Dear Reviewers and Editor,

We want to thank you for putting the time and effort into reviewing our manuscript and for your valuable comments and questions. We are certain that the revised manuscript benefited from your input.

The main points we took away from both reviews is that the manuscript needs (1) a more thorough evaluation of the model uncertainties, (2) more contextualization of the model regarding existing modeling approaches and (3) a clearer structure and explanations.

To address these points, we made the following main revisions:

- We added a quantitative point-wise evaluation of wind direction and snow redistribution for the case study similar to the evaluation of wind speed. We also expanded the analysis with a contextualization of the SNOWstorm errors with regard to the errors of the benchmark HEF-LES, errors in the coarse-scale input data, as well as errors reported in comparable modeling approaches. We also want to point out that a direct comparison of reported error metrics in other modeling approaches is problematic, due to the differences in settings (e.g., temporal resolutions) and validation strategies. This last point leaves room for a coordinated model intercomparison experiment in the future (for blowing snow models on the one hand, and near-surface wind models on the other hand).
- We added a new section on model limitations, which explicitly summarizes the model uncertainties as presented in the ideal-case experiments and the case study. Additionally, this section summarizes which settings the model is designed for, and which settings are not represented in the model. These include the atmospheric boundary condition, topographic limitations, and limitations in the surface structure (e.g. focus on non-vegetated areas).
- We added additional explanations on the model structure and specific choices, where the reviewers pointed out questions in the text. We also re-worked Fig.1 to facilitate understanding of the topographic analysis.
- We agree with the reviewer comment that the case study is rather short, and that conclusions drawn from this are only valid for this one case. However, we believe that evaluation against the HEF-LES (which introduces the short-period constraint) can provide valuable insights on a process level. We therefore put an even stronger focus on the process perspective in our evaluation. This includes an analysis of which flow features in the HEF-LES are, or are not, represented in the SNOWstorm predictions.

In the following, we address all specific points; please find our answers in blue, with the changed text in italic.

Yours sincerely,
Manuel Saigger on behalf of all co-authors

Response to Reviewer 1

Summary

This manuscript presents a new machine-learning-based approach to simulate near-surface winds, snow drift sublimation and transport over complex topography at a high spatial resolution of 50 m. The model is trained on controlled large-eddy simulations (LES) conducted over synthetic topographies, which supports its broad applicability. An independent validation on a glacier in the Austrian Alps demonstrates comparable wind fields and snow redistribution patterns to those obtained from numerical LES results.

This novel downscaling model for wind and snow redistribution opens the door to longterm, high-resolution glacier mass-balance applications. However, I have several major and minor concerns in its current form. Most importantly, the proposed approach is not yet sufficiently situated within the context of existing models. The manuscript would benefit from clearer benchmarking against established wind and snow redistribution models, a more explicit discussion of how specific modeling choices relate to existing approaches (and the motivation for differing choices), and a clearer definition of the applicability range and limitations of the SNOWstorm framework and its individual components. In particular, additional detail on the training dataset (e.g., the representation of atmospheric conditions and terrain types) would help assess whether model mismatches may stem from limited coverage of certain flow regimes or synthetic topographic characteristics. In this context, the rationale for adopting power-law spectral scaling based on parameters derived from real terrain should be more explicitly justified, especially given the potential uncertainties when extrapolating the model beyond the training range in terms of topography or spatial resolution. Clearer methodological explanations (e.g., regarding the sequential execution and interaction of model components) and a more detailed quantitative evaluation would aid the manuscript. This could include benchmarking modeled snow redistribution rate errors against errors from previous studies and incorporating available observational datasets, such as the mentioned terrestrial laser scans. Finally, the manuscript would benefit from a more transparent discussion of model capabilities and limitations, ensuring that conclusions remain closely aligned with the evidence presented. If these aspects are addressed, the study would make a valuable contribution to the field.

We want to thank the reviewer for their valuable feedback. As stated in the general response above, our revision concentrated on a more thorough presentation of model uncertainties and contextualization in the framework of existing models, as well as on a modified structure that aims to improve readability of the manuscript. This includes a new section on model limitations and applicability.

We agree, that a direct comparison of the simulated snow redistribution with the observed snow height change could strengthen the validation. We can unfortunately not do this here, as the observed snow height change is a difficult signal to interpret. It is not only due to redistribution but confounded by compaction and avalanching, and additional restrictions in the scanning geometry. Disentangling these effects would be out of scope of this work, and has been the focus of dedicated studies like Voordendag et al. (2024). However, the HEF-LES data were validated against the laser scans in the original publication and we use these validated HEF-LES as benchmark in our study. In the section 3.3.1 (case study overview) we emphasize this validation approach more clearly. Apart from that we added a quantitative validation of snow redistribution with respect to the HEF-LES.

Major comments

Please try to be more specific to increase understanding and avoid overgeneralizing results. Below, I refer to my concerns at the corresponding text positions.

Abstract:
Line 7: ”[...] to be applicable over a wide range of atmospheric conditions and for a wide range of mountain regions.”: Be more clear: Aren’t the atmospheric conditions only for winter time conditions of glaciated mountain regions in mid- to high latitudes and the mountain regions for rather moderate slopes of maximum 40°?
That is correct, we agree with your suggested specifications.

“to be applicable for winter-time conditions in glaciated mountain regions in mid- to high latitudes.”

Line 11: ”In a first real-world application study in the European Alps, SNOWstorm predicts wind fields [...]”. : Be more specific: ”on a glacier located in the European Alps”.
We changed that.

“on a glacier in the European Alps,”

Introduction:
Line 71-72: ”The model of Le Toumelin et al. (2023) used the data set of idealized numerical simulations of Helbig et al. (2017) as training data.”: Be more specific: [...] idealized numerical simulations across diverse synthetic topographies [...]
We changed that.

“The model of Le Toumelin et al. (2023) used the data set of idealized numerical simulations across diverse synthetic topographies of Helbig et al. (2017) as training data.”

Line 72-73: ”Despite the successful implementation, this model has the shortcomings of assuming a neutral stratification of the atmosphere, neglecting turbulent motions and assuming a linear dependence in the wind velocity.” : Please be more specific. While Le Toumelin et al’s wind downscaling model assumes linear scaling with respect to coarse wind velocity, consistent with linear flow theory, the model itself is also a nonlinear convolutional neural network and does not assume linear terrain-flow interactions.
You are correct on the non-linear nature of the CNN, our point refers to the linear scaling with respect to the coarse input wind (Eq.3 of Le Toumelin et al. (2023)), which we see as problematic, as linear flow theory assumes velocity perturbations to be small compared to the background flow (e.g., Nappo, 2013), and this might not hold in situations of highly turbulent terrain-modified flow. Considering this, we decided to include a large range of background flow velocities, to not rely on linear velocity scaling.
We agree, that this point could be misunderstood in our text, and we stated this more clearly in the revised manuscript.

“Despite the successful implementation, this model has the shortcomings of assuming a neutral stratification of the atmosphere, neglecting turbulent motions and applying a linear scaling with respect to the coarse-scale wind velocity.”

