment, the virtual increments, denoted by a v exponent, are not directly calculated from an experiment but diagnosed from a complementary experiment, and therefore include compounds effects that cannot be isolated.

These different increments will be used in the Results and Discussion sections to analyze the effects of heterogeneities in the observational network in Alpine terrain, on the assimilation in Arome.

2.7 Another issue is related to the analysis in Sect. 3.1.1, in particular the author's statement that the approximation T2m ~ T5m is incorrect (line 306). From their results this conclusion is a bit misleading as on average, considering the whole winter periods, they show that this approximation is valid. The approximation is valid in mean, meaning that the mean difference between T5m and T2m is within the (quite high) observational error assumed at the CLB site and at the CDP for the T5m only. But the approximation is certainly not valid when regarding diurnal amplitudes, diurnal cycles or **error scores** like the RMSE. We now say this more clearly in the manuscript.

2.8 The authors base their conclusion on the analysis of a specific case study in Figure 5, covering only a few days. To make their conclusion statistically stronger, I think the authors should also show a diurnal cycle of temperatures (or some statistics) computed only for anticyclonic conditions or clear- sky periods.

We will incorporate a statement and statistics on this aspect in the revised version of the manuscript, also referring to Gouttevin et al., 2023 who analyzed the diurnal cycles in temperatures at different heights above the surface under clear-sky vs cloudy conditions at CLB in winter. We here present an analysis performed over one winter period (DJF) at CLB, distinguishing between clear skies and cloudy skies based on an atmospheric effective emissivity criterion similar to Gouttevin et al 2023 (**Figure R3**): the 25% days with strongest (resp. lowest) effective emissivities are classified cloudy (resp. clear-sky). Additionally, we distinguish the low-wind periods within the clear-sky days, by selecting the moments when wind speed is lower than 4 m/s. Figure R3 shows that the daily cycles in temperature differences between 2m and 5m at the CLB, largely differ between cloudy conditions and clear-sky conditions. In the former there is almost no difference between t5m_obs and T2m_obs, while in the latter the mean difference is strong especially at night, reaching -0.6°C between 0 and 8 UTC. This difference is larger than the measurement uncertainty. It is even stronger in clear-sky, low-wind conditions when the mean DJF difference at night is below -0.7°C for several hours.

