Reviewer #2

General comments

Reviewer: The manuscript describes a hybrid approach that exploits the structure of temperature-index (TI) models to derive a machine learning based model that predicts daily surface mass balance. The approach relies on differentiable programming which allows building a model whose structure is inspired by existing and well established MB models but which allows computing gradients and incorporating learnable components which are trained to account for the spatiotemporal variability in the parameters of these TI models. Two flavors of the model, named GlaUnTI, are presented and they differ by the input features that are used.
The proposed architecture is compared against a differentiable TI model and a fully data-driven architecture which does not integrate a physically driven prior in the model architecture. The four models are trained and compared on various glaciers in Europe to illustrate the transferable properties of the approach.
Finally the studies in the appendices showcase applications of the differentiable approach for inverse modelling, model explainability as well as uncertainty quantification.

The manuscript is well written and very interesting to read. The development of transferable mass balance models that leverage both glaciological and remote sensing based products are timely and quiet new in the literature. Training a model at a daily time step with a term in the loss function that aggregates yearly and at the glacier scale is far from being easy. The comparison of the proposed approach with a fully data-driven model is appreciated, especially since there starts to be multiple works in the literature that build MB models on this class of fully data-driven architectures.
The validation of the trained model is performed in a robust way through independent validation and test sets. Some of the potential caveats in model comparison (models C vs D) are properly treated. However the comparison with other models could be improved.
The illustration of the potential of the approach for inverse modelling thanks to differentiable properties is much appreciated and is part of the contribution of this work, which in the opinion of the reviewer should be more tightly integrated into the main body of the manuscript.

On the overall this is a good contribution that perfectly fits into the scope of The Cryosphere. However there are a few issues that should be addressed and that would strengthen the comparison and the contribution.

Response: We sincerely thank the reviewer for the careful and constructive assessment of our manuscript. We are grateful for the positive evaluation of the proposed modelling framework and the appendix-based demonstrations. The comments were very helpful in improving the clarity and rigour of the study. We have addressed the points raised below and suggested the revisions accordingly.

Specific comments

1. Major comment: Quantify the impact of ice dynamics on the SWE representation.

The SWE variable represents the accumulated snow within a grid cell but this assumes that there is no ice dynamics and that a column of snow does not move. Incorporating ice dynamics would obviously make the problem much harder but at the same time the initial state of SWE is obtained with a spin-up of 5 years and for some glaciers this can change a lot the topographical and climate conditions. Could the authors quantify the impact of not representing the glacier dynamics?

Response: We thank the reviewer for raising this important limitation, which we admittedly overlooked. We agree that the SWE state in our model is not advected with ice flow, indeed introducing additional modelling errors. In the present implementation, glacier geometry is time-dependent through annually varying surface elevation fields and glacier outlines, but SWE itself doesn't move horizontally. Directly quantifying the effect of ice advection on SWE would require coupling GlaUnTI to an ice flow model, which is beyond the scope of the present study (but can and should be done in the future).
Nevertheless, we expect this limitation to have a secondary effect for the present experiments. First, SWE is used only internally to separate snow and ice melt through the fractional snow cover term. We do not notice obvious performance drops that would be related to that. Second, persistent SWE is expected mainly in accumulation areas, where glacier surface velocities are typically lower than in the main ablation tongues, while in lower elevation areas the seasonal snow column is often removed by melt before long-distance ice advection becomes important. Third, we do not include fast flowing or surging glaciers, where the static column assumption would be particularly problematic. We also note that fixed-geometry assumptions are common in mass balance modelling studies with comparable or longer simulation periods (e.g., Laan et al., 2025; Schmidt et al., 2023).
As an indirect sensitivity test, we trained an additional one-DDF TI model in which the SWE state, snow/ice partitioning and multi-year SWE memory are removed entirely:

$$\widehat{\mathrm{SMB}}_{ijd} = \beta_1 P_{ijd}^{\mathrm{solid}} - \beta_2^* T_{ijd}^{+} ,$$
$$P_{ijd}^{\mathrm{solid}} = P_{ijd} \cdot \sigma\left(\tau_{P,\mathrm{s}}\left[\tau_{P,\mathrm{c}} - T_{ijd}\right]\right) ,$$
$$T_{ijd}^{+} = \mathrm{softplus}_{\varsigma}\left(T_{ijd}\right) ,$$

where β_2^* is a single DDF replacing β_2 and κ in the two-DDF formulation, and the remaining parameters correspond directly to the TI implementation used in the manuscript. This model was trained identically to the two-DDF TI baseline. Since the one-DDF model contains no SWE state, it provides a sensitivity test of whether the local SWE representation is essential for the reported TI model performance.
The comparison shows that the performance differences are small (Fig. 1). Thus, removing the SWE state altogether does not substantially degrade point-level performance. It indicates that the static snow column is not a dominant source of error in our results.

