Response to the Referee#1 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.*

Summary:

1.1 - This study addresses an important topic related to discrepancies in measurement and modeled surface air temperature heights. The study provides evidence supporting that the height differences between surface air temperature measurements should be accounted for when evaluating model performances and assimilating data. These contributions will be valuable to publish and account for in future research and operational modeling; however, there are major revisions required prior to this paper being suitable for publication that are addressed below.

We thank the reviewers for careful reading of the manuscript and helpful suggestions.

Overarching comments/concerns:

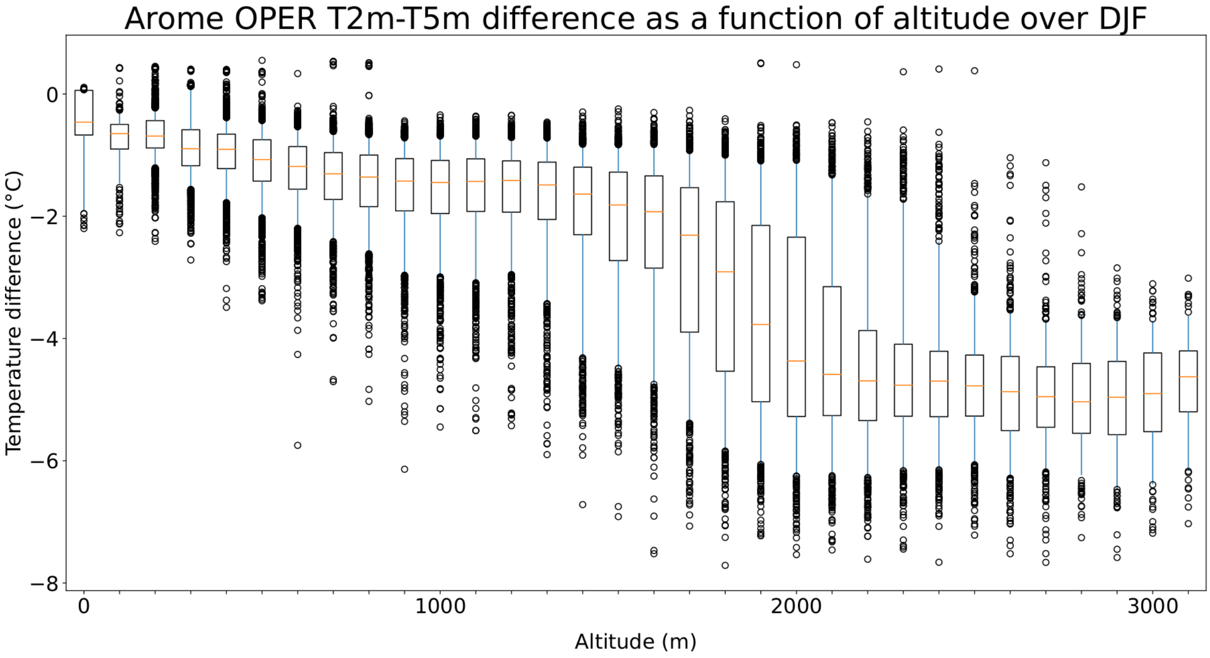
1.2 - The manuscript requires thorough editing by a native English speaker prior to its resubmission. There are common wording errors, citations outside of parenthesis, grammar issues, and awkwardly worded sentences that need to be resolved prior to publication. Examples are provided in the first 5 specific comments below, but this comment applies throughout the manuscript.

We carefully revised the wording, references and grammatics of the manuscript.

1.3 - Analyses for Section 3.1 are only conducted at 2 sites. This seems lacking and would require a justification of this limitation. Why are the other stations (e.g., from Figure 6) not included in this initial analysis? Even if both 2m and 5m temperature observations are only available in a few locations, it seems this analysis can and should be broadened by: (i) comparing modelled 2m and 5m temperature across a broader spatially continuous domain, and (ii) comparing modelled data with more ground observations, and group results by station height to evaluate the potential discrepancies between simulated T2 and T5, and provide deeper insights on how validating modelled T2 with observed T5, or DA practices, can induce issues. (i) Could be further used to evaluate how discrepancies between modelled T2 and T5 vary with geographic, climate, and vegetation conditions.

The Col de Porte and Col du Lac Blanc are to the best of our knowledge the only mountain stations with observations at both at 5m and 2m above the surface in France. This indeed poses an intrinsic limitation to the observation-based evaluation of model temperatures at these height levels, especially for a joint evaluation at these 2 levels. However we do fully agree that a broader evaluation of AROME 2m and 5m temperatures, with the stations from Fig6 and as performed in Fig 8, would strengthen our conclusions and broaden their impact in this section.

