REVIEWER 3

MAJOR COMMENTS:

1. *One major missing element from the manuscript is the discussion of elevation mismatch between the coarse reanalysis gridcells and stations. Both in terms of evaluation, because from all previous studies it emerges that if one accounts for these differences, errors drop considerably. But also for the RF, it could be a key input variable.*

   Answer:

   Elevation mismatch is certainly the primary factor contributing to systematic error of snow depth over complex orography. If in Fig. 4a-b, the mean bias is plotted against the elevation difference instead of absolute elevation value, there will be a distinct linear relationship between them (see Fig. 3 below). However, bias changes during the season (increases with snow accumulation), so trend parameters derived from the plots would not be relevant to calculate a daily correction. Hence, apart from elevation mismatch, other factors (e.g., absolute value of snow depth, accumulated precipitation since the beginning of the season) should also be explanatory. It is definitely more common in literature to account on elevation mismatch when correcting bias of air temperature using lapse rate since the method is very simple and the parameter is instantaneous (Bouallègue et al., 2023; Keller et al., 2021). When it comes to snow fields, it is common by dynamical downscaling to apply lapse-rate-based corrections to fields that determinates snow at most, i.e. air temperature and precipitation (Baba et al., 2021; Dalla Torre et al., 2024). However, we are not aware of any publication where a snow field would be corrected using some linear relation based on elevation mismatch. We can imagine that it could possibly reduce the error in complex terrain to some extent, however, please notice that even for stations with little elevation difference, there are still non-negligible systematic errors. An example of such site is Puczniew (Fig. 4), a climate station which lies in central Poland in nearly flat terrain, with only 3 metres of elevation difference against ERA5 and a perfect match in case of ERA5-Land. The station lies around 20 km away from a synop station which probably was assimilated and therefore the mean bias for the atmospheric reanalysis is relatively low. However, RF is still able to make it better.

   Summing it up, the approach we used is to tackle the systematic error in total, regardless weather it was predominantly driven by elevation mismatch, simplifications in parametrizations of snow physical processes or any other factor (although some part of the error could be reduced with some simple linear method). Such an approach is common when correcting snow bias using ML methods.

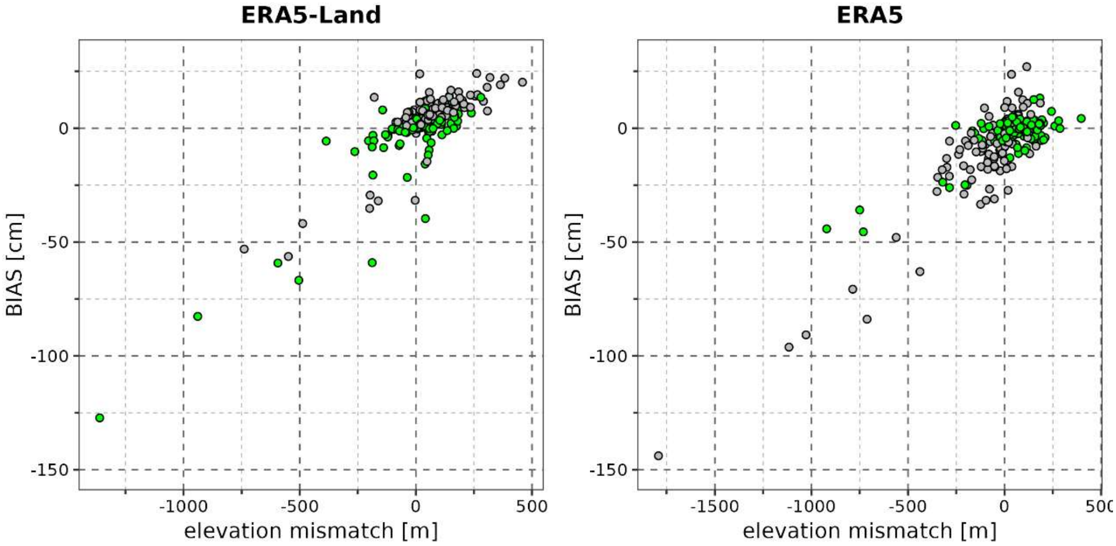

Fig. 1 Mean bias of snow depth for every station in relation to elevation mismatch (reanalysis minus real) for ERA5-Land (right) and the atmoshperic ERA5 (left).

