Reviewer 1

This manuscript presents a deep learning-based statistical downscaling model for surface winds over complex terrain. The model consists of two parts, correction of winds from a regional atmospheric model, and conversion from a coarser grid to a finer grid, based on information of high-resolution topography data and atmospheric conditions. Results indicate that the proposed model better represent winds over Western Alps, for which the model is trained. The manuscript is generally well written, and conclusions are clear. In particular, analyses on the explainability increases its potential for real applications, where reliability of the model matters. There are, however, still room for improvement in the presentation quality. Therefore, my recommendation is publication after minor revisions are made.

We thank the Reviewer for the time dedicated to the review and for his constructive comments that we believe help improve the manuscript. Please find below our answers to all of your comments.

Minor comments:

1. I understand that the model names, such as AROME and ARPS, are well known in the community, but it would be better to present their full names somewhere in the manuscript, such as the lines at which they appear for the first time or a list of names at the end of manuscript.

AROME and ARPS acronyms are now fully explained at their first occurrence in the manuscript.

As an illustration, Le Tourmelin et al. (2022) observed that AROME (the NWP "Application of Research to Operations at MesoscalE" operated by Meteo-France) wind fields are frequently underestimated at elevated and exposed areas.

More specifically this model was trained at replicating the behavior of the atmospheric model ARPS ("Advanced Regional Prediction System") over complex Gaussian topographies Helbig et al. (2017).

2. What is the source of high-resolution topography data?

The digital elevation model used in this manuscript is a combination of DEM from different sources, used for current studies at our research lab (Snow Research Center - CEN/CNRM/CNRS). Within France boundaries, the “RGE ALTI” DEM |[IGN], i.e. a 5m grid-spaced DEM from the national cartography institute, is resampled to 30m and used. Outside of France, the “GLO-30” DEM, i.e. the 30m grid-spaced Copernicus DEM, is used (Fahrland et al. 2020). Both DEM have been previously processed to be assembled in a coherent way.

The following text has been added:

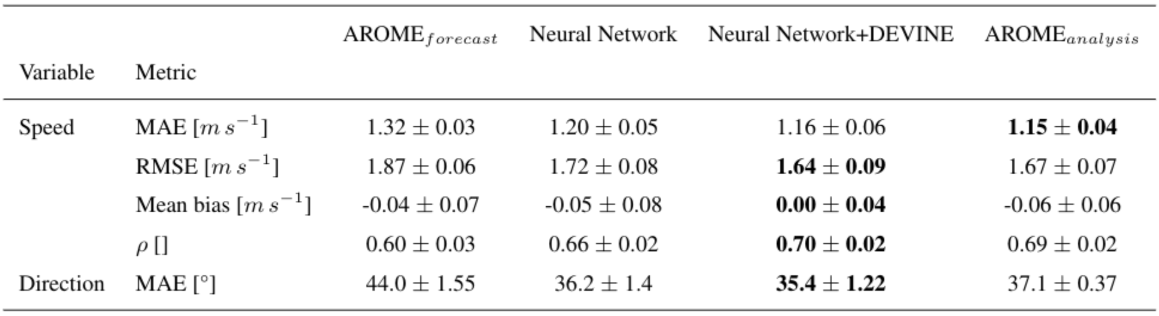
In this study, the DEM used has been obtained after merging RGE Alti DEM resampled to 30m (IGN) inside of France boundaries and GLO-30 DEM in Switzerland (Fahrland et al. 2020).

3. In this manuscript, the data are divided into training and test datasets. However, in many practices, people divide data into three sets: training, validation, and test, where the validation dataset is used to tune hyperparameters. How did you tune the hyperparameters?

Initially, we started the study using a validation dataset composed of a few stations in order to properly calibrate hyperparameters. However, we found that the results on the validation dataset were highly sensitive to the selected validation stations. This is explained because wind conditions and associated metrics are highly variable in mountainous terrain. Even though we did not stick to this evaluation procedure, this step enabled us to derive a first sketch of the selected architecture.

Another procedure could have consisted in selecting hyperparameters using cross validation. Given the fact that the derivation of the selected architecture required hundreds to thousands of numerical experiences, each experience requiring several hours of computations, multiplying the number of training instances to compute cross validation would not have been possible with the resources at our disposal. These many experiences are explained because we did not use a standard deep learning architecture but built a rather more complex data flow that required much experimentation.