We therefore added **a new subsection within section 3.1, entitled “Assessment of forecasted T2m and T5m across the Alps”**. This section shows that the differences seen at the research sites between forecasted T5m and T2m in Arome, are actually quite representative of the entire Alps, where a negative difference between T2m-T5m is a generalized pattern and grows with altitudes across the mountain range (**new Figure 6 a,** see below). We furthermore now show in this section the evaluation of Arome against all T2m and T5m observations available across the French Alps, that was previously shown in the Discussion section 5.1 (now Figure 6b in the revised manuscript, Figure 8 in the original manuscript). This figure illustrates that the model biases highlighted in Sect 3.1.2 at the research stations, are representative of the general model behaviour in the French Alps.



New Figure 6 a*. Arome-OPER mean temperature differences between 2 m and 5 m as a function of altitude, for each model grid-point over the study area for winter 2021-2022. Orange line denotes median, boxplots mark the 25%-75% percentiles, blue whiskers the 5%-95% percentiles and dots the values outside this range.*

1.4 - It was not clear why results were presented in the order they were presented, and it is not particularly easy to follow. A clear explanation for the paper’s logical flow to start the results section, e.g., focusing on addressing specific science questions, would be very useful.

As an answer to the concern of clarity expressed by both reviewers, we thoroughly revised the structure of the manuscript. The main modifications are the following :

- First, **also as an answer to a concern expressed by Referee#2**, we made **the research questions** addressed and the hypothesis tested within the manuscript **much clearer in the Introduction**. We formulated them so that they provide the overarching structure of the manuscript, see in particular the modifications in the last paragraph of the Introduction, reported hereafter :

“In a nutshell, the present study intends to draw the light on some pitfalls affecting the use of the near-surface air temperature observations in mountain terrain for numerical weather forecasting, through addressing a series of research questions :

Taking the example of the Arome-France NWP system that operationally runs over a large alpine region, we will first address the question of the impact of varied sensors' heights above surface, on the assessment of model performances. One of the underlying questions is whether observations acquired at 2m to about 5m above the snow surface, can be used without specific treatment to evaluate model performances, or whether they should be considered separately, as revelatory of different model behaviors. Through this analysis, we intend to provide guidelines for the use of temperature measurements for model evaluation in mountain regions.

In a second subsection of the Results, we will evaluate the effect of this height heterogeneity on the way the model is corrected by assimilation. This subsection will answer the question of whether the height of the observation above the surface matters for assimilation, or whether it is not necessary to discriminate between temperatures from 2 to 5m above the surface for the assimilation. In particular, we will examine the assimilation of mountain near-surface temperatures as a possible cause for the cold bias of Arome.

Finally, another question poorly addressed in existing literature, is how the relief mismatch between observation stations and model grid-cell, and valley-vs-mountain heterogeneities in terms of observational density, affect the efficiency of data assimilation. We will address this question in a final Results subsection of this study, through the use of dedicated assimilation experiments.

The plan of our manuscript addresses these items sequentially, after a section dedicated to material, method and study area. To the best of our knowledge these questions have not thoroughly been addressed in mid-latitude mountain regions of the world. We focus on winter conditions as the period when the model biases are the strongest. We also take the opportunity to propose in a Discussion section perspectives to circumvent the problems highlighted, for the benefit of weather forecasting in complex terrain.”

- Second, the **Results section 3** is now clearly structured into **3 subsections** dedicated to each of these 3 research sections:
    - **3.1 Impacts of heterogeneous sensors’height** (t5m vs t2m) **on model evaluations**
    - **3.2 Effects of heterogeneous sensors’height** (t5m vs t2m) **on the 3DVar assimilation**
    - **3.3 Effect of other heterogeneities within the mountain observation network** (namely : altitude differences between stations and model, and station density heterogeneity), **on the 3DVar assimilation**
- As already mentioned earlier, a new subsection was created within 3.1 to demonstrate the spatial generalization of the differences between the modelled t5m and t2m seen at research sites (see comment 1.3). This subsection also integrates a full evaluation of the t2m and t5m model biases at all stations available within the study area, showing consistent results with the evaluation performed at the 2

research sites. This general evaluation was previously part of the Discussion, and we agree with both reviewers that it rather belongs to the results section.