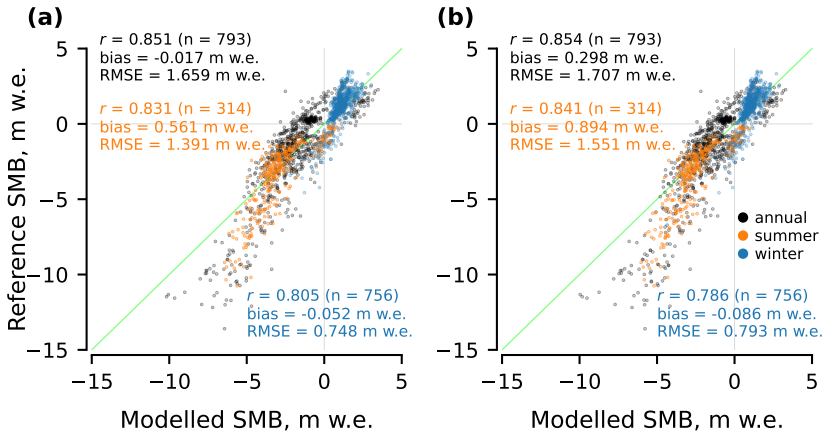

Figure 1: Point-level performance of (a) the one-DDf TI formulation and (b) the two-DDF TI formulation.

Notably, in our response to Reviewer #1, we added qualitative examples showing that the modelled SWE states are spatially coherent and that the model (Model D) generally reproduces plausible equilibrium-line positions relative to the end-of-ablation-season facies maps. This provides an additional qualitative check that the SWE mechanism behaves consistently. We also use this fact to justify the choice of the two-DDF formulation as the basis for GlaUnTI, even if one-DDF showed slighly better results here.
We prefer to keep the one-DDF sensitivity experiment in the rebuttal rather than adding it as a new appendix/subsection, because it is an auxiliary check of one modelling assumption rather than part of the main model comparison. However, we agree that the limitation itself should be stated explicitly in the Discussion. We therefore suggest adding the following sentences:

> . . . In a related way, point measurements themselves are not necessarily representative of the mean SMB of a 100 m grid cell resolution, particularly on large, topographically complex glaciers with strong microclimate gradients, wind redistribution and variable shading. The signal at the point scale mixes true process mismatch with unresolved subgrid variability and measurement representativeness. Another unresolved process is the absence of explicit advection of the SWE state. SWE is not transported with glacier flow, so coupling GlaUnTI to an ice dynamics model would be required for long-term projections or fast flowing glaciers. Similarly, uncertainty in surface elevation propagates into the temperature forcing, introducing additional errors at the point scale. This matters . . .

2. Major comment: Comparison with the autodiff-friendly TI model is not fair.

Usually TI models are calibrated per glacier which is not the case in this study where the parameters are calibrated across different regions. In classical TI models these parameters reflect the various climate conditions which are expected to vary across different glaciers.
While the choice of tuning TI parameters across different glaciers is understandable to be able to assess the performance on an hold-out set, this leads to a poorer performance. A comparison with the autodiff-friendly TI model with parameters tuned per glacier would allow assessing the real performance gain, beyond the intrinsic transferability property machine learning models have over TI models.