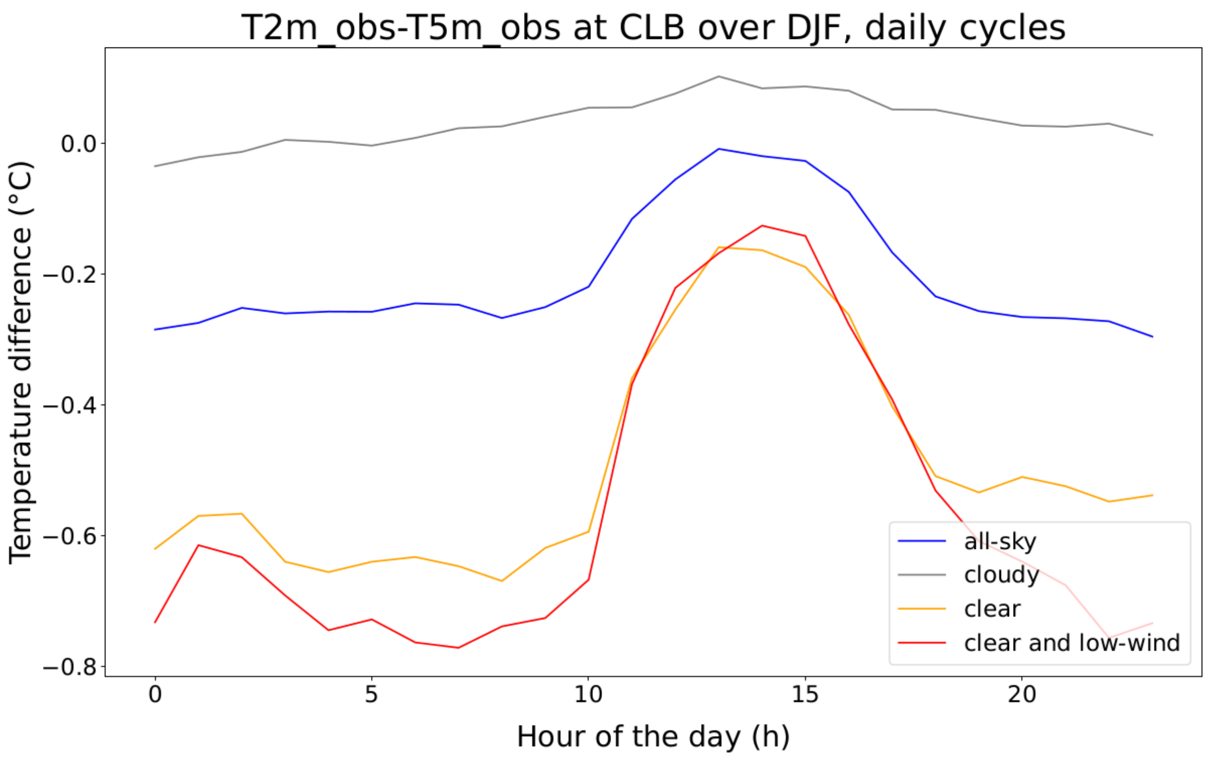


Figure R3: Daily cycles of the T2m_obs minus T5m_obs difference at CLB over one winter period (DJF), distinguishing the **clear-sky** and **cloudy** weather conditions. '**all-sky'** represents the mean DJF daily cycle (all-sky conditions). The low-wind periods within the clear-sky days are also distinguished (**clear and low-wind**).

Specific comments

Abstract: "rôle" à "role". This has been corrected.

Line 28: "… they are originally designed for Rudisill et al. (2024); Gouttevin et al. This has been corrected.

(2023)." Some parentheses might be missing for the references. All references have been corrected

Line 35: "While primarily strong over peaks and ridges, it often comes with a warm

bias in valleys. " is not clear. Do you mean "…it is often associated with a warm bias in valleys"? Yes, and this is now corrected.

Line 38: I think a few more references on the literature review and/or the forecaster's reports would be useful to better justify the 3 types of biases described. References to Arnould and Preaux, 2021; and Beauvais, 2018, have been added.

Arnould, G. and Préaux, D.: Study of AROME Temperature in mountain regions, ACCORD Newsletter, 2021.

Beauvais, L.: Fronts chauds sur les Alpes : Hiver 2017-2018. Comportement des modèles, ateliers de la prévision du Centre-Est, 2018.

Line 44: “wood fire heating Aymoz et al. (2007)” some parentheses are missing for the reference. This has been corrected.

Line 46-52: Not clear, could you please reformulate? Also It seems that the reference justifying the argument is put at the end (Beauvais, 2018) but should it be placed at the beginning of the paragraph? We reformulated the sentence and positioned the reference at the beginning:

“The second Arome warm bias manifests in valleys when a warm front encounters the relief, especially in the direction perpendicular to the valleys and ridges (Beauvais, 2018). In these situations, the warm front penetrates too rapidly or too deeply in the valleys, leading to a modelled rise in temperature that is too strong and often generates an altitudinal upward shift in the snow-rain transition in the model. As a result, the model can forecast rainfall instead of snowfall in the valleys, where the major roads are.”

Line 75: what are Nivose stations? Please introduce them properly. This has been corrected:

“This is typically the case in France, where the sensors of the high-altitude observation network for snow and mountain meteorology, the so-called "Nivose" stations, are about 7 m above the snow-free ground.”

Line 83: “have been a preoccupation for numerous modelers”, not clear. Do you mean “have been an issue recognised by numerous modelers”? We rephrased following this suggestion.

Line 83: “kind of covering up the height-above- surface adjustments that we here mention. “, please reformulate this. This has been reformulated:

“*It may have until now obliterated the possible issue of height-above-surface adjustments*”

Line 99: “mountain” à “mountains” OK done

Line 108: “pitfall” does not sound very scientific, do you mean “challenges”? Yes we switched to “challenges”

Line 142: “T5m_ mod refers abusively to the temperature..”, not clear. What do you mean by “abusively”? Also I think there is a trailing space between T5m_ and “mod” We suppressed tha abusively as it was misleading :

“For convenience, in the diagram and in the rest of the article, T5m_mod will refer to the temperature at the first level of the model, which is approximately at 5~m above the surface.”

Line 154: “This stage eliminates observations that are considered doubtful because they come from a non-qualified source or are too far away from the design.” Is not clear, please reformulate. Do you mean that differences with the first guess are larger than a certain threshold?