- **A better distinction was made between Results, Discussion and Conclusion:** as stated above, we relocated parts of the Discussion to Results, but we also relocated parts of the original conclusion, into the Results section (see the penultimate comment by Referee 1, also raised by Referee 2). We completely rewrote the Conclusion, following the 3 overarching research questions and synthesizing the main results relevant to each of them.
- Finally, we added within the Data and Methods “Assimilation experiments” subsection a new paragraph dedicated to expliciting the methods used to analyse these experiments, based on the analyses increments. The questions raised by both reviewers regarding the original Figure 7 of the manuscript, showed that an explanation of these methods would be a beneficial addition to the text.
- We are currently finalizing the revised version of the manuscript where the referees will be able to see these changes.

Specific comments:

L18: “Becken (2010)” citation should be inside parenthesis. The reference is now corrected.

L25: Also, at local & global scales. The sentence has been rephrased.

L28: “were” rather than “are”, and citations in parenthesis. This has been corrected.

L34: “high” altitude regions. This has been corrected.

L33-36: Awkwardly worded sentences, suggest revising. The sentence has been rephrased.

L95-96: This is a crucial statement for the paper’s scope and therefore requires citation(s).

It is difficult to find a citation that mentions that the specific characteristics of mountain stations, and in particular the height of the measurement, are not taken into account, because the Arome assimilation system currently treats all stations identically. This seems so obvious that it is not even mentioned. As an illustration, the recent Marimbordes et al., 2024 paper that describes the forthcoming evolutions for the surface assimilation (CANARI part) in AROME, presents in its Figure 3 the map of the 2-m temperature observation stations assimilated within CANARI. This map features stations above 2700 m altitude, all of which are actually Nivôse where air temperature is measured at about 7 m above the snow-free ground (a zoom helps notably distinguish the Ecrin-Nivôses (2970 m a.s.l.) and the La Meije-Nivôse (3100 m a.s.l) as white dots across the 45° parallel on the Figure). But the height of the measurement is not mentioned. We added this example in the revised manuscript to anchor our statement in recent scientific literature on assimilation incl. over mountain regions :

“As an illustration, Figure 3 in Marimbordes et al. (2024) shows a map of so-called "2-m temperature observations stations that are assimilated" in the surface assimilation. This map includes high-altitude (> 3000m a.s.l.) stations from the Meteo-France "Nivôse" observation network, that measure air temperature actually at roughly 7.5 m above snow-free ground”

Paragraph starting in L53: This paragraph seems to focus on cold biases, but biases reported as positive values. If the bias is a cold bias, then it should be reported as a negative number (i.e., model – obs). This has been corrected.

The introduction could also benefit from including the motivation of the snow-albedo feedback. That is, surface air temperature biases can propagate to snowpack biases (e.g., in snow cover) which can have albedo feedbacks due to the high albedo of snow that in turn feedback to and increase the original temperature biases.

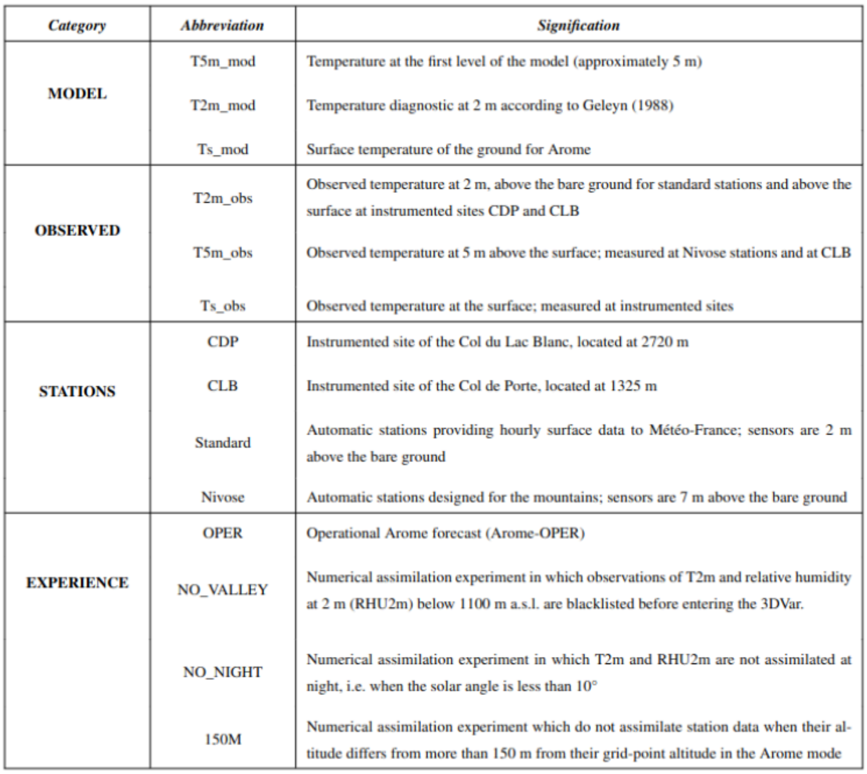
We recognise the snow albedo feedback as an important motivation for our work and added a dedicated paragraph :