Response: We thank the reviewer for this comment. We agree that classical TI models are often calibrated independently for each glacier, and that such a setup can improve their local performance because the parameters can adapt to glacier-specific climate, geometry and forcing biases. However, this comparison answers a different question from the one targeted in our study. Our main evaluation focuses on spatial transferability. Models A–D are trained without using any observations from the test glaciers, and are then evaluated on these fully withheld glaciers. A per-glacier calibrated TI model, by construction, uses observations from the same glacier on which it is later evaluated, and therefore no longer represents a spatially transferable model in the same sense. Such a comparison is not a fully fair transferability benchmark against GlaUnTI, which does not see the test glaciers during training.
We also note that, operationally, the motivation for GlaUnTI is precisely the limited availability of glaciological measurements. If point SMB measurements are already available for a glacier, a locally calibrated TI model

is a reasonable and strong baseline. The more difficult and practically relevant question is whether a model trained on monitored glaciers can be transferred to glaciers where such local glaciological calibration data are unavailable. This is the setting addressed by our spatially independent test split.
Nevertheless, we followed the suggestion and performed an additional diagnostic comparison. We finetuned Model A separately for each test glacier, following the transfer learning setup used in Appendix A, while reserving 2019–2024 for evaluation. This provides a deliberately favourable setting for the TI model, because it is allowed to use earlier observations from the same glacier. We then compared this per-glacier finetuned TI model with the original spatially transferable Model D over the same 2019–2024 evaluation period.
We were positively surprised by the result. Even in this favourable setup for the locally calibrated TI model, Model D remains highly competitive and shows consistently lower biases and comparable RMSEs across annual, summer and winter point SMB (Fig. 2), despite never having seen the test glaciers during training.

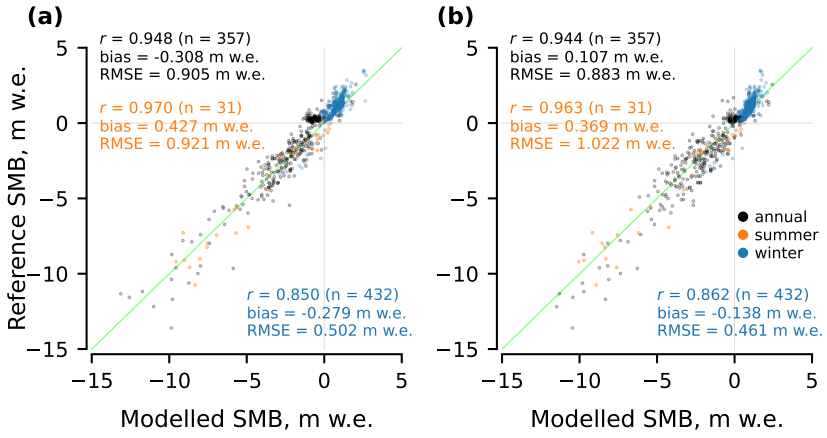

Figure 2: Point-level performance of (a) the per-glacier finetuned TI model and (b) Model D for 2019–2024.

This comparison supports the central goal of the study. GlaUnTI approaches the performance of a glacier-specific TI calibration while retaining spatial transferability. We suggest including this comparison as an additional subsection or appendix in the revised manuscript, depending on how it fits during preparation of the revised version.

3. Major comment: Number of training iterations for the fully data-driven architecture is insufficient.

Given that updates are performed at every epoch there is no stochasticity in the training, which is usually a property that helps to explore the parameter space. The low number of epochs is probably not enough for the fully data-driven architecture to capture the relationship between the inputs and the output. A training with more iterations should be performed.

Response: We thank the reviewer for raising this point. Of course, we monitor loss-epoch curves during training. Because of that, we used 100 epochs to train the GRU, as 50 epochs were not enough (unlike for the rest of the models). After 100 epochs, the GRU had almost reached the plateau, and we did not expect significant improvements after that. Nevertheless, as both reviewers raised the understandable concern of not having enough epochs to train the GRU, we increased the computational budget for the GRU to 300 epochs ($\sim$ 36 hours). This led to an improvement in the validation performance (Fig. 3).
Ironically, while the validation performance improved, the testing performance slightly dropped, indicating a lower tolerance of the purely data-driven baseline to domain shifts (Table 1).

Table 1: RMSE of Model B (the GRU) on the test subset before and after training for more epochs.

Type	Period	Before, m w.e.	After, m w.e.
Point	Annual	2.032	2.066
	Summer	2.012	1.969
	Winter	0.869	0.875
Glacier-wide	Annual	0.930	1.009
	Summer	0.752	0.804
	Winter	0.904	0.921

Given that the validation loss is lower after longer training, we *have to* report the results for this new model in the revised manuscript. We therefore will update all figures for Model B in the manuscript. This, however, doesn't change any derived conclusions.