Line 76-84: ”[...] we specifically aim for these characteristics:” Are these goals fully achieved with SNOWstorm or not? Be more specific, several goals seem really vague, e.g. ”-large speed up rate compared to conventional numerical simulations of several orders of magnitude,” or ”representation of turbulent motions in the atmosphere,”.
We agree that these goals should be stated more specifically:

- *“Large speedup rate compared to conventional numerical simulations to be feasible for multi-seasonal applications on a regional scale”*
- *“explicit representation of near-surface large turbulent structures in the wind field”*

Data and methods:
Line 105-110: While it is briefly mentioned that the approach for the SNOWstorm development strategy was based on work from Helbig and Löwe, 2012, Helbig et al, 2017 and Le Toumelin et al, 2023, this should be clearly indicated in other sections such as the introduction in the manuscript. Le Toumelin et al, 2023 developed a Convolutional Neural Network model for near-surface winds based on Helbig et al, 2017's database consisting of thousands of atmospheric model simulations on Gaussian Random Fields as topographies.
Please be also more concrete in ”Atmospheric simulations run on such topographies can inform downscaling tools for, e.g., wind (Helbig et al., 2017; Le Toumelin et al., 2023).”
We agree with the reviewers comment that the basic idea for our model builds on the previous work mentioned, though expanding on it. We re-structured this section slightly, to clearly state the underlying idea in the opening of this section (2.1, before 2.1.1)

“The main goal of this work is to build a model that is applicable for a wide range of atmospheric conditions and over a wide range of regions, focusing on winter-time mountain environments in mid- to high latitudes. Work by Helbig and Löwe (2012) showed that essential characteristics of real terrain like slope statistics can be represented by artificial topographies such as Gaussian Random Fields. Atmospheric simulations run on a large set of these synthetic topographies have been used to develop downscaling tools for near-surface wind, subsequently applicable for real terrain (Helbig et al., 2017; Le Toumelin et al., 2023). Building on this idea, of representing real-world characteristics in a synthetic setting, with the goal of a more universally applicable tool, we develop our training data. The focus of our approach, however, lies in capturing the harmonics in the atmosphere-terrain interaction for terrain-induced flows with large-scale forcing, as previously done for modeling approaches building on linear mountain wave theory (e.g., Smith and Barstad, 2004; Sauter, 2020). For this, the synthetic topographies used in our approach are designed to reflect the range of spectral characteristics of real terrain. In the same way, the atmospheric conditions represent the range of harmonic properties found throughout winter-time mountain regions in mid- to high latitudes.”

Line 105-110: The motivation and reasoning for choosing topographies with power law spectral scaling are not made clear. Particularly, given that your approach is built on that from Helbig et al, 2017, who used Gaussian over power law topography models based on a better statistical agreement with real terrain slope characteristics, and given that power laws do not hold across spatial scales.
See comment above, we re-structured the first paragraph to emphasize the idea in more detail in the revised manuscript.
The main idea is to represent the harmonics of the terrain-atmosphere interaction. This is the reason why we want to represent the whole range of static stability and wind speed in the training data on the one hand and terrain harmonics on the other hand.

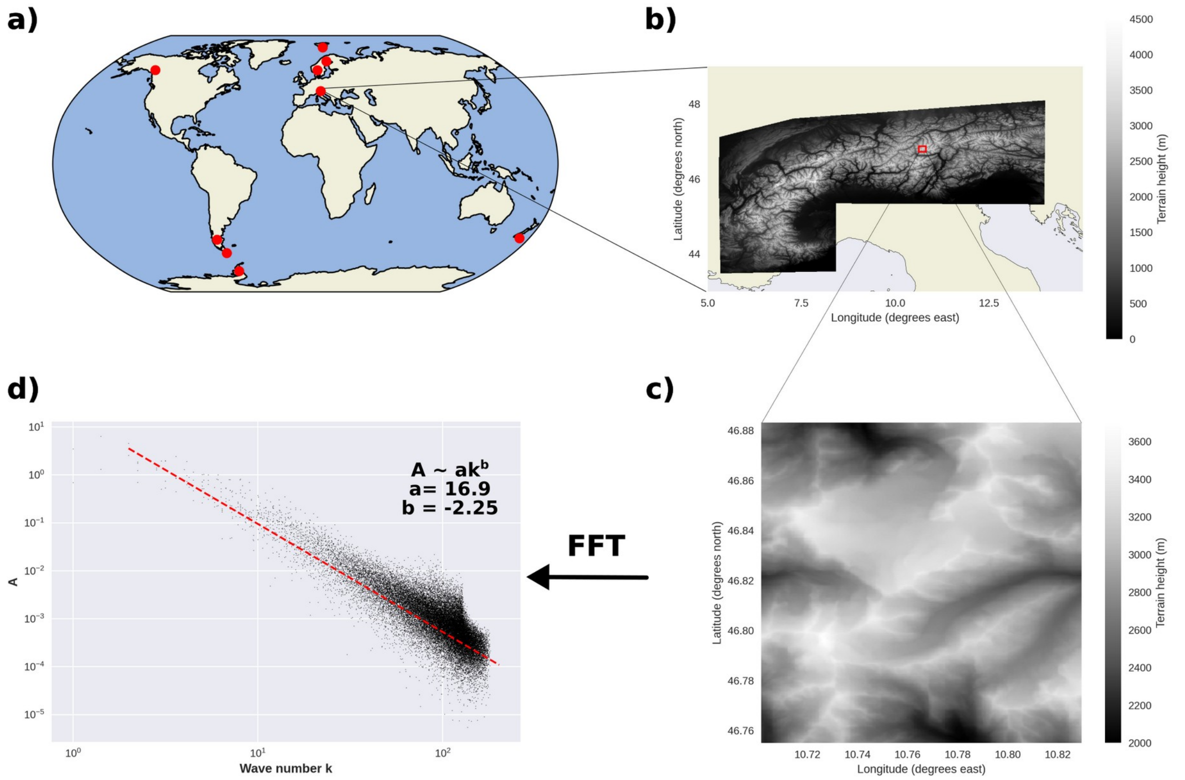
Please also note, that the output of SNOWstorm is not transferable to other spatial resolutions (see comment below).

Eq. 1: If I am not mistaken you used the 9 regions shown in Fig. 1 and cut them in 256x256 tiles with 50 m resolution. Based on the spectral parameters a, b from these topographies you derived the 72 power-law spectrally scaled topographies. Please indicate how many 256x256 tiles were used for the a, b parameter pool. Please also clearly state that the 72 resulting synthetic topographies only represent the observed topographic scaling range. Thus, training a statistical model on processes computed on these topographies does not guarantee similar statistical model accuracy on topographies that fall outside this observed scaling range. Also, isn't the wavenumber k bound by domain and grid cell size? I believe this also introduces quite some uncertainty in model applications at different domain and grid cell sizes then used for the training data, i.e. the model may not be that widely transferable. Please clarify and discuss. Please also better describe a, b, set them in context and provide the formula for k.
You are correct in how you understand our procedure.
In total we have 54022 tiles: 16416 (Alaska range), 3157 (Alps), 20272 (Antarctic Peninsula), 1485 (New Zealand), 2088 (Southern and Northern Patagonian Icefield), 1224 (Cordillera Darwin Icefield), 1488 (Scandinavia north), 1292 (Scandinavia south), 6600 (Svalbard). Across these regions we only found little difference in the distributions of the spectral slope characteristics. The main difference was between the more mountainous regions and Antarctic Peninsula and Svalbard, as these had flatter areas and smoother terrain.
You are correct, that, similar to any other statistical model, predictions for situations not represented in the training data will be problematic. In terms of spectral slope characteristics, the training data are designed to represent the natural range, in terms of real terrain slope characteristics we cannot represent the full natural range, as we have to avoid very steep slope angles for numerical stability. However, as we show in the manuscript, SNOWstorm is able (at least for the case study presented) to predict realistic fields, even with input terrain that is steeper than anything presented in the training data. Nevertheless, future users should be cautious when very steep slopes or even close-to-vertical faces are present in the domain. We stated this restriction more clearly in the limitation section.

Regarding the concerns about the domain size and resolution: SNOWstorm is specifically designed for output tiles of 256x256 points with 50 m grid spacing. For larger domains, predicting multiple tiles is necessary and a transfer to other spatial resolutions is not possible. On the one hand, this is due to the reasons in the spectral characteristics you described above, and on the other hand to account for the non-linearity in the wind field. We will state this restriction in resolution more clearly throughout the manuscript.

As stated in the general introduction, we re-worked Fig.1 into a conceptual figure on the topographic analysis procedure.

"As with any statistical model, future users of SNOWstorm are advised to be cautious when applying the model outside the range of the training data. Due to the design of the training data, this implies several limitations of the model:
– To ensure numerical stability, steep slope angles (> 40 ◦) had to be avoided in the training data. Although the case study results show consistent results even for slope angles outside the training range, future users should be cautious in cases of large steep slopes or near-vertical faces present in the domain."



“Figure 1. Overview of topographic analysis: Regions marked by red dots in a) are analyzed, tiles of 256x256 grid points are extracted form DEMs with Δx = 50 m (b–c), for each tile, a 2D-FFT is calculated, the spectrum is approximated by Eq. 1 (d).”

Fig. 1: Please indicate the reasoning for choosing exactly these nine regions as well as their terrain characteristics.