"Several publications have pinpointed the links between snow cover and near-surface temperature (cold) biases, with the persistence of a too-extended snow cover and possible limitations in snow-atmosphere exchanges and representation of ablation processes in the models, invoked as possible sources for too cold temperatures over snow (Vautard et al., 2013 ; Kotlarski et al., 2014). In particular, near-surface air temperature is involved in the estimation of the snow-albedo feedback (Scherrer et al., 2012), a mechanism by which snow aging and/or disappearance, enhancing the surface albedo, leads to an increased absorption of solar radiation by the surface and further surface warming or melt (Peixoto and Oort, 1992). Winter et al. (2017) and Kotlarski et al. (2015) have among others highlighted the links between temperature biases in high-resolution climate models and the magnitude of this feedback, with models suffering from negative biases over snow and ice artificially overestimating the temperature response upon snow disappearance."

Figure 1 should be presented more clearly, (e.g., with (a), (b), (c), etc) labeling to show the flow of the figure. This has been corrected.

There are many definitions and abbreviations used throughout the paper. There should be a table in Methods which clearly defines these.

We will introduce one of such tables in the revised version of the manuscript, in the Material and Method section or as an Appendix, based on the example below but with shorter descriptions in the last column :



Category	*Abbreviation*	*Signification*
MODEL	T5m_mod	Temperature at the first level of the model (approximately 5 m)
	T2m_mod	Temperature diagnostic at 2 m according to Geleyn (1988)
	Ts_mod	Surface temperature of the ground for Arome
OBSERVED	T2m_obs	Observed temperature at 2 m, above the bare ground for standard stations and above the surface at instrumented sites CDP and CLB
	T5m_obs	Observed temperature at 5 m above the surface; measured at Nivose stations and at CLB
	Ts_obs	Observed temperature at the surface; measured at instrumented sites
STATIONS	CDP	Instrumented site of the Col du Lac Blanc, located at 2720 m
	CLB	Instrumented site of the Col de Porte, located at 1325 m
	Standard	Automatic stations providing hourly surface data to Météo-France; sensors are 2 m above the bare ground
	Nivose	Automatic stations designed for the mountains; sensors are 7 m above the bare ground
EXPERIENCE	OPER	Operational Arome forecast (Arome-OPER)
	NO_VALLEY	Numerical assimilation experiment in which observations of T2m and relative humidity at 2 m (RHU2m) below 1100 m a.s.l. are blacklisted before entering the 3DVar.
	NO_NIGHT	Numerical assimilation experiment in which T2m and RHU2m are not assimilated at night, i.e. when the solar angle is less than 10°
	150M	Numerical assimilation experiment which do not assimilate station data when their altitude differs from more than 150 m from their grid-point altitude in the Arome mode