4. Major comment: Incorporate the appendices more closely into the main body of the paper.

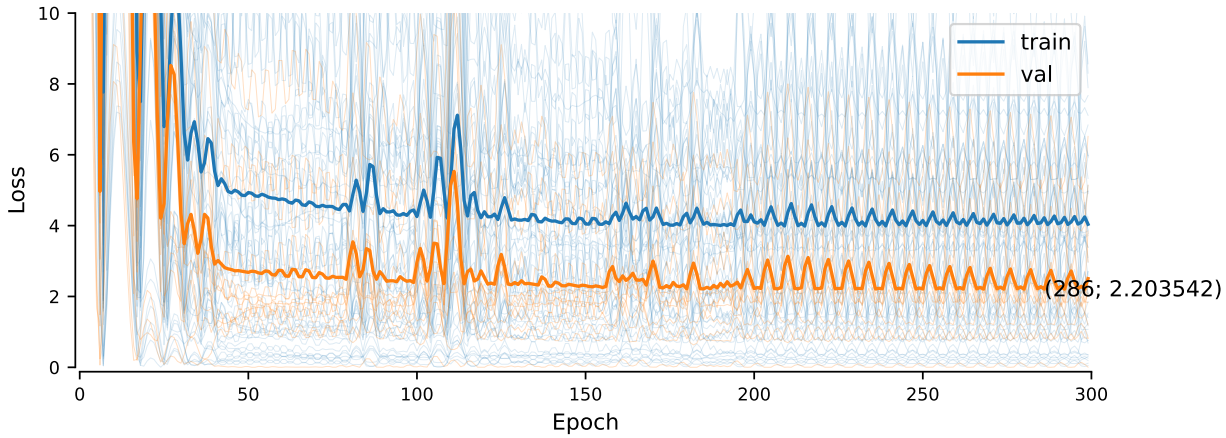

Figure 3: Loss function dynamics for Model B (the GRU). The plot area before Epoch 100 corresponds to the initial submission. Semi-transparent curves show loss dynamics for individual glaciers. The best validation loss value is labelled.

Beyond having a transferable MB model, differentiability is key for inverse modelling. Many scientific questions in glaciology require the representation of both ice dynamics and mass balance over long simulation periods. Having a differentiable MB model is one of the key components to tackle these questions. The authors illustrate this potential in a convincing way through different and complete experiments in the appendices. In the opinion of the reviewer this should be included as one of the main messages of the paper and in the current version the appendices are decorrelated from the rest of the manuscript.

Response: We thank both reviewers for highlighting this point. We, however, decided to keep these experiments in the appendix. The central aim of the manuscript is to introduce and evaluate GlaUnTI as a transferable SMB model, with the main results focused on predictive performance, spatial transferability and the comparison with the TI and data-driven baselines. The differentiability experiments are intended as proof-of-concept demonstrations of additional capabilities enabled by the autodiff-friendly implementation, rather than as full-scale case studies evaluated across the complete glacier sample. Moving them into the main results would risk shifting the emphasis of the paper towards these demonstrations relative to the main transferability analysis. We think that keeping them in Appendix A optimally preserves the narrative flow of the manuscript, while the main text makes clear that differentiability is an important methodological implication of the overall framework.

5. Minor comment: More details should be provided on the glacier facies maps.

More information should be given in section 3.5 about the glacier facies maps. The classes, which are ice, snow, debris, firn and refrozen-like according to Maslov et al. (2026), are not defined and giving them would make it clearer for the reader of what information these maps carry.

Response: We thank the reviewer for this comment. We suggest the following modificatons to the manuscript.

> Hence, in cases of available optical satellite imagery from Landsat or Sentinel-2 at the end of summer (31st July–29th September), we employed the ~~glacier facies classification models~~compact convolutional neural networks trained in the cross-validation folds of Maslov et al. (2026) to classify glacier facies and ensembled their predictions ~~as follows~~. This classification workflow yields five glacier facies classes (ice, snow, debris, firn and refrozen-like) and three auxiliary classes (shadow, water and cloud). For each pixel and class c, we obtained aggregated predictive confidence as: ...

We also presented several examples of the derived glacier facies maps and how they compare with the distributed annual SMB in a new Results subsection (see correspondence with Reviewer #1).