We selected these regions for a list of reasons: glaciated regions, that could be interesting for future applications; large spread over the world; dominance of large-scale forcing.

As stated above, we found the spectral slope characteristics very similar throughout the mountainous regions and also fall in line with earlier studies on regions not included in our analysis (Salvador et al. (1999): Spanish east coast; Young and Pielke (1983): Colorado, Rocky Mountains). This allows us to assume a good transferability of these characteristics to other regions.

Fig. 2 and Line 130-132: How can the Fourier Land b parameter range be broader than the b range extracted from the real topographies from which they were randomly drawn?

The range between 10^{th} and 90^{th} percentile (colored bars) is wider, however the full range (small dots) for b is for Fourier Land within the range of the observation.

Line 145: ”With the ideal-case setup of our simulations, only [...]”: What do you mean by ”ideal-case setup” here? And how does this compare to the mentioned ”semi-idealized” simulations in abstract and conclusions?

Thank you for pointing out this mismatch. This stems from the characterization of WRF simulations into real cases and ideal cases, which changes the pre-processing procedure.

We used “semi-idealized” in the rest of the manuscript, to emphasize that the simulations are run in an idealized setting, however with conditions set to reflect the range of conditions and interactions in the real world.

Line 16-163: Please detail your model setup with 10 simulations per topography:
1. One simulation describes one atmospheric condition with randomly drawn pressure, temperature and relative humidity. How you can draw random values of pressure, temperature and humidity with the ranges given in Table 1 and ensuring physical meaningfulness from a random combination at the same time?
For each simulation we draw one value of surface-level pressure, temperature and relative humidity out of the ranges presented in Table 1, which are designed to reflect the physically meaningful range for winter-time conditions in mid to high latitudes. From these values and the drawn values of static stability, and wind speed and direction, the initial profiles of wind components, temperature, and specific humidity are calculated. With this approach we want to reflect the variability seen in nature, while keeping in mind, that some combinations in nature (e.g., especially towards high stabilities) might occur less often. However, we still want to reflect these rare cases in the training data to present the maximum variability. Within the value ranges used we did not identify any combinations that would be completely meaningless.

2. When you are shifting the wind direction by 1 degree, does this mean you have 360 additional simulations for each topography or does this mean you have each wind direction twice in the 720 simulations?
The second interpretation is correct: each wind direction (full degrees) is used twice in the 720 simulations.

3. Were the same 10 atmospheric conditions run for all 72 topographies, or did every topography had their own conditions? I think it is essential to detail what atmospheric conditions are in the training data sets. Can you provide a table for the conditions if the same 10 conditions were used for each topography and if not, can you provide a histogram describing the atmospheric conditions of the data set, e.g. a barplot showing how often neutral, stable were drawn, which wind directions and also which snow density etc.
Every simulation has its own unique atmospheric conditions with a unique combination of wind speed and static stability, additionally with randomly drawn values of surface-level pressure, temperature, relative humidity, snow density, and roughness length, drawn from the range indicated in Table 1, and the wind direction as explained in the comment above.
With the 72 individual topographies used, it means that on every topography 10 different simulations with unique atmospheric conditions were performed.
We re-worked this part and show explicitly the distributions of atmospheric input settings.

"In total we run 720 individual simulations with unique atmospheric conditions. For each simulation the terrain height is defined by one of the artificial topographies described above, which means that ten simulations with different atmospheric conditions are conducted on each topography. All simulations are initialized with horizontally homogeneous profiles of potential temperature, specific humidity and the meridional and zonal wind component. These profiles are calculated from values of wind speed and direction, static stability, expressed as Brunt-Väisälä Frequency, and ground-level pressure, temperature and relative humidity. The values for ground-level pressure, temperature and humidity are taken randomly from within the ranges shown in Fig. 3c–e, reflecting the variability over winter-time mountain environments. For each simulation a unique combination of wind speed and stability (ranging from neutral to isothermal conditions) is drawn. Every full degree of wind direction is represented twice throughout the simulations.

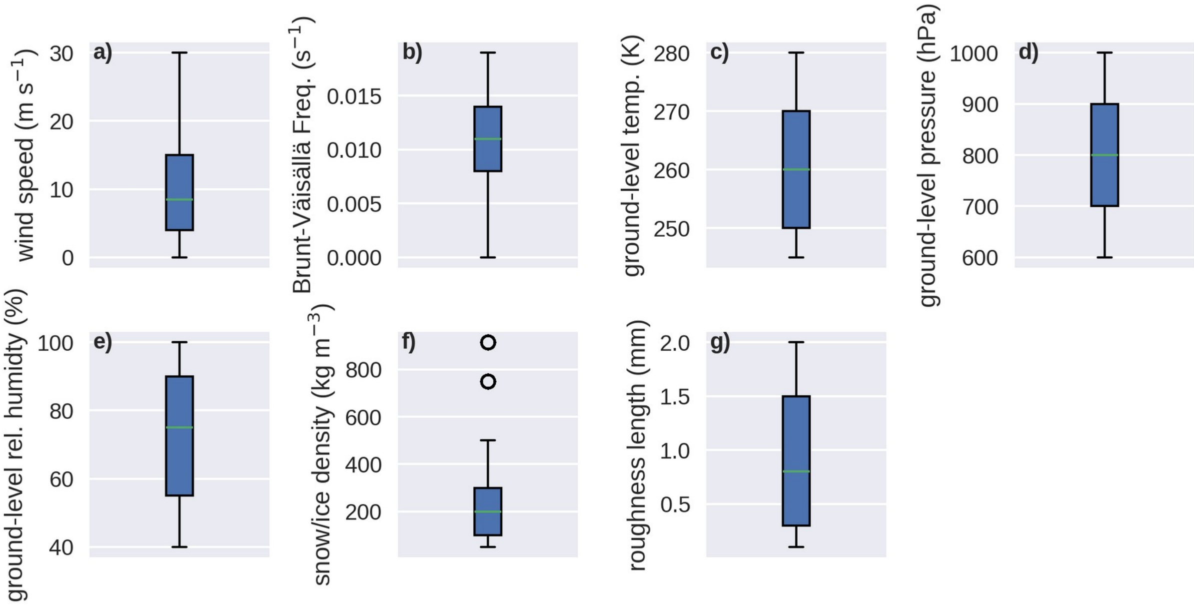


Figure 3. Distribution of atmospheric input variables and surface conditions to drive the numerical simulations."

Line 174-176: What is the reasoning and goal to average and accumulate modeled fields over 2 hours for training?

As we assume stabilized steady-state conditions for our training data, we average over two hours to smooth out potential un-steady fluctuations.

Line 203-205: Why is this application of a square-root transformation on terrain height done? Doesn't this mean resulting slopes are smoothed and the resulting topographies are not comparable anymore to the ones generated with Eq. 1? Please clarify if any implications for the topography characteristics result.

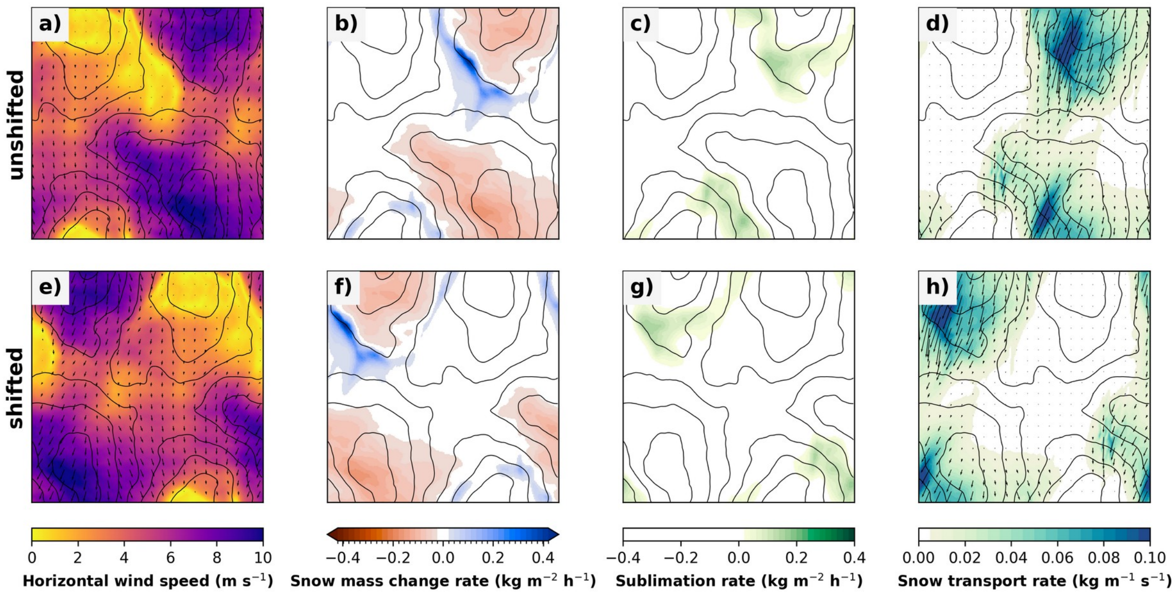
The square-root filter is applied only for the input data for SNOWstorm, not for the terrain height of the WRF simulations. Therefore, the full range of slopes is seen in the numerical simulations. We applied this filter to the input data of SNOWstorm to deal with the positive skewness in the distribution of terrain height.