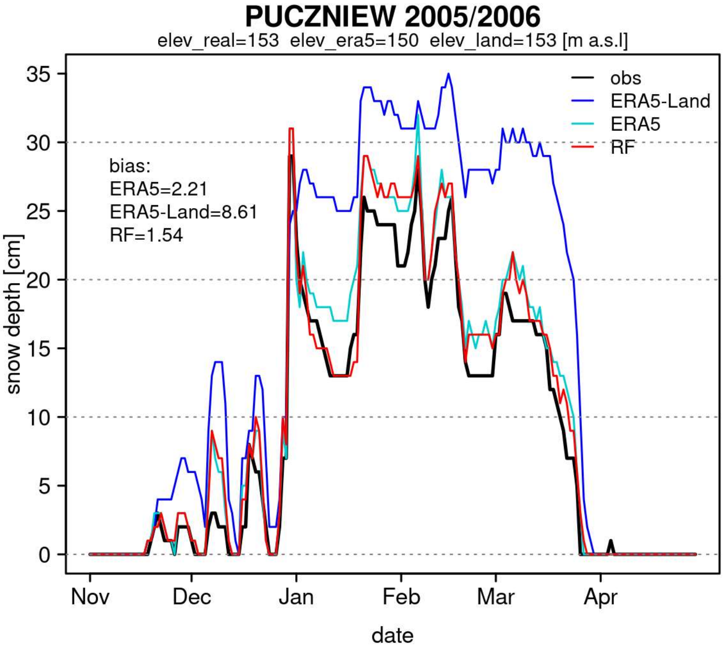

Fig. 2 Snow depth variability in the 2005/2006 season at station Puczniew. Despite negligible elevation mismatch, systematic errors still exist.

Regarding the inclusion of elevation mismatch at the stage of a RF model training – this piece of information is indirectly provided to the model with three variables: elevation from the Digital Elevation Model (as a proxy of real elevation) and model elevation of ERA5 and ERA5-Land. If we used relative fields instead (difference instead of absolute values), they would probably be the most important features during training. However, some information regarding absolute altitude might be lost. Actually, we conducted such an experiment and the difference in training MSE is negligibly small (6.321 versus 6.315 cm). Thus, the two alternative forms of information about elevation could be considered as equivalent.

2. *In the RF date, day, month, and year are used as input. In an operational setting, year and date would not be available? From the variable importance analysis, they seem to have some influence. Would it make sense to test a model without these variables?*

   Answer:

   We cannot see any reason why temporal variables like day, month, year and date would not be available when running operationally. However, it was not the goal of this study to propose a tool that could be run operationally. The main reason for it is that ERA5 reanalyses are publicly available with a delay of around 5 days.

3. *Sec 2.5 unclear how you split into training, test and validation sets. Was a validation set used at all? Similar to the previous reviewers, I strongly suggest including a validation set in the spatial domain. Moreover, it could be useful to give summary metrics for the different sets (training, test, validation), to see how well the model generalizes.*

   Answer:

   There is no separate validation set in a sense that occurs by other machine learning methods, e.g. artificial neural networks. In contrast to them, random forests (RF) does not use a separate, fixed validation dataset due to different training methodology. RF is a decision-trees-based method that uses bootstrap aggregating – a resampling method that randomly generates multiple data samples with some data replaced by duplicated samples of the original set (Boehmke and Greenwell, 2019). The part of the original dataset not included in a bootstrap sample (around 27% on average) is considered out-of-bag (OOB). The OOB samples are then used to validate the model. Hence, the score is called the OOB error. Breimann (2001) explicitly stated that the error "*removes the need for a set aside test set*". Under "test set" he meant "validation set" – these two terms used to be often confused in literature (Ripley, 1996). The OOB error is proved to be a good estimation of the model generalisation error (Breiman, 2001). Hence, the error is commonly regarded as a training error, while *sensu stricto* this is a validation error.