In the end, we kept the hyperparameters both inherited from the first experiments using a validation dataset and from the next experiments using only a test dataset. When the architecture and hyperparameters were fixed, we ensured that we did not overfit our network to the test database by training two alternative models with the same architecture and inputs data on two alternative train sets. For each alternative train set, alternative test sets were also obtained after selecting new test stations with the same selection procedure as described in the article. The average three-fold evaluation is presented below:



		AROME$_{forecast}$	Neural Network	Neural Network+DEVINE	AROME$_{analysis}$
Variable	Metric				
Speed	MAE [$m\,s^{-1}$]	1.32 ± 0.03	1.20 ± 0.05	1.16 ± 0.06	$\mathbf{1.15 \pm 0.04}$
	RMSE [$m\,s^{-1}$]	1.87 ± 0.06	1.72 ± 0.08	$\mathbf{1.64 \pm 0.09}$	1.67 ± 0.07
	Mean bias [$m\,s^{-1}$]	-0.04 ± 0.07	-0.05 ± 0.08	$\mathbf{0.00 \pm 0.04}$	-0.06 ± 0.06
	ρ []	0.60 ± 0.03	0.66 ± 0.02	$\mathbf{0.70 \pm 0.02}$	0.69 ± 0.02
Direction	MAE [°]	44.0 ± 1.55	36.2 ± 1.4	$\mathbf{35.4 \pm 1.22}$	37.1 ± 0.37

Notations in this Table follow the specifications from Table 3 in the main manuscript.

This choice is summarized in the main manuscript as follows:

Diverse architectures and hyperparameters were tested in order to converge to the final model. We checked that our model doesn't overfit the test set by computing metrics using a three-folds cross-validation strategy, presented in Table S1 in the supplementary material.

4. Figure 4 and 5: Labels are too small. Please enlarge them.

   The labels are now bigger.

5. Figure 6: In the main text, the bottom panel (Fig. 6b) appears first, and the top panel (Fig. 6a) later. Because this is a bit confusing, I would suggest swapping the panels. In addition, I would suggest citing the figures in the main text as Fig. 6a and Fig. 6b by explicitly indicating the panel ID.

   We followed your advice and updated the text and figures accordingly.

6. Figure 7: Please enlarge the labels. The lines are difficult to distinguish. You can use different line types, such as dashed or dotted.

   We updated the graphics accordingly.

7. Why does this study start downscaling at the forecast lead time of 6 h? I am curious how these models perform at shorter lead times. In other words, do you assume the same model error in AROME_forecast for the lead times from 6 to 29 hours?

   We downloaded modeled data using the same procedure as in Vionnet et al. (2016), Quéno et al. (2016), Le Toumelin et al. (2022) and similarly to Gouttevin et al. 2022. In more detail, we only extracted lead times between +6 and +29h, issued from the 00:00 analysis. This allows us to reconstruct continuous atmospheric forcings over the study period, as typically found in the forcing files of snow models such as CROCUS, that are commonly processed at our lab. The choice of +6h was originally selected in Vionnet et al. (2016) as a way to minimize the effect of the analysis at 00:00 on the simulations. Here, we adopted the same procedure in order to have corrected and downscaled simulations comparable to the previously used set of data. This offers the possibility to update previously used forcing files including AROME data in the most consistent way.

   However, this strategy does not assume that errors between lead time +6h and lead time +29h are the same. Input variables used by the model allow the neural network to access information correlated to the time of the day, which in our case also corresponds to a given lead time. This way, the correction can take into account varying performances following different lead times. We tested to include the lead time among the input variables but did not observe any improvements. Constructing an architecture that would correct shorter lead time (<6) or larger lead times (> 29) or other analysis cycles (different from 00:00 UTC) would probably require training different models, or informing the models about the lead times/analysis cycles used.

   We added the following modifications to the main manuscript:

   “Firstly, we built a continuous time series by extracting +6 to +29h AROME forecast lead times, initialized with the analysis of 00:00 UTC, as in Quéno et al. (2016); Vionnet et al. (2016); Le Toumelin et al. (2022). This was done as a way to obtain continuous time series, typically used to force snow and surface models as in Quéno et al. (2016); Vionnet et al. (2016); Gouttevin et al. (2023).”

8. Related to the question above, have you applied “Neural Network+DEVINE” to AROME_analysis? “AROME_analysis+Neural Network+DEVINE” may serve as a good analysis product.

   We didn’t apply “Neural Network+DEVINE” to “AROME_analysis“ because we didn’t have more data than used in the test set when it concerns AROME_analysis. With the storage system hosting modeled data at our disposal, it requires months of continuous downloading to obtain years of model outputs. As a consequence we can not perform additional training within a short period of time. However, we fully agree with the reviewer that training our model using AROME_analysis as inputs would be of high interest for the community.