Line 212-216: Following the data augmentation with one shift per topography-wind field, please indicate the total of wind fields that were used for training. If I understand correctly, the data augmentation was only applied for the snow U-Nets not for the NSW U-Net? This means the models have different amount of training data? Please clarify.

Of the 720 simulations 6/8 are used as training data for the model (540), which are used once in un-shifted and once in the shifted way, which makes 1080 individual training samples per epoch. This augmentation is applied to all U-Nets, including the NSW one.

Line 212-216: Please consider showing a few spatial examples for shifted wind fields on topography, perhaps in the Supplement?

We did that, thank you for the suggestion.



"Figure S2. Example of original (a–d) and shifted (e–h) fields of near-surface wind, snow mass change rate, sublimation rate, and vertically integrated snow transport rate for data augmentation, fields have been shifted by 120 grid points in x direction."

Line 229-223: How are the low-resolution fields interpolated to the fine-scale grid?
In the first step, the low-resolution fields are interpolated bi-linearly to the fine-scale grid. In the second step, surface-level pressure and temperature are corrected for the height difference between the low-resolution and high-resolution topography.

Line 230-231: Are predictions of DSM, SUBL VI and SNOW VI sequential? How are interactions between sublimation and snow transport considered, as it is the case in the real-world?
The predictions for DSM, SUBL_VI and SNOW_VI are sequential to the predictions of NSW but not to each other. The interaction between sublimation and snow erosion and transport is considered in the numerical simulations for the training data and thus learned by the ML model. Having for example SUBL_VI sequential to DSM would pose problems, as the sublimation not only depends, but also interacts with the snow erosion and transport. As stated in the text, we experimented with a combined U-Net to output both SUBL_VI and DSM (also using mass-conservation aware loss functions), however, in the end separating the U-Nets produced the better results.

"Apart from that, the predictions of wind surf are given as an additional predictor to the other U-Nets."

Line 252-253: To strengthen this hypothesis an overview of actual atmospheric conditions used in the training data set would be beneficial. See my comment above.
We agree and extended the section on atmospheric conditions of the training data set by the new Figure 3 (shown above).

Line 255: Compared to studies with a similar spatial resolution to that achieved here (30 m, 50 m), the phrase "fall in line" may be somewhat optimistic, as the MAE of 0.8 m/s is notably higher than the value reported by Le Toumelin et al. (2023), and the mean bias is higher than the ones reported

by Dujardin and Lehning as well as Le Toumelin. In addition, given the substantially coarser target resolution of 1 km in Dupuy et al., a direct comparison may not be entirely appropriate.
We re-worked this paragraph to be more specific. We agree, that errors in Dupuy et al. are not directly transferable and therefore removed the comment.

"*Although transferability is limited, the uncertainties reported here are in a comparable range as the ones found for other ML-based models for high-resolution winds in complex terrain. Dujardin and Lehning (2022) find for their model (trained on weather station data, high-resolution terrain descriptors, and meso-scale numerical weather model output) an MAE of 1 to 1.5 m s-1 with a negative bias between -0.16 and -0.04 m s−1 , depending on surrounding terrain characteristics. Correlation coefficients here range between 0.42 and 0.66. Le Toumelin et al. (2023) report for their ML model (trained on simulated data over synthetic topographies) a bias below 0.01 m s−1 , and a correlation coefficient of 0.96 in their cross validation experiments, though with a markedly lower MAE of 0.16 m s−1 .*"

Line 263-271: Can you set these errors in context or, if possible, compare to errors from previous studies modeling or measuring these rates, as the magnitudes of these errors alone are not intuitive.
We agree that the magnitudes are not intuitive, however, with the differences in overall settings, and validation strategies, comparison to previous studies could be highly misleading.
We therefore present the uncertainties in our testing environment as well as the error behavior with respect to other variables like the wind field.

Line 308-309: Is this validation referring to what we see in Fig. 9, i.e. green versus black lines?
Yes, this refers to Fig. 9 and the more in-depth analysis in Voordendag et al. (2024), which contains a similar figure. In the revised manuscript we made clearer that this paragraph is only a short overview on the results of the original publication.

"studied in detail by Voordendag et al. (2024), of which the main results are summarized in the following."

Section 3.2: Could you provide terrain characteristics for these three cases, similar to the background conditions described in the figures? Could it be informative to investigate three different atmospheric background conditions over the same topography? I am unsure about the origin of the mismatches: do they arise from an insufficient representation of flat terrain, or from a lack of neutral and low-wind conditions in the training dataset? This is why I suggested above providing more detailed information about the training dataset as to understand the model behavior.
Thank you for the suggestion, we provided the terrain characteristics in the figures.
It would definitely be interesting to show the model sensitivity in examples, where only one parameter is changed, in otherwise similar conditions. However, we used exemplary cases, taken from the test data set, that show the model's behavior in a wider range of different circumstances.

In our experience, one reason for the mismatches especially in weak-wind regions, is already in the training data. Many of the instances of weaker winds are in wake regions in the lee of regions of flow separation. Here the weak wind pattern are more influenced by chaotic turbulent structures, are less clearly aligned with the terrain shape and, thus, seem to be harder to predict.

Line 325-327: "are in agreement" and "slightly smoother" seem a bit overly optimistic for the spatial patterns shown in Fig. 8b,c verus Fig. 8d. Perhaps change to "agree overall well" or similar.

We agree and changed this paragraph in the revised manuscript.

"Here, the flow fields predicted by SNOWstorm driven with input from the two meso-scale domains (S_WD1_W and S_WD2_W, Tab. 1) overall agree with the HEF-LES (Fig. 9 b–d, Fig. 10)."

Line 329-330: 1-4 m/s is a rather large absolute error range. Please be more specific in saying which experiment led to better performances and please also set the errors in context to errors from previous studies. From what I see in Table 3, the correct error range is 2-4 m/s not 1-4 m/s.
Please note that the main spread is in between the stations and not in between the experiments with meso-scale input.
We provided more context of these errors, on the one hand comparing to previous studies, on the other hand comparing to the errors of the benchmark HEF-LES. As shown in Tab.4 they are in a similar range of 2-4 m/s and, therefore, can provide an estimate of the overall predictability of this event. Additionally, we give some explanation for the overestimation of wind speed as requested in the next comment. With the additional evaluation of wind direction and errors in the HEF-LES, we decided to change the table in to a figure for better readability.