Yes, there was an error in our sentence, and we corrected “design” for “background”.

Line 154: “unfairly” is not a scientific term here. We reformulated:

“However, if this background is biased, the screening can also reject observations that come from accurate measurements and contain valuable information for the assimilation”

Cost function Equation: Please add an equation number to the text. All the equations are now labelled.

Sect. 2.1.2: Why CANARI and the 2D-OI are introduced? It does not seem this is used at all in the manuscript, as the analysis focuses on 3D-Var, as far as I understood. Indeed the focus is on the 3DVar in terms of assimilation scheme. However, we also examine the temperature biases of AROME-oper at 2m above the surface. This temperature, T2m_mod, is the result of a diagnostic that considers both T5m_mod and Ts_mod, the latter being affected by the surface analysis CANARI. It therefore seemed necessary to introduce CANARI to present all the schemes affecting our temperatures of interest. We added this to the manuscript for more clarity.

“The analyzed surface temperature is involved in the estimation of the analyzed temperature at 2 m via a diagnostic (Figure 2c)”.

Line 161: “Increments are calculated for the surface observations and for the upper-air observations. Then, J is minimized using these increments.” I am not sure if here “increments” is the right terminology. Do you mean the x-xb terms in the cost function equation?

Indeed, *‘increment’* is not the right term: what is meant is *‘background departure*’. We have corrected this error in the manuscript.

“Background departures are calculated for the surface observations and for the upper-air observations. Then, J is minimized using these background departures.”

Line 168-172: This is not clear. How B is defined in the 2D OI equation? What do you mean “that very few observations are important in determining the analysis increment”? Please clarify and/or reformulate.

In the case of CANARI, B is a static, univariate matrix (meaning that it does not integrate any inter-variables correlations, e.g. no correlation between T2m and Hu2m). It relies on a very large correlation-length of 100 km, used almost isotropically as only modulated by sea/land surface mask and topography (Marimbordes et al., 2024). The large correlation length works so that only a few observations are enough to have an impact on the analyzed state.

OI was traditionally used for its numerical efficiency and flexibility, via the possibility to reduce the system to a “small” matrix inversion, considering only a limited number of relevant observations for the correction of the model state (see Durand et al., 1993 for an early time application in data-scarce mountain areas).

We added these elements in the revised version of the manuscript and rephrase the sentence on the few observations for more clarity : “*OI is an assimilation method particularly suited in the context of rather scarce data, when a limited number of observations are used to determine the analyzed state (Durand et al., 1993).*”

Durand, Y., Brun, E., Merindol, L., Guyomarc'h, G., Lesaffre, B., and Martin, E.: A meteorological estimation of relevant parameters for snow models, Annals of glaciology, 18, 65–71, 1993.

Line 177: "1D IO" I think should be "1D OI". This has been corrected.

Line 180: "(Figure 2: map on the right) ", please add a label to each panel and refer as "Figure 2b". This now done in the revised manuscript.

Line 204: "…, with for the latter a measurement co-located with every other observation due to the high spatial variability of the snow height at this site." It is not clear, please reformulate. We reformulated the paragraph.

Line 213: I think the authors refer to the Nivose stations in the text before these are described/introduced. Please adjust the text. This has been adjusted.

Line 221: This hypothesis would be better placed at the end of the introduction to better illustrate the scope of the work. I think this is somehow described at line 96-97, but it is not clear enough, in particular the link to the model cold bias. The authors should also explain why an assimilation deficiency should cause a forecast bias (at which lead times?). This is now accounted for through the research questions and hypotheses stated in the Introduction, that specifically mention this hypothesis.

The effect of assimilation on the forecast clearly depends on the variables/situations/phenomena at stake, and could last for one to a few hours or even more, but it is hard to precisely tell this in the present context without a dedicated evaluation. We changed the formulation as the question of effects on the forecasts was actually not addressed in our study:

" *In particular, we will examine the assimilation of mountain near-surface temperatures as a possible cause for the cold bias of Arome*."

Line 224: "These numerical simulations are be compared to a reference. ". Please reformulate. We corrected the typo: are to be compared.

Figure 3 caption: "exemple" à "example" This has been corrected.

Line 283: "et" à "and" This has been corrected.