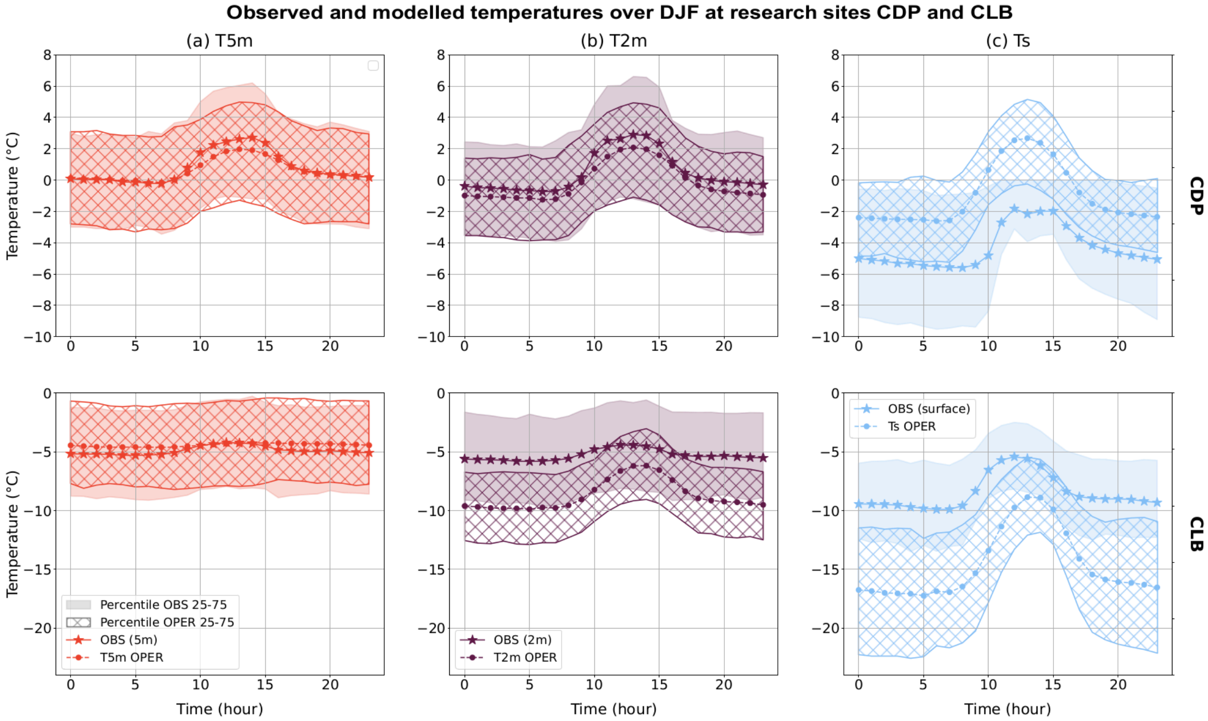
Table A1. Key abbreviations used in the study for modeled and observed temperatures, type of stations and numerical assimilation experiments

Figure 4: It may be more useful to have OPER and OBS lines on separate panels, and show shading for respective lines to represent temporal variability.

One of the purposes of Figure 4 is to compare model behaviour (OPER) to observations, as analysed in the submitted version in section 3.1.2.

As separating OPER and OBS on different panels makes this comparison much uneasier, we preferred to stick to representing them on the same panel. However we welcome the suggestion of the reviewer to also represent variability. To do so while limiting the complexity of the figure, we now propose 3 separate panels for T5m, T2m and Ts, enabling the representation of variability as well as a comparison between model (Arome-OPER) and observations for each of these temperatures. See this new Figure 4 and its caption below.

This new Figure seems to us much easier to read, but has the drawback that it makes it difficult to compare T5m with T2m and Ts (be it observation or model) at each site, so that we propose to keep the original Figure in Supplementary.



***New Figure 4.** Diurnal cycle of the 5 m (a, red), 2 m (b, violet) and surface (c, blue) observed (OBS) and modeled (OPER) temperatures averaged over the winters of the study period at the CDP and CLB research sites. The shaded (resp. hatched) areas represent the observed (resp. modelled) variability through the 25-75% percentile range.*

Can you provide an explanation for the differences between the measurement heights, particularly why max daily T2 is larger than max daily T5, but T2 is lower than T5 in most other time steps (at CDP); whereas at CLB, T5 is higher at all time steps relative to T2. Importantly, because only 2 sites are analyzed, and the sites show differences in patterns, how can results be generalizable?

The different behaviour between the mid-altitude CDP site, and the high-altitude CLB site, can be explained by the differences in dynamics of the lower atmospheric boundary layer between both topographic, meteorological and physiographic settings.

CDP is located at a mid-altitude pass surrounded by elevated mountains (> 2000 m on the eastern side). The site is furthermore located in a large meadow surrounded by ~35 m-high coniferous trees, experiences moderate wind speeds (1.4 m/s mean annual windspeed over 1993-2023) and sometimes exhibits patchy or no-snow conditions even in winter. This, in conjunction with the surrounding forest with much lower albedo than snow, enables the development of a convectively-driven mixed layer on the course of the morning on clear-sky days, whereby the 2m air temperature becomes ephemerally higher than the 5m one at midday. We insist however that the difference between t2m-obs and t5m-obs at midday is lower than 0.5°C and on the order of magnitude of measurement uncertainty.

On the other hand, CLB is a more open and higher altitude site. Because nearby relief is less present, the lower atmospheric boundary layer is more influenced by the nearby free atmosphere and less by the surface. Furthermore, larger wind speeds (mean winter wind speed over 2006-2026 : 4.9 m/s) contribute to a shallower temperature inversion over the snow surface and reduced daily range in temperatures (Oke, 1987). The signature of this is visible in the t2m and t5m daily cycles that exhibit a much reduced amplitude w/r to what happens at CDP. The continuous snow cover around the site over the winter and the absence of surrounding surface elements subject to solar heating, contribute to maintaining a temperature inversion at least up to 5 m height and a shallower boundary layer all day long during winter.