"Predicted wind velocities at the three observation sites generally agree with measured overall velocities, though with absolute errors between 1.6 – 4.4 m s^{-1} and a mean overestimation of 3 – 4 m s^{-1} at the stations IHE and AWS28, while errors at STH are generally lower (Fig. 11a–b). Pointwise comparison to the HEF-LES shows an MAE of about 3.9 m s^{-1}, with negligible bias, however, localized velocity maxima in summit and ridge regions are underestimated by SNOWstorm (Fig. 9 b–d). Errors in wind direction are in line with the aforementioned underestimation of the valley channeling and the misprediction of the local flow field around IHE described above and are in the range of 70 to 100 $^\circ$ (Fig. 11c). These errors are higher than the overall errors seen in the cross validation experiments (Sec. 3.1) and for other comparable modeling approaches: in real-world settings, Le Toumelin et al. (2023) report an MAE in wind speed of 1.1 to 1.4 m s^{-1}, a bias of -0.17 to -0.24 m s^{-1}, and an MAE in wind direction of 14 to 57°. Dujardin and Lehning (2022) find for wind speed an MAE of 1.0 to 1.5 m s^{-1}, and a bias of -0.04 to -0.16 m s^{-1}, as well as an MAE in wind direction of 32 to 58°. Keeping in mind the short duration of the case study and the resulting limited transferability, these errors have to be viewed in context. Observed and simulated velocities have not been corrected for the height difference between the model output of HEF-LES and SNOWstorm (10 m agl) and the measurement heights (STH: 3.3 m, IHE: 3 m, AWS28: 2 m), which can account for roughly 1 – 2 m s^{-1} of bias. While unsurprising and in line with Le Toumelin et al. (2023), that errors in the real-world application are larger than in the semi-idealized testing environment, absolute errors in the more comparable, higher velocity classes (wind speed higher 10 m s^{-1}) in the cross validation experiments are between 1.1 and 1.5 m s^{-1} (Fig. 5 a), thus closer to the errors found in the real-world experiment here. Errors of the HEF-LES benchmark to the observations are in a similar range as for SNOWstorm, providing an estimate of the overall predictability of this event.

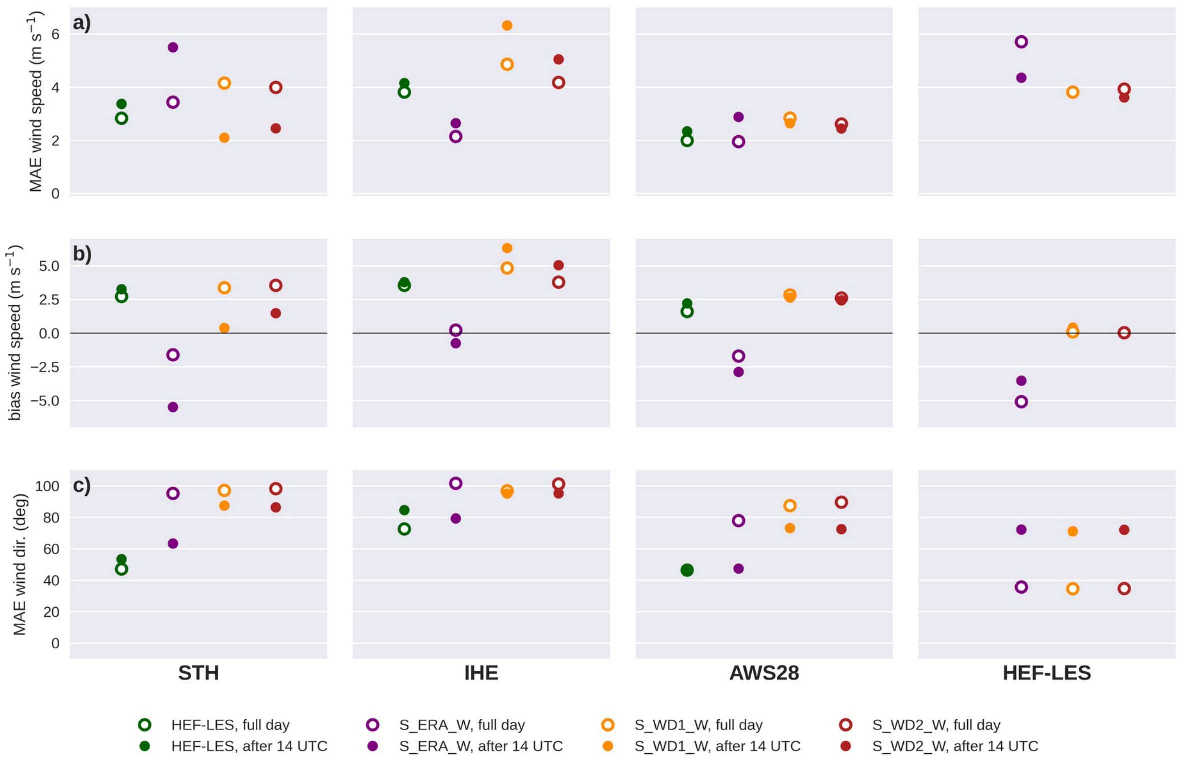


Figure 11. Error statistics for SNOWstorm experiments and HEF-LES validated against automated weather stations and point-wise against HEF-LES."

Line 330: Can you give an explanation for this overestimation?

We can try to offer some explanation:

One factor that was not mentioned in the manuscript (but will be in the revised version) is the measurement height of the observations: IHE, STH and AWS28 measure at roughly 3, 3.3 and 2 m above the ground, respectively. Additionally, the sensor heights change throughout the day, as the snow depth changes. HEF-LES wind fields and SNOWstorm predictions are at 10 m above the ground. As we do not have a prediction of near-surface stability in SNOWstorm, we did not correct for this height difference, as this would introduce more uncertainties. Assuming a neutral stratification with a logarithmic wind profile and the roughness length used in the simulations, this height difference would explain ~1-2 m/s of bias.

Table 3: Please provide the errors for ERA5 and WRF low-resolution input as well as for HEF-LES at the stations. Set in context, this would allow to better understand the errors. Please also provide the errors for wind direction, as both wind speed and wind direction, are used as input for redistribution and drifting snow sublimation in SNOWstorm.

Thank you for raising this idea. In table 3, we provide the errors of the coarse-scale input data interpolated to the coordinates of the AWS and to the grid of the HEF-LES. Additionally, we also provide the errors of the HEF-LES to the AWS for contextualization (table 4 in this document, Figure 11 in the revised manuscript). The errors of SNOWstorm are in a similar range as the errors of the HEF-LES, which shows some context on the predictability in this case study. Interestingly, based on the simple error measures presented, the errors of the interpolated low-resolution input are in a similar range or even lower than the SNOWstorm predictions and HEF-LES. This does not mean that the low-resolution fields are better in every sense, as the HEF-LES and SNOWstorm provide more information on e.g. terrain-induced flow deflections and spatial wind speed distribution.

We agree that the evaluation of wind direction errors is necessary, which is in line with comments by Reviewer 2, and we added a quantitative evaluation of the errors in wind direction.
We also expanded the process-based analysis on the representation of specific flow features.

"Errors of the coarse-scale input wind interpolated to the observation sites are in a similar range or even slightly lower than the SNOWstorm predictions and HEF-LES (see table S3), though with the missing additional information on local flow features"

Line 337-340: Can you provide error measures to demonstrate this improvement? Please consider showing the error statistics similarly to Table 3 for all "* G" experiments presented in Table 2 (e.g., in supplementary).
We provide the errors for the experiments "*G" in the supplement and also in this document (table 1)
As you can see, the errors are remarkably similar to the "*_W" experiments (table 2). We also want to apologize for the error in presentation of the ERA5 experiment on STH in the original table (values for full day and afternoon were switched).

Line 347-448: While the overall spatial patterns seem to agree well, the agreement in magnitude appears less strong. It would be helpful to be a bit more specific.
Thank you for raising this point. We added a point-wise error evaluation of the snow redistribution comparing to the HEF-LES similar to the evaluation of wind speed. We will also discuss the spatial patterns and magnitude of redistribution more thoroughly (e.g., underestimation of erosion in the summit area of Weißkugel is consistent with the underestimation of wind speed here)

"Regions of maximum erosion are slightly more localized with higher amounts in the HEF-LES, consistent with the underestimation of maximum wind velocities in the summit regions by SNOWstorm. For example, the snow erosion in the summit region of Weißkugel is underestimated by roughly 4 kg m^{-2} in SNOWstorm with meso-scale input."
"Pointwise validation against the HEF-LES shows an MAE between 2.3 and 2.8 kg m^{-2} and a correlation coefficient between 0.4 and 0.6 for the different experiments."

Line 370-373: Same comment as above in the abstract: "wide range of regions world wide", please be more clear.
We did that

"synthetic topographies and atmospheric conditions representative for glaciated mountain regions in mid- to high latitudes during winter time."