Line 306-310: "Thus, the approximation is invalidated: …" please reformulate. It is not clear at this stage what approximation the authors are referring to, even though it is clarified in the remaining of the paragraph. We now reformulate this clearly, by defining the commonly made "error in measurement height" at the beginning of Results section 3.1, and reformulating the sentence:

"We conclude from this section, that considering T2m and T5m as fully equal temperatures is an invalid approximation: the difference between T2m and T5m is weak and within the measurement uncertainty on average over winter, but is not so during certain weather situations".

Line 306-310: Can this argument be better generalised for instance by computing a diurnal cycle only for the anticyclonic conditions in the considered 4-year period? We will incorporate a statement on this aspect in the revised version of the manuscript, see for more detail the response to point 2.8.

Line 343: The discussion in the text “jumps” from Figure 4/5 to Figure 8, which is a bit confusing. Please readjusts the figure order or the text. This is now corrected in line with a complete reformulation of the Discussion.

Section 3.1.3 is difficult to follow. Could you please reformulate the Section to better distil the main message the authors would like to convey? For instance, reducing where possible references to “see above text” or if necessary, better explaining where a reader should focus, e.g. a figure or a particular subsection. This has been taken into account. Section 3.1.3 is now section 3.2.1 in the revised manuscript, and we invite the Referee to read this new section in the revised manuscript that we will provide shortly upon agreement of the Editor.

Line 369: “Secondly, we note that the analysed T5m is worse at Nivose stations”. Worse than what? This is now better explained:

“Secondly, we note that at Nivose stations, the analysed T5m performs poorly at night, and especially worse than the forecast at 5 m (at mid and high altitudes) and even then the background at 5 m (at high altitudes) (Figure 8 (b,d)).”

Line 383: “esp. “. please correct this. This has been corrected.

Line 387: “Mountainous areas are complex to instrument and model.”, please clarify what “complex” means in this context. This sentence disappeared in the restructuration of the manuscript as redundant with the enhanced description of the research questions and their motivations in the Introduction.

Table 2: typos in “Biais”, no closing parenthesis in figure caption. This is now corrected.

Line 495-497: Please rephrase. This has been rephrased.

“It is therefore one of the least biased models according to the synthesis by Rudisill et al. (2024) and is close to the Canadian limited area model GEM-LAM evaluated by Vionnet et al. (2015), featuring a "0.5°C cold bias at high elevations” (Rudisill et al., 2024; Vionnet et al., 2015).”

Line 513 – 520: This discussion is too much related to Meteo-France’s models and using a lot of details that are not of interest to an external reader. The authors should give the information that would be of interest for the community, not the namelist settings used in operational AROME. This has been improved and we invite the referee to read the new version of the Discussion in the revised manuscript.

Line 588: “Correcting T2m biases would also enable the model’s physical parameterisations to be improved.”. Is it not the other way around, improvements in physical parameterisations reducing the T2m biases in the first guess and hence reducing the “activity” of the data assimilation system? What we meant by this sentence, is that a progress was needed to more accurately characterize t2m, as a way to identify and henceforth limit “true” biases and error compensation. We rephrased for more clarity.

“Having a more accurate T2m estimate, not affected by e.g. the error in measurement height, would enable a better knowledge of the true model biases, the formulation of relevant hypotheses for these biases and henceforth favor the improvement of the model's physical parameterisations. “

Line 596: Is there any reference to previous work that could back up this hypothesis of misrepresented katabatic winds? Gouttevin et al. (2023) highlighted the possible role of Arome wind biases in the T2m bias of the model, but didn’t specifically target thermal, especially katabatic winds. This is a quite recent hypothesis in our research group and has not been previously formulated in published literature.

Line 599-606: This long paragraph mixes perspectives and hypotheses for future work with conclusions from this work. I would suggest reorganising it so that the conclusions from this work are clearly divided from the perspectives. E.g. the recommendation to properly consider the height above the surface of the observations in assimilation and evaluation should come as a conclusion, whereas reducing temperature biases through improved model physics (using Isba-ES-DIFF etc.) should come as a perspective. We followed your suggestion and now better serapare the Conclusions (Section 5) that synthesize our main findings and recommendations on how to properly use mountain temperature measurements for assimilation and model evaluation, from the Perspectives that now belong to a dedicated subsection “4.1 Summary and general perspectives” opening the Discussion section. We refer the referee to the revised version of the manuscript for the Discussion, as reproducing this part here would be very long.