6. Minor comment: Clarify the two "regimes" that produce SMB predictions.

According to section 4 the models have two "regimes" to produce SMB predictions but how these two regimes are obtained in practice is not detailed. Are the authors referring to the aggregation in the loss function over a time window like in Eq (5)? If so the statement "considerably reducing the memory footprint" (L193) is only partially true since even though the predictions can be aggregated recursively the inputs still needs to be stored somehow.

Response: By "two regimes", we referred to two implementation modes for the forward pass. In both cases, daily SMB is computed internally using the same model equations. The first mode returns the full daily SMB

trajectory and is used when daily outputs are required, for example for evaluation over exact point-measurement dates. The second mode recursively accumulates the daily SMB over predefined seasonal windows, as in Eq. (5), and returns only the accumulated SMB. This does not eliminate the need to store the input forcing data, which are kept *in regular RAM* and fetched lazily for the relevant glacier and time window, but it substantially *reduces the VRAM footprint* because dense daily SMB fields and intermediate activations do not need to be materialised over the full trajectory. We adjusted our wording accordingly:

> ... All models use daily fields of temperature and precipitation as primary predictors and have two regimes of producing SMB predictions. The first regime predicts daily SMB fields, while the second one predicts accumulated SMB by aggregating daily predictions on the fly over the simulation period without materialising full daily SMB trajectories, considerably reducing the GPU memory footprint and hence enabling training on large grids and long temporal trajectories. Regardless of the model, ...

We also rely on rematerialisation a lot to make gradient computation possible at all.

7. Minor comment: Clarify if normalization is applied to the inputs of the deep learning corrector.

This is not detailed in the manuscript, but there is probably a normalization given the heterogeneity of the input data ranges. The bounds should also be given for reproducibility.

Response: Yes, input normalisation was employed for the neural components. For the GRU baseline, the daily temperature and precipitation inputs were divided by fixed maximum-absolute scaling constants, 45.744850158691406 °C and 0.16851751506328583 m w.e., respectively, both specified in `constants.py`. For the neural corrector in GlaUnTI, the daily temperature and precipitation fields entering the TI backbone were not normalised, because these variables appear in the physical TI equations in their native units. However, the auxiliary inputs of the corrector were scaled: Z, Z_σ, T_M, and P_M were divided by their corresponding maximum absolute values, which are provided with the released dataset in `normalisation_factors.pkl` at `https://doi.org/10.4121/5ea53bc3-2c85-42bb-89d1-606c8ed1d80a.v1`. The initial TI SMB estimate (SMB_0) was divided by 10 m w.e. The perturbation-derived sensitivity fields, binary glacier mask, and facies/confidence predictors were kept in their original ranges. We clarify this in the revised manuscript as follows:

> ..., where $\boldsymbol{x}_{ijd} = [T_{ijd}, P_{ijd}]^\top$ is the input vector, $\boldsymbol{z}$ is the update gate, $\boldsymbol{r}$ is the reset gate, $\widehat{\boldsymbol{h}}$ is the candidate hidden state, $\boldsymbol{h}$ is the hidden state, $\odot$ stands for the Hadamard product, and $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W}_h, \mathbf{U}_z, \mathbf{U}_r, \mathbf{U}_h, \boldsymbol{b}_z, \boldsymbol{b}_r, \boldsymbol{b}_h$ are trainable parameters. In the implementation, the GRU receives normalised inputs, corresponding to the temperature and precipitation values divided by their maximum-absolute values. Similar to the TI model, we initialise $\boldsymbol{h}_0$ as a trainable vector at the year $y-5$ and run the simulation for five consecutive years to obtain the starting conditions for the period of interest...
>
> ...
>
> ... This simple resampling is computationally efficient and provides a regularising effect due to the enforced smoothness of the outputs. Before entering the corrector, $\mathbf{Z}$, $\mathbf{Z}_\sigma$, $\mathbf{T}_\mathrm{M}$ and $\mathbf{P}_\mathrm{M}$ are divided by feature-specific maximum absolute values. $\mathbf{SMB}_0$ is divided by 10 m w.e., corresponding to the typical magnitude order of measured point SMB. The perturbation-derived sensitivity fields, glacier masks and facies predictors are kept in their original ranges. Notably, Eq. (6) remains the main backbone of the GlaUnTI model, and the overparameterised corrector only influences its core parameters, hence, keeping the interpretability of the model high by design...

