Second review of "SNOWstorm (v1.0) - a deep-learning based model for near-surface winds and drifting snow in mountain environments", by Saigger et al.

General

This is my second review of this manuscript. Following the first revision, the authors have improved the manuscript considerably. However, below I list several issues that still need to be addressed, along with a few new ones that have arisen as a result of the additional explanations. Please ensure that the explanations provided to the reviewers are also incorporated into the manuscript, as this has not been done consistently in the revised version.

Specific concerns

Introduction

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. Answer: 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."

Le Toumelin et al. (2023) trained Devine on a dataset consisting of ARPS simulations under idealized flow conditions with turbulent flow intentionally suppressed. In that context, the linear scaling with background wind is consistent with the assumptions of the training data and intended. Your mentioned "shortcomings" for Devine, refer to potential limitations when extrapolating beyond Devine's intended regime of applicability. This is meant as clarification here, I leave it up to you what to do with this.

Data and methods:

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. *Answer:*
"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 realworld 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."
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).
I'm not fully agreeing with the rewritten paragraph. The large ensemble of synthetic topographies used for training a wind downscaling model in Helbig et al., 2017 and Le Toumelin et al., 2023 covers a very wide, diverse range of terrain characteristics in order to describe the local interactions of large-scale wind flow with topography, i.e. terrain-induced flow patterns independent of a geographic region. How does your above paragraph answer the raised concern of choosing power laws for topographies that do not hold across spatial scales over Gaussian topographies? I am assuming the comment that SNOWstorm is not transferable to other spatial resolutions was meant to respond to that. Please be specific and also include this explanation in the manuscript (I didn't see it there but maybe I missed it).
Secondly, the reference to covering the whole range of static stability doesn't seem to be fully justified, given that your training data covers mainly stable conditions with marginally extending the conditions to neutral cases (now clear from the new Fig. 3b): "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 clarify this in the manuscript.

Fig. 3: The caption of Fig. 3 is not clear on its own. If I understand correctly the shown

distribution covers the possible range from which actual values are randomly drawn. Please adapt: "Distribution of atmospheric input variables and surface conditions to drive the numerical simulations."

Fig. 3: While I appreciate the newly added Fig. 3, showing the distribution range of input parameters from which the parameter values are randomly drawn, I'm still convinced that it is key to understanding the SNOWstorm model performance and its limitations to know which parameter combinations are actually used during training, e.g. how often was a neutral atmosphere combined with low or high wind etc. Please indicate this additionally in Fig. 3 or in a new figure and incorporate it in your discussion and limitation section.

Line 174-176: What is the reasoning and goal to average and accumulate modeled fields over 2 hours for training?
Answer: As we assume stabilized steady-state conditions for our training data, we average over two hours to smooth out potential un-steady fluctuations.
Does this temporal average reduce turbulence in the training dataset? Please indicate.

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.
Answer: 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.
I cannot understand what this means "deal with the positive skewness in the distribution of terrain height". Please clarify.

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.
Answer: 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 in-

fluenced by chaotic turbulent structures, are less clearly aligned with the terrain shape and, thus, seem to be harder to predict.

I am a bit confused from your answer. Above you explicitly say one project goal is to include large turbulent structures in the flow (likely why you are using WRF in LES mode), which is different from earlier model development approaches. However, later you mention you averaged the wind flow training data over two hours and here you say chaotic turbulent structures for weak events are harder to predict. Can you please clearly indicate which turbulent processes are still accounted for in your training wind dataset after applying the preprocessing etc. Furthermore, for which cases do you expect improved performances with SNOWstorm compared to the earlier model approaches? Can you explicitly demonstrate that SNOWstorm captures a case study experiencing turbulent flow conditions, similarly as done for Fig. 6 and 7? This is highly relevant to understand potential applicability of your model in contrast to the existing wind downscaling models and also in view of your mentioned project goal.

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.

Answer: 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 (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 realworld 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 cor-

rected 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.
I'm curious to see if height corrected wind speed and your estimated bias reduction of 1-2 m/s would alter overall bias errors. Please provide the resulting errors on the height corrected wind speed and clarify. Also, the comment on that prediction for larger velocity classes are performing better is vague. Please be more specific, is this related to missing terrain impact description in the model, to turbulent flow or to the parameter range of the actually used training data (Fig. 3 versus the actually covered range)? This is also related to my previous comment. Where do you see an impact/advantage of using LES mode data for training?

Line 330: Can you give an explanation for this overestimation?
Answer: 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. Please provide the resulting errors including the height correction. Given that the missing height correction is now provided as the main reason for the larger errors compared to previous studies (see also your summary and limitations section), I would like to see the resulting errors after corrections. Your model is trained on predominantly stable winter conditions (Fig. 3, albeit the combined randomly selected parameter space is unclear), could this be a reason? Also, please check whether the studies you are using to contextualize the errors have applied that height correction (if I am not mistaken, they actually do not apply it either).

Conclusions and Outlook
Line 465: "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 higher values are partly stemming from a positive bias of SNOWstorm for several stations." Please clarify what you mean with this explanation. Why is there a positive bias?