In the end, the main differences between the 2 sites in the winter, lie in a more developed boundary layer for the mid-altitude, that manifests through different amplitudes in daily cycles (well explained by differences in wind speeds and differential heating of the surroundings), and the ephemeral crossing between t5m and t2m at midday at CDP that lies within measurement uncertainty. Otherwise, the behaviours of the lower boundary layer at CDP and CLB do not differ much. As the identified differences are in line with processes described in literature, we do not put in question their general validity. Much more different are the model behaviours at these sites, which require a deeper scrutiny developed in subsection 3.1.2.

Finally, would these discrepancies in diurnal cycles look different for periods of snow cover vs. no snow cover (e.g., winter vs. summer)?

To answer this question we here provide a figure (Figure R1) similar to the New Figure 4, but over summer months JJA:

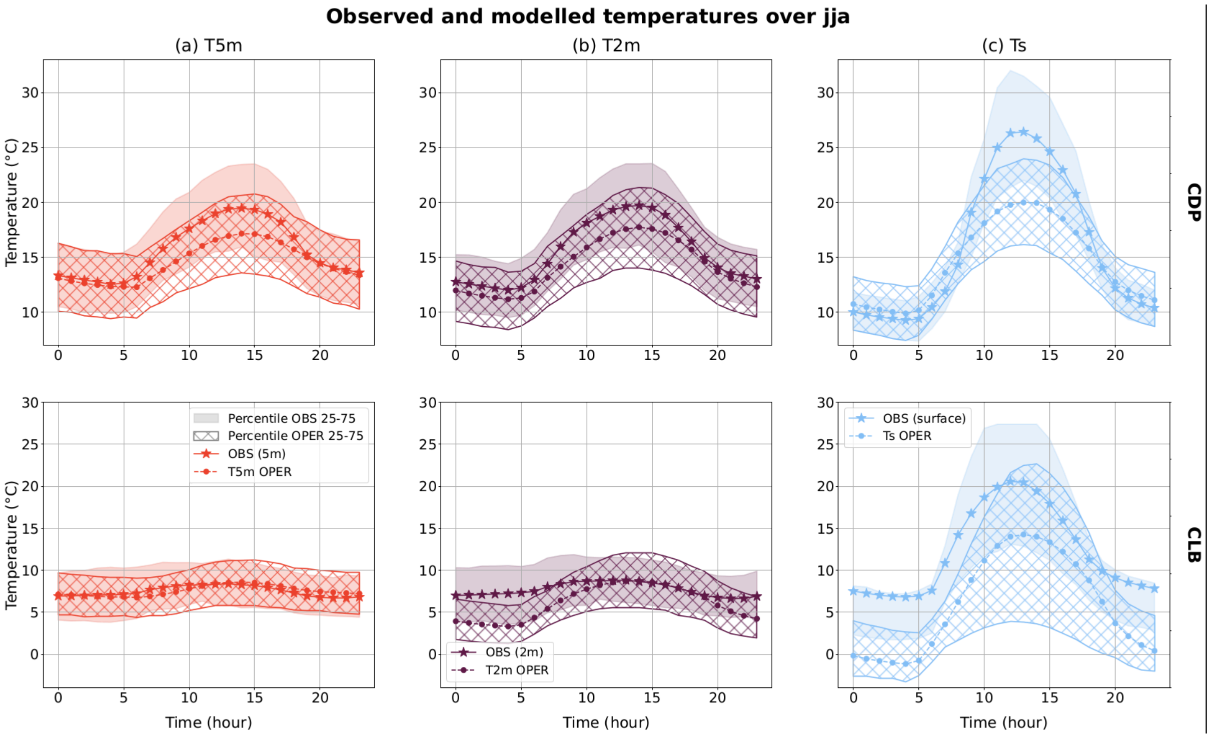


***Figure R1**: Diurnal cycle of the 5 m (a, red), 2 m (b, violet) and surface (c, blue) observed (OBS) and modeled (OPER) temperatures averaged over the winters of the study period at the CDP and CLB research sites over JJA 2020-2022. The shaded (resp. hatched) areas represent the observed (resp. modelled) variability through the 25-75% percentile range.*

These discrepancies in diurnal cycles are mostly different at CDP for the surface temperature, as induced by the contrast between the presence vs absence of snow. Indeed snow is majorly present at this site in winter while completely absent in summer. While the winter surface temperature is therefore capped at 0°C, the summer surface temperature is much higher than T2m and T5m as consistent with the diurnal pattern of radiation and convection development. As the signature of these effects is already present at CDP in winter except for observed surface temperature, due to the presence of surrounding canopies considered in the model grid point, the general pattern in air temperatures above the surface and ranking between model and observations is the same as in winter, with enhanced diurnal amplitudes.