9. Figure 9: The color bars for (d) and (e) have no labels for negative values.

   The label is now incorporated.

10. Page 25, line 3: A link to a reference is broken.

    The link is now corrected.

11. Section 5.2: In this section, each paragraph looks very long. I would suggest splitting the paragraphs for readability.

We thank the reviewer for this advice and agree that this section was hard to read as a single uniform paragraph. We splitted the long paragraph into multiple paragraphs to improve readability.

12. Figure 10 and 11: The yellow shadings and lines are difficult to read.

    We changed the contrast and modified the color to a darker yellow to make the shadings and the line more visible.

13. Figure 10: ALE increases as the wind speed increases at 515, 126, and 10 m. In contrast, ALE is neutral for wind speed at 50 m, and ALE decreases as wind speed at 5 m increases. Do you have any interpretation on this behavior?

    By construction, the ANN_speed computes a correction that is added to 10m AROME wind speed. Consequently, we understand that model outputs will increase or decrease in phase with this variable, which is confirmed by the ALE plot.

    Then, ALE plots for wind speeds from other atmospheric levels give us insights on how input values relate to the output values. This is informative to understand how decisions are made inside the model but not necessarily why decisions are made or if they relate to physical processes.

    We note that AROME wind fields result from numerical computations that account for processes from the free atmosphere down to the surface. It is very probable that errors observed in the simulations of 10m wind fields result from modeling error originating in different parts of the code. For example, inaccurate parametrizations of atmosphere surface interaction can translate to errors in low level wind fields (e.g.

5m wind fields). Similarly, erroneous estimations of higher altitudes wind fields can lead to errors in surface wind estimations. We hypothesize that these types of behavior might be (partly) integrated in the decision making of the model regarding wind variables.

14. Page 26, line 558-: Although this paragraph states that input variables perhaps have large interactions, in the following sentence, it is also stated that input variables are not correlated, which is a bit confusing. Could you clarify this point?

    Interactions between two variables define how a specific set of values are combined in order to determine the output of the model. In other words, the variables interact one with each other within the model in order to create the output. On the other hand, correlated variables are variables that can be linearly related. These two concepts are independent.

    To illustrate this point, we can hypothesize that a model downscaling wind fields depends on large scale wind direction and atmospheric stability. For a given large scale direction, the model might present various outputs given the air stability: the feature “large scale wind direction” interacts with the feature “air stability” to produce the model output. However, large scale wind directions and air stability can be completely independent (i.e. not correlated): given a fixed large scale wind direction, any type of air stability can be encountered in the input data.

    We now specify in the text that the interaction between variables designates interaction within the model:

    Input variables of ANN_direction present scattered individual effects, probably evidencing large interactions among input variables within the model when computing the output, as visible on PDP (Fig. 11).

15. There are many typos and grammatical errors. Please doublecheck.

We apologize for these errors. We carefully read the article and removed all the observed errors.

References:

Fahrland, E., Jacob, P., Schrader, H., & Kahabka, H. (2020). Copernicus digital elevation model—Product handbook. *Airbus Defence and Space—Intelligence: Potsdam, Germany*.

IGN: RGE ALTI ® Version 2.0, https://geoservices.ign.fr/sites/default/files/2021-07/DC_RGEALTI_2-0.pdf.

Quéno, L., Vionnet V. , Dombrowski-Etchevers I. , Lafaysse M. , Dumont M. , and Karbou F. , 2016: Snowpack modelling in the Pyrenees driven by kilometric-resolution meteorological forecasts. *Cryosphere*, **10**, 1571–1589, doi:10.5194/tc-10-1571-2016.

Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Quéno, L., Seity, Y., & Bazile, E. (2016). Numerical weather forecasts at kilometer scale in the French Alps: Evaluation and application for snowpack modeling. *Journal of Hydrometeorology*, *17*(10), 2591-2614.

Gouttevin, I., Vionnet, V., Seity, Y., Boone, A., Lafaysse, M., Deliot, Y., & Merzisen, H. (2023). To the origin of a wintertime screen-level temperature bias at high altitude in a kilometric NWP model. *Journal of Hydrometeorology*, *24*(1), 53-71.

Le Toumelin, L., Gouttevin, I., Helbig, N., Galiez, C., Roux, M., & Karbou, F. (2023). Emulating the Adaptation of Wind Fields to Complex Terrain with Deep Learning. *Artificial Intelligence for the Earth Systems*, *2*(1), e220034.