Line 373: "European Alps", please be more specific.
We did that

"on a glacier in the European Alps"

Line 375: Please rephrase as this does not fully agree with the results shown. Particularly, strengths and amounts compared less well. What do you mean with ”ground truth”? Why do you not show any evaluations with laser scan observations given that they were available for you case study (Line 306)?
Please note that the first three bullet points in this list only refer to the results of the validation experiments in the idealized training and testing environment. In this case we refer to the numerical simulations as the ground truth. We clarifyed this in the revised text.

As discussed in the beginning of our response, we cannot easily use the laser scanning data for validation, as the redistribution signal is mixed with compaction and avalanching, with additional restrictions in the scanning geometry.

Line 375: What do you mean with ”sharp gradients”? Where do you show that SNOWstorm predicts turbulent flow well and can particularly well predict atmospheric conditions that haven’t been included in existing models?
We agree, that this sentence (second bullet point) is too unspecific.
As we show in the analyses of the experiments in the training environment, the general shape and location of e.g. zones of flow separation are predicted by SNOWstorm. However, turbulent wake regions are predicted worse. We do not to claim that SNOWstom can predict certain flow situations better than other models, as this would most probably require a coordinated model intercomparison experiment.

Line 380: Please consider to be more specific here. See my comment to errors in the results section.
We changed this sentence to be more specific.

“Overall absolute errors in the cross validation experiments are about 0.8 m s −1 with a bias of -0.12 m s −1 , and increase with stronger background wind speeds. The predicted wind fields tend to be smoother than the ground truth, especially in regions of sharp gradients like zones of flow separation, or in wake regions.”

Line 380: ”SNOWstorm generally succeeds to capture the overall flow structure and redistribution patterns by the LES both with the smoothed topography from the LES and an un-smoothed high-resolution DEM.”: Again, this is general statement. Please be more specific and open. Also, I don’t see error statistics demonstrating this for the unsmoothed topography, only the spatial fields are shown in S4 and S5.
We re-worked the conclusion section to be more specific. As this is the opening sentence on the part of the conclusions focusing on the case study, we feel one sentence on the general and overall patterns is appropriate; however, we agree that the following points should be more specific.

“– SNOWstorm overall succeeds to capture the general flow structure and redistribution patterns in this case study. Features like the acceleration and deflection of the flow in summit and ridge regions are reflected by the SNOWstorm predictions, while the channeling effect inside the valley is underpredicted. The general placement of zones of snow erosion and deposition agree between SNOWstorm and HEF-LES; mismatches, like an underestimation of snow erosion in summit and ridge regions are in agreement with an underestimation of wind speed here.”

Line 385: ”Validated against automatic weather stations, errors in the wind speed are slightly higher than in the crossvalidation experiments with MAEs between about 1.5 and 4 m/s”: Please rephrase.

I don't agree that 1.6 - 4 m/s absolute errors are only ”slightly higher” than 0.8 m/s. Please also set these errors in context to errors obtained from previous ML-based models.
We agree that the increase in MAE is not only “slightly higher” and adapted the manuscript accordingly. One aspect that was not mentioned here, but is discussed more in the revised results section is that the 0.8 m/s are the value for the entire range of velocities. If we compare the errors presented here with the errors for the higher velocity classes in the cross validation experiments (Fig 4 a) the errors are much more in line.
As we want to keep the conclusion section concise and already discussed pre-existing models in the result section and the new limitations section, we only present the uncertainties of our own model here.

“Validation against automated weather stations shows absolute errors in wind speed between 1.6 and 4 m s−1 , notably higher than in the cross validation experiments. These values are partly stemming from a positive bias of SNOWstorm for several stations.”

Line 390: ”Overall amounts of snow mass change as well as the placement of zones of erosion and deposition agree between SNOWstorm and the LES.” : Please be more specific. See my other comments to this regards.
We changed the text throughout the conclusions to be more specific.

“The general placement of zones of snow erosion and deposition agree between SNOWstorm and HEF-LES; mismatches, like an underestimation of snow erosion in summit and ridge regions are in agreement with an underestimation of wind speed here.”

Line 405-408: ”semi-idealized”. Why do you call it semi-idealized? Compared to previous models that used model experiments, I would say the approach here is idealized as well, although you explicitly cover broader atmospheric stability and turbulent conditions? Please clarify and, if applicable, adapt in the manuscript.
As discussed above, we use “semi-idealized” to emphasize that the training data are designed to reflect the natural range and interactions of processes, rather than isolating individual processes in classical idealized studies of process understanding. Nevertheless, we agree that our simulations still could be considered “idealized”.

Minor comments

Line 58-61: Please verify referencing and be specific about developments and applications of wind downscaling models. As far as I know the studies, Schirmer et al used the Winstral approach, Vionnet et al and Marsh et al used wind libraries from WindNinja. WindNinja is the diagnostic model described in Wagenbrenner et al. Dadic et al 2010 used atmospheric model wind data, Helbig et al 2017 presented a statistical wind downscaling, Helbig et al, 2024 a statistical wind and snowfall downscaling approach.
We specifyed this in the revised manuscript.

“These include, e.g., various approaches for extrapolation of observed wind based on topographic 2descriptors (Winstral and Marks, 2002; Liston and Elder, 2006; Strasser et al., 2008; Schirmer et al., 2011) or wind library approaches with pre-computed wind fields from diagnostic downscaling tools like WindNinja (introduced by Wagenbrenner et al. (2016), used and adapted by Vionnet et al. (2021) and Marsh et al. (2023)), or pre-computed fields from a numerical model (Dadic et al., 2010). Statistical models building on data sets of numerical simulations under idealized conditions were introduced by Helbig et al. (2017) and Helbig et al. (2024).”

Line 70-71: "For this, Dujardin and Lehning (2022) trained their model on data from weather stations and high-resolution digital elevation models in Switzerland.": Dujardin and Lehning trained on COSMO, weather station and terrain data.
Thank you for pointing this out, we added this.

Line 196: What is meant with "above-crest height"? Please indicate which level or height above ground should be used for the low-resolution wind input.
Thank you for pointing out this mismatch. As we initialize the training data simulations with constant profiles of wind speed, direction and static stability, "above-crest height" does not make sense here.
For the real-world coupling, the default setting is for the low-resolution input to be taken from 600 hPa. However, the low-resolution model topography should not intersect this height. In settings with higher mountain ranges, the low-resolution input should therefore be taken from higher levels.

Technical comments

Line 273: Do you mean "than" instead of "that"?
No, "that" is used here as the preposition for the following relative clause.

Line 355-358: I think it should be S4 and S5 here.
Yes, that is correct (but changed with the inclusion of more figures in the supplement)

Table 1: SNOWstorm errors to AWS (GLO-30 topography):

		STH	IHE	AWS28
After 14 UTC	MAE wind speed	5.5o / 2.09 / 2.45	2.64 / 6.32 / 5.05	2.88 / 2.64 / 2.44
	Bias wind speed	-5.5 / 0.39 / 1.48	-0.75 / 6.32 / 5.05	-2.88 / 2.64 / 2.44
Full day	MAE wind speed	3.43 / 4.15 / 3.99	2.14 / 4.86 / 4.18	1.95 / 2.84 / 2.61
	Bias wind speed	-1.61 / 3.37 / 3.54	0.23 / 4.83 / 3.79	-1.7 / 2.84 / 2.61
After 14 UTC	MAE wind dir.	63.4 / 87.5 / 86.5	79.3 / 95.2 / 95.3	47.4 / 73.2 / 72.5
Full day	MAE wind dir.	95.3 / 97.2 / 98.3	101.7 / 96.9 / 101.3	78.0 / 87.4 / 89.7

Table 2: SNOWstorm errors to AWS (HEF-LES topography):

		STH	IHE	AWS28
After 14 UTC	MAE wind speed	5.98 / 1.57 / 1.96	3.80 / 3.52 / 4.43	2.46 / 3.77 / 3.40
	Bias wind speed	-5.98 / 0.32 / -0.84	-3.80 / 3.05 / 4.43	-2.28 / 3.77 / 3.40
Full day	MAE wind speed	3.11 / 3.09 / 3.08	2.55 / 4.02 / 3.93	1.69 / 4.09 / 3.71
	Bias wind speed	-2.39 / 2.52 / 1.79	-1.17 / 3.80 / 3.89	-1.37 / 4.09 / 3.71