At CLB the differences between both seasons is less marked as snow is regularly present until early to mid-July at the site, making half of the summer months similar to winter in terms of snow surface conditions. The winter and summer dynamics at the site are therefore closer, though with a much weaker T2m negative bias in the model, completely disappearing at mid-day and possibly induced by the development of convection and more air mixing facilitated in snow-free or patchy snow conditions. The model biases in terms of surface temperatures remain of similar magnitude, which could be caused by numerous reasons (soil thermal and optical characteristics and their representation in the model, snow staying too long in the model as assessed by e.g. Monteiro et al., 2024) that are beyond the scope of this paper.

We took the opportunity of this discussion to clearly state in the Abstract and Introduction **the focus of the paper on the winter season when the model biases are the strongest.**

Throughout the paper I recommend using different wording than “guess” which is confusing (e.g., in Figure 6). Guess is also not clearly defined making the results related to this wording difficult to follow

We acknowledge that the term “background” is indeed more common in data assimilation, although the wording “first guess” or just “guess” is also regularly used, at least in the NWP community. All occurrences of “guess” have been replaced by “background”, in the text as well as in the Figures.

L363:366: I am not sure if this makes sense, because the guess at 2m is also much lower than the diagnostic analysis and forecast at 2m as well.

In the Fig 6 of the submitted manuscript, we indeed see that the background at 2 m is colder than the analysis and forecast at 2m at standard stations. There are actually a few differences in the estimation of the background vs the analysis/forecast: first, the former is estimated based on the four nearest neighbours, while the latter are taken at the closest point to the stations. Second, the background comes from a 1h lead time forecast, launched each hour, while the forecast itself comes from the 00h run for its 24 first hours of prevision (24 first “terms” of the prevision). Both effects are responsible for the differences between guess at 2 m on the one side, and analysis/forecast at 2m on the other side (the analysis being warmer than the forecast thanks to the correction induced by the assimilation of observations).

At our 16 mid-altitude standard stations this difference is weak (0.4°C mean difference between guess and forecast) as should be expected as differences between 1h forecast vs 1-to-24h forecast, and between the 2 spatial interpolation procedures used, are usually not major. However it becomes stronger at the 2 high-altitude stations, on the order of 1°C. We think that this is primarily an effect of the 4-nearest-neighbour vs closest-point algorithm applied to extract the background vs the analysis/forecast, and also a side-effect of only 2 stations being used in this Figure for T2m at high-altitudes, namely La Masse and Mont Cenis stations (**Figure R2** below). Indeed, because so few stations are available, "local effects/configurations" have an important impact on the statistics. This is particularly striking in the case of the La Masse station (2800 m a.s.l) that exhibits a strong altitude difference with the model grid-point, making Arome not very representative of the ground-truth conditions. The nearest neighbour point in Arome is at 2548 m a.s.l while the mean of the 4 nearest neighbours are at 2506 m a.s.l.. At this station the difference between guess at 2 m, and forecast, is particularly high and we think that beyond differences in lead-times between both modelled fields, this may likely be due to differences in altitude between the neighbouring points and modelled altitudinal temperature gradients between these points. Altitude differences are lower for the Mont Cenis stations and come together with more limited difference between guess and forecast.