8. Minor comment: Explain how the folds are constructed.

Given the small number of groups, the performance of the model depends a lot on the construction of the folds. Are they randomly split or is there a smarter construction strategy that leads to better representation in the train, validation and test sets?

Response: We thank the reviewer for this comment. The folds were not obtained by a purely random split. We used a glacier-level approximate stratification procedure to balance regional representation and data availability across folds. Specifically, glaciers were grouped by region and by quantile-binned indicators of the number of point SMB measurements, the number of observation years and the number of available glacier facies maps. Glaciers were assigned greedily to the fold with the lowest current representation of the corresponding stratum. This produced folds with generally balanced regional and observational coverage. We clarify this in the revised manuscript as follows:

> We performed a stratified split of the dataset into six folds, balancing the number of glaciers, the number of glacier-wide SMB estimates, the number of point SMB measurements, the number of available glacier

facies maps and the appearance of the five regions among the folds. The split was performed at the glacier level. For the stratification, each glacier was assigned to a stratum defined by its region and quantile-binned descriptors of data availability, namely, the number of point SMB measurements, the number of observation years and the number of available glacier facies maps. Glaciers were then assigned iteratively to the folds. At each step, the next glacier was placed into the fold with the lowest current number of glaciers from the same stratum. After that, we selected the first fold as the test one, the second one as the validation one and the rest as the training ones. Overall, ...

Technical corrections

Reviewer: L308: "stands for the mean value" → "stands for the spatial mean value" to be more precise

Response: We adjusted the sentence accordingly:

..., and $\overline{\cdot}$ stands for the spatial mean value.

Reviewer: L312 "so that the training dynamics is optimal at the beginning": in which sense? Loss function value? On the training or the validation set?

Response: Yes, we monitor validation loss values in this case. We clarified this moment:

We selected $\lambda_2 = 0.1$, $\lambda_3 = 5$, $\lambda_4 = 20$, $\lambda_5 = 5$ with grid search by running training for a couple of epochs so that the ~~training~~validation loss dynamics is optimal at the beginning; a more fine-grained tuning was not possible due to the high computational requirements.

Reviewer: Section 3.1: The glacier-wide MB targets are model outputs. Since the approach mixes glaciological point measurement with glacier-wide MB values, this should be clearly stated.

Response: We agree that this distinction should be stated explicitly. The point SMB records are direct glaciological measurements, whereas the glacier-wide SMB values are spatially extrapolated estimates. We therefore use them rather as spatial regularisers and also downweight their contribution during training. We clarify this in the revised manuscript as follows:

... We harmonised all data to be largely compatible with the WGMS format. We distinguish between point SMB measurements, which are direct glaciological observations, and glacier-wide SMB estimates, which are derived by the corresponding monitoring programmes through spatial extrapolation of the glaciological observations over the glacier area. The latter are therefore used here as a weak constraint rather than as direct measurements in our modelling framework. All point measurements ...

Reviewer: Section A1: Emphasize that in a perfect modelling framework we would have to change also the distribution of precipitations since in a changing climate we expect the distribution of climate variables to change.

Response: We thank the reviewer for this comment. Indeed, temperature is quite unlikely to change completely independently from precipitation. We suggest the following modification:

... Minimising L w.r.t. $\Delta T_{2015...2024}$ then answers the question: "What temperature change during 2015–2024 is required to bring the modelled glacier-wide SMB into equilibrium with the mean climate state over 2020–2024?" This setup should be interpreted as a deliberately reduced inverse experiment rather than a complete climate-change scenario. In a more complete setting, the joint distribution of temperature and precipitation would need to be modified, including changes in precipitation amount and frequency. ...

References

Laan, L. N. van der et al. (2025). "Decadal re-forecasts of glacier climatic mass balance". In: *The Cryosphere* 19.9, pp. 3879–3896. DOI: 10.5194/tc-19-3879-2025. URL: https://tc.copernicus.org/articles/19/3879/2025/.

Schmidt, L. S. et al. (2023). "Meltwater runoff and glacier mass balance in the high Arctic: 1991–2022 simulations for Svalbard". In: *The Cryosphere* 17.7, pp. 2941–2963. DOI: 10.5194/tc-17-2941-2023. URL: https://tc.copernicus.org/articles/17/2941/2023/.