After 14UTC	MAE wind dir.	75.9 / 81.4 / 81.5	57.0 / 93.9 / 97.5	47.0 / 70.6 / 75.4
Full day	MAE wind dir.	94.0 / 96.9 / 96.5	87.1 / 101.8 / 103.5	83.4 / 86.0 / 92.0

Table 3: Low-resolution input errors interpolated to AWS (ERA, WRFD1, WRFD2) and HEF-LES grid:

		STH	IHE	AWS28	HEF-LES
After 14UTC	MAE wind speed	8.80 / 1.90 / 1.34	7.88 / 1.50 / 2.42	5.53 / 2.43 / 4.29	8.87 / 3.36 / 2.83
	Bias wind speed	-8.80 / -1.14 / 0.68	-7.88 / 0.52 / 2.25	-5.53 / 2.36 / 4.29	-8.86 / -0.60 / 0.01
Full day	MAE wind speed	4.41 / 3.57 / 3.22	4.66 / 2.40 / 2.80	3.50 / 2.86 / 3.85	6.07 / 3.11 / 2.98
	Bias wind speed	-3.85 / 2.05 / 2.92	-4.66 / 1.72 / 2.65	-3.48 / 2.78 / 3.76	-6.00 / -0.06 / 0.39
After 14 UTC	MAE wind dir.	19.2 / 75.2 / 66.7	27.1 / 90.6 / 77.2	15.4 / 71.1 / 60.8	51.3 / 20.9 / 20.7
Full day	MAE wind dir.	37.7 / 53.7 / 71.1	53.2 / 64.9 / 68.5	34.5 / 52.2 / 71.0	62.2 / 43.5 / 43.4

Table 4: HEF-LES errors to AWS

		STH	IHE	AWS28
After 14 UTC	MAE wind speed	3.37	4.16	2.34
	Bias wind speed	3.26	3.79	2.22
Full day	MAE wind speed	2.83	3.81	1.99
	Bias wind speed	2.72	3.55	1.61
After 14 UTC	MAE wind dir.	53.5	84.6	46.5
Full day	MAE wind dir.	47.2	72.6	46.5

Response to Reviewer 2

This is a review of SNOWstorm (v1.0). Overall, I really like the idea of going after blowing snow with deep learning. Blowing snow is critical to include, and it's very tricky. However, I think this manuscript needs more work and contextualization to be a strong contribution.

I found that the manuscript layout to be difficult to follow, with the mix of results and discussion together making for a difficult read to really figure out what was tested and where. For example, the conclusion notes that direction and speed were tested against observations, but I'm struggling to find this in the text. The heavy use of model variables, e.g., SUBL_VI, SNOW_VI, makes parsing the paragraphs difficult and makes it easy to get lost. I would encourage the authors to have clear numerical model methodology, the tests, and the case study results, with a synthesis discussion section to tie them together. And then use this to tighten the text as much as possible.

Broadly, I am left with an uncertainty as to how good this model really is. I find both the wind velocity and blowing snow analysis underwhelming. Specifically, I would really like to see figures showing model speed and direction, compared to both a model but also observations. I see figure 9, but I would prefer this as a wind rose and to see >24h of observation analysis. How are the ridges and lees /exactly/ compared ot the more complex model or obs? The lee is where the snow deposition will happen and if this is wrong, the deposition is wrong (which seems like the case). Can the model perturb the wind direction very much? It looks like it barely does, but this might just be a visualisation limitation. How large of a domain can this scale up to? How does this approach compare to other blowing snow approaches likes those of Snowmodel, ICAR, CHM, SnowPappus, and Alpine3D? For example, Figure 5 d-e shows a ~4x under estimation in transport rate. Figure 6 is completely missing deposition zones. This seems like a lot. I would also like to see side by side plots of the lidar snowdepth observations and the model outputs. Is the spatial variability we observe, that all models struggle to get 100% right, at least qualitatively present in the model outputs? Without these contextualizations, it is difficult to evaluate the manuscript.

I think having a core set of research questions that are answered throughout the manuscript would help tighten and refine the story. As is, this is a statement of what was done, and the clear common thread tying it all together, including the advancement versus existing approaches is lacking. Getting the wrong answer faster is not particularly useful unless the error in that approximation is well bounded, understood, and justified.

We want to thank the reviewer for their valuable feedback. As stated in the general response above, our revision concentrated on a more thorough presentation of model uncertainties and contextualization in the framework of existing models, as well as on a modified structure that aims to improve readability of the manuscript. This included a new section on model limitations.

Please find our answers below to your specific points raised in the upper paragraphs :

First, we changed the abbreviations for the model output to more intuitive ones (NSW → $wind_{surf}$; DSW → ΔM_{snow}; SUBL_VI → ΔM_{subl} ; SNOW_VI → f_{snow}). For the abbreviations of the individual experiments in the case study, we used the description of the experiment setup wherever possible, and repeat the explanation of the abbreviation otherwise.

We agree that a real-world validation for longer than 24 h would be desirable. Sadly, the HEF-LES as one part of the benchmark data only exists for this period and this data basis is not easily extendable due to the large computational demands. We therefore concentrate on this rather short case study, in which we can still learn a lot about the model's behavior in various flow scenarios as we can validate against the highly resolved LES as benchmark.

We want to argue against the change to wind roses in Figure 9, as we would lose the temporal variability information, which is one important aspect in our analysis of the case study. As stated in the introduction of our response and in the specific comment below on the evaluation of wind direction, we agree that a more thorough analysis of the wind direction is necessary, which is added in the revised manuscript.

We agree, that a direct comparison of the simulated snow redistribution with the observed snow height change could strengthen the validation. We can unfortunately not do this here, as the observed snow height change is a complex signal to interpret. It is not only due to redistribution but confounded by compaction and avalanching, and additional restrictions in the scanning geometry. Disentangling these effects would be out of scope of this work, and has been the focus of dedicated studies like Voordendag et al. (2024). However, the HEF-LES data were validated against the laser scans in the original publication and we use these validated HEF-LES as benchmark in our study. In the section 3.3.1 (case study overview) we emphasize this validation approach more clearly. Apart from that we added a quantitative validation of snow redistribution with respect to the HEF-LES.

"How large of a domain can this scale up to?" The output of SNOWstorm is fixed to tiles of 256x256 points with spatial resolution of 50 m. For larger domains, multiple tiles have to be predicted, and in this way any domain sizes are possible. Note that the model is not transferable to other spatial resolutions (see also response to Reviewer 1). We point this out more clearly in the revised manuscript.

Unfortunately, we could not follow some remarks of the reviewer, in particular we do not see the missing deposition zones in Fig. 6 (blue shaded areas), which the reviewer is referring to.

Specific comments:

L1 "and (the) resulting" missing the

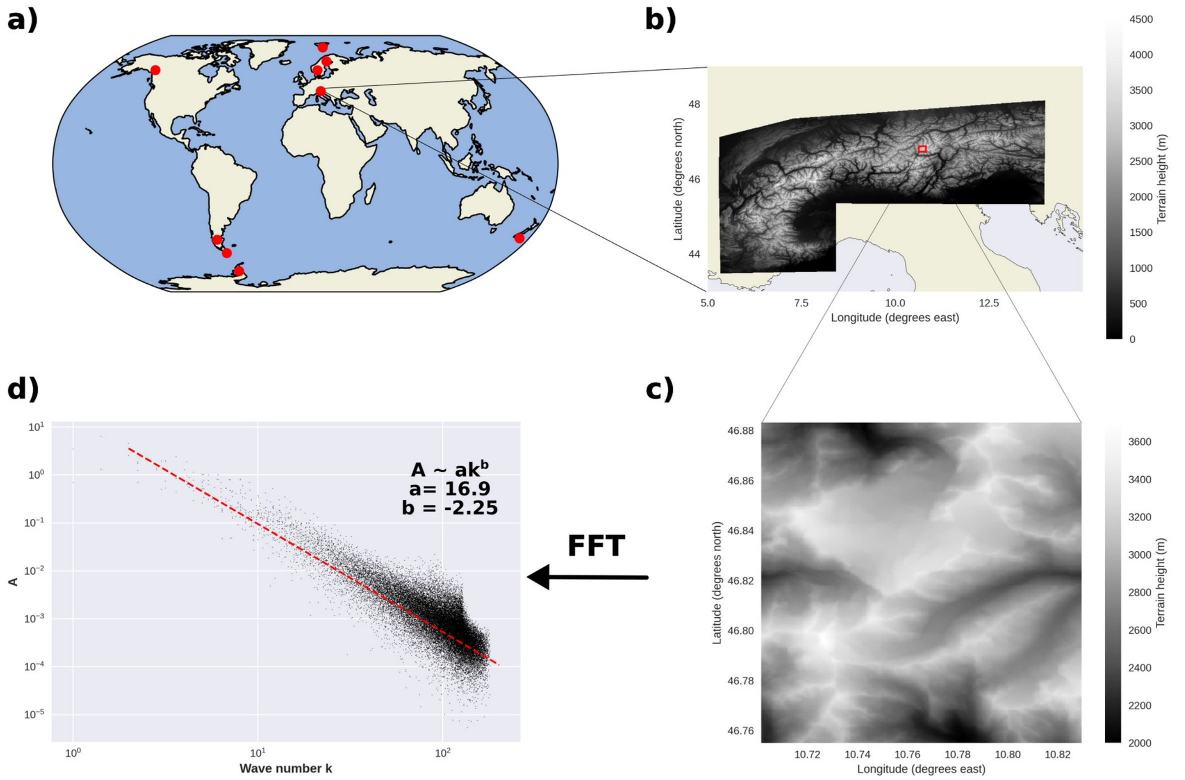
We added that.

L22 e.g., add the comma after e.g. throughout

We added that.

L 91, Figure1 these need scale bars, north/south arrows, and lat/lon reticules. As is, it's not at all clear what we are looking at.

We agree, and we re-worked this figure into a conceptional figure to showcase the topographic analysis procedure.



“Figure 1. Overview of topographic analysis: Regions marked by red dots in a) are analyzed, tiles of 256x256 grid points are extracted form DEMs with Δx = 50 m (b–c), for each tile, a 2D-FFT is calculated, the spectrum is approximated by Eq. 1 (d).”

L95 The synthetic topographies are neat, but why go through the effort when global 30m DEM exist? I suspect it's for the periodic boundary conditions? If so, this needs to be stated up front and not as a buried lede

As described in L170-172, the periodic nature of these is a useful feature for our application, though not the main reason for working with this approach. It is correct that high-resolution DEMs are globally available and that building a training data set based on real terrain would be possible as well, however, then the problem would be how to choose individual regions and how to avoid sampling bias.

The main idea is for our model to learn the fundamental behavior of terrain-atmosphere interaction. For this reason, we build the training data set on synthetic topographies that reflect fundamental terrain characteristics.

L 114 "to fit the requirements of the later model" where in the MS? what section?

This refers to the design choices in model resolution and domain extend of SNOWstorm. We will make this more clear in the revised manuscript.

“in order to fit the requirements in terms of domain size and resolution of the later model”

L120 `a` term is what?

Thank you for pointing this out, “a” is the intercept of the spectral slope, we will add this in the revised manuscript. Additionally, it is shown in the re-worked Fig.1.

L145 "is used" -> "are used"

We corrected this.

L155 Why are the particle sizes not a distribution as commonly noted in the literature?

In the current version of the snow drift WRF module the particle size is described for simplicity as the mean particle size at ground level with decreasing size with height, the size at ground level is either prescribed or calculated as a function of friction velocity.

L155 Is this domain masked for vegetation? how is that handled? I realize that the authors are interested in non-vegetated areas, but some description of how this is handled and impact on flow fields is needed
Yes, the entire model domain is set up without vegetation. As we run the numerical simulation in ideal-case setup with no land-surface model, this is reflected in the prescribed roughness length. It is correct that this is not a problem in our application, as we are interested in non-vegetated areas, but it is a limitation for other applications. We stated this more clearly here and in the additional section on model limitations.

“In our simulations, the entire domain is assumed to be uniformly snow covered, and free of vegetation.”

“– The domains in the training data are designed fully snow-covered and free of vegetation, representative for high-alpine winter-time environments. Vegetated areas can therefore not be appropriately simulated by SNOWstorm.”

L 156 table 1, how are these sampled while maintaining physical coherence? Please cite ranges for these values

As the values of Tab.1 are chosen to reflect the variability seen in nature, we did not identify any parameter combination as incoherent. Nevertheless, some parameter combinations (especially with very strong static stability) might occur less frequent. We still want to include these settings in the training data, as we want to expose the model to a training data set that is as variable as possible.

L174 "model run time" are these wall clocks or internal model hours?

We refer here to internal model hours, we changed this in the manuscript.

L229 how is T changed but RH is not updated? this will dry/wet the atmosphere out and impact the sublimation estimates

We are convinced that this should not be a problem. In this step, we extrapolate the temperature from the coarse-scale input to the height difference between the coarse-scale model topography and the fine-scale DEM, under the assumption of a moist-adiabatically stratified atmosphere of constant relative humidity. You are correct, that this will change the absolute humidity, however, the sublimation calculation in SNOWstrom is dependent on the relative humidity.

L236 I would like to see clear wind direction eval here to

As stated in the general introduction, we added a point-wise evaluation of the wind direction error compared to the AWS and HEF-LES. Additionally, we will ed a more process-based evaluation of how flow features are spatially captured. This will include the flow deflection and flow splitting on the windward side of Weißkugel summit, the flow deflection and acceleration at the ridges north and south of the glacier, and the channeling inside the valley.

As you pointed out in your general comments, it is crucial for the model to represent the wind direction at ridges to correctly to identify wind-ward and lee-ward slopes correctly. This is happening largely throughout the domain, except for the area around station IHE on the ridge south of the glacier. Here, the ridge is almost aligned with the large-scale flow direction and therefore both HEF-LES and SNOWstorm fail to predict the observed southerly wind direction. We discuss this more critically in the revised manuscript.

"General flow features, such as the flow deflection and splitting on the windward side of Weißkugel summit (see Fig. 8), the deflection on the ridge north of the glacier, as well as the acceleration of the flow in summit and ridge regions, are represented in the SNOWstorm predictions (Fig. 9 b–d). The channeling effect in the valley is underrepresented by SNOWstorm, leading to a too strong westerly component here (Fig. 9 b–d), which is also evident in the comparison to the two observation sites of Station Hintereis (STH, Fig. 10 b) and the temporary station on the glacier (AWS28, Fig. 10 f). With the ridgeline of Rofenberg south of the glacier (see Fig. 8) being aligned almost parallel to the ambient westerly flow, both the HEF-LES and SNOWstorm fail to reproduce the local flow field here. HEF-LES and SNOWstorm simulate a south-westerly and westerly flow on this ridge, and thus fail to reproduce the southerly flow and therefore the main direction of overflow (Fig. 9 b–d, Fig. 10 d)."

L308 has results in the overview

This is a result of the original publication of Voordendag et al. (2024), for which we are giving the overview here. In the revised manuscript we stated more clearly, that this paragraph only gives a brief overview on the analysis of the original publication.

L320 how sensitive are the results to this light-ish density? Windslab layers are of course also missing here

We ran a sensitivity experiment with higher density of 400 kg m-3. In this experiment the average erosion is reduced by roughly a factor 3, the maximum erosion by a factor 1.4. With the fresh snowfall immediately before the drifting snow phase, we consider the 200 kg m-3 realistic. As this density is also present in the HEF-LES, we also use it in our experiments for consistency.
You are correct, that snow surface hardening and densification is not included in our approach, and that this is an important aspect to include. In our experiments, we wanted to be consistent with the HEF-LES, where it is not included either. However, for future application with SNOWstorm coupled to a snow model, this aspect will be included.

L330-> would love to have S_WDI_W and other variables described or reiterated for the reader. By the time I got here, I had no idea what these were anymore. There is some description on L 325 but just they are inputs, of what?

Thank you for pointing out these difficulties, we re-introduced variable names for better readability. The structure of the abbreviation in your example is:
S → SNOWstorm experiment
WD1 → atmospheric input for the SNOWstorm prediction taken from WRF domain D1
W → SNOWstorm run on smoothed topography of HEF-LES