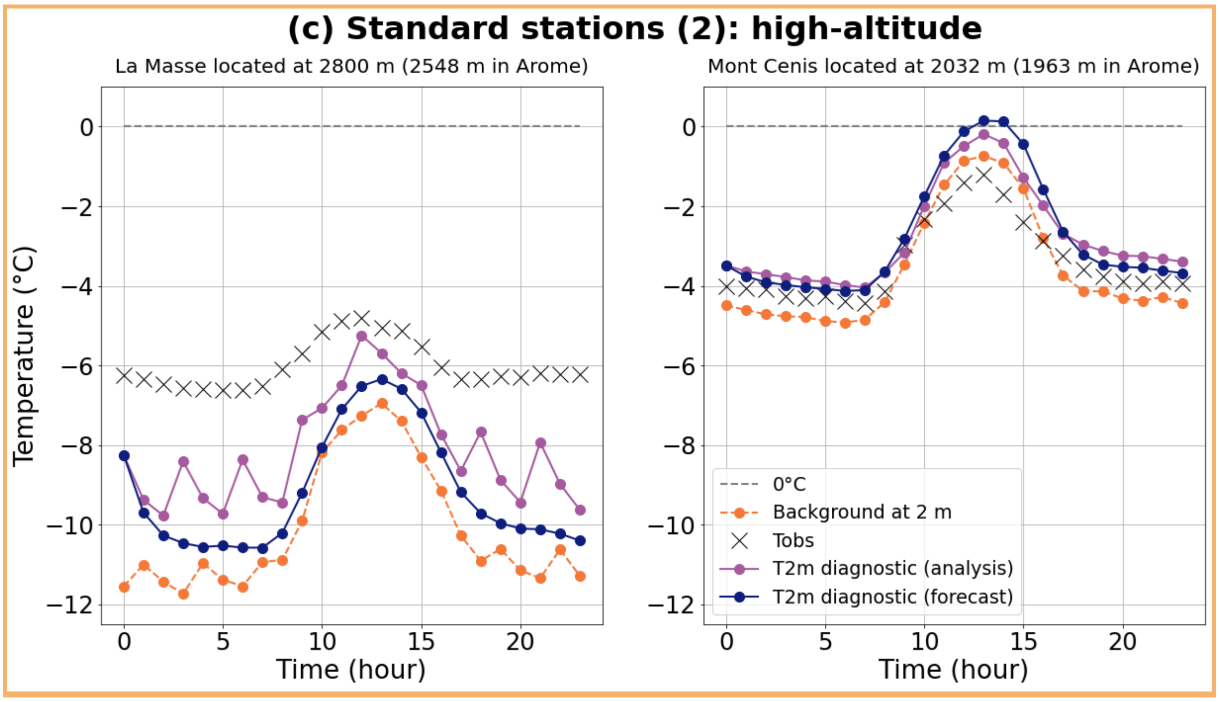


Figure R2: Diurnal cycles of temperature observed (Tobs, crosses) or calculated at different steps within the assimilation workflow of Arome-OPER for the 2 high-altitude standard mountain stations, similar to Figure 6 of the original manuscript.

We added a short explanation of this in the manuscript:

"Note that for technical reasons, the interpolation procedure for the background temperature at 2 m involves the 4 nearest grid-points to the station, and differs from that used for the other model products (nearest model grid-point only). This induces a structural difference between the background at 2 m and for instance the forecast at 2 m, that is usually below

0.5°C but can be enhanced by local effects when only few stations are considered like in Figure~\ref{cycle_assim_REF}c, resulting in that case in a background at 2 m being distinctively colder than the forecast at that height."

Figure 8: it does not seem to make sense that the symbols should be connected with dashed lines. These results may be better presented in a table format than a figure.

We feel that the dashed lines connecting the symbols, as well as the presentation of the results in a Figure, help capturing the altitudinal evolution of the model biases at the different heights. It also makes clear how instrumented sites fill gaps in the standard observation networks, by enabling to assess biases at heights usually not scrutinized in some altitudinal ranges (e.g. the T5m bias at the mid-altitude site CDP). Therefore we stand for keeping this figure as it is, but we also propose to relocate it to Section 3.1 in accordance with the manuscript structural changes recommended by both reviewers, and put it together with another new figure (new Figure 6a, see above) showing the difference in T2m-T5m in Arome-OPER as a function of altitude, in support of the assessment of the generalized behavior of Arome T2m vs T5m across the study area.

Figure 9: pseudo-biases are not clearly defined and therefore it is difficult to make sense of this figure. This is now better explained.

Overall, much of the discussion section seems more like additional results sub sections, rather than a true discussion of the authors' perspectives on the results and insights for future research.

The Results and Discussion sections have been reorganised for greater clarity, see point 1.4.

The Conclusions section should be shortened to more concisely highlight the key takeaways and implications. Much of the discussion that is currently in the Conclusions section may be better placed in the Discussion section.

We agree with the reviewer and relocated most of the previous "Conclusion" into the Discussion section, by creating a first subsection entitled "Summary and general perspectives". We entirely rewrote the conclusion, which now concisely highlights the main findings of the study in link with the research questions.

Please make data used for this study publicly available to support reproducibility.

All data and code to analyze them were provided. We took the opportunity of this revision to better distinguish between code and data availability in the revised version of the manuscript.