   Furthermore, we appreciate your suggestion regarding the spatial split. The training strategy has been extended so that, beside temporal split, also spatial split is concerned. Taking into consideration latitudinal pattern of land relief in the study area, the domain has been split into 5 longitudinal bands (every 2°), so that every band includes some mountainous stations which provide extensive data. Little variability of training error in both temporal and spatial split proves good generalization skill of the RF model. A detailed list of the errors for every combination is put in the Appendix C. In the manuscript (section 2.5), the information was given in an aggregated form (mean + standard deviation). In addition, the description of splitting was put in a separate paragraph in the Section 2.5 so that it is more comprehensible.

4. *Example downscaling: it unclear if the authors used interpolation of surface meteorology from stations using bicubic? Or what variables were interpolated to perform the downscaling? Note that simple bicubic is not appropriate for variables that have a strong elevation dependency such as temperature, humidity, … I don't know if this might be an explanation for the errors found. Of course it is difficult to validate such a dataset, but have you considered remote sensing products based on MODIS, such as globalsnowpack from the DLR or ESA snow CCI? Of course you'd have to convert snow depth to snow presence, but it could give you some independent spatial information.*

   Answer:

Thank you for this question. None of the station-measured parameters have been interpolated. Bicubic interpolation was performed over predictors from reanalyses in order to prepare a test set. As they already were continuous 2D fields before this operation, it was actually regridding rather then interpolation. We are aware that some of the fields are meteorological fields with distinct elevation dependence which is not accounted for while interpolating bicubicly. However, please notice that during retrieval of a point value from a reanalysis (as it was done for every station in the study area for training), it is often interpolation from the nearest node which is performed (alternatively, it could be a raw value from the nearest node, but in our case it wouldn't make sense). Therefore, the predictors are not really downscaled, but rather bicubically regridded. The description of the experiment setup in Section 2.6 has been modified so that it is hopefully more accurate and clear.

Regarding using an independent dataset, we are very grateful for this suggestion. At the initial stage of research, we found the remote-sensing-based products not relevant to our research as they mostly provide qualitative information about snow, not quantitative. However, it is indisputable that in cases where snow does not cover the entire study area, information about snow presence do provides added value. In order to fully benefit from it, date of the presented results had to be changed (at the initial date, the whole domain was covered by snow). Consequently, major modifications have been introduced in Section 2.4 and 3.3, including the figures. The snow mask from the GlobSnowPack database was added upon Fig. 8b.

MINOR COMMENTS:

1. *L69: Avanzi and Fontrodona are not appropriate references for the statement.*

   Answer:

   Thank you for pointing it out. Indeed the work of Avanzi et al. is an example of reanalysis that does not involve any numerical modelling. The second study we referred to does employ calculation using a regional snow model ΔSNOW, however only point-wise (for stations). Therefore, it cannot be considered dynamical downscaling. The incorrect references have been replaced with the relevant ones.

2. *L90: I guess topographic complexity can also be high in the Americas and HMA, depending on where you are.*

   Answer:

   What we intended to convey is that spatial resolution of the output of the ML models proposed in referred papers (specifically, Cui et al. (2023), King et al. (2020), Tanniru (2025)) is still insufficient regarding the scale of topographic complexity that occurs in the highest mountain range in our study area. No doubt that such complexity occurs also in mountains of North America or in the Himalayas. However, the papers concerning these regions propose spatial resolution that is indeed finer that the output, but with primary goal to accurately reflect snow depth (or SWE) over a large mountainous area, rather than offering horizontal resolution corresponding to the scale of topographic complexity. The sentence has been reformulated for clarity.

REFERENCES

Baba, M. W., Boudhar, A., Gascoin, S., Hanich, L., Marchane, A., and Chehbouni, A.: Assessment of MERRA-2 and ERA5 to Model the Snow Water Equivalent in the High Atlas (1981–2019), Water, 13, 890, https://doi.org/10.3390/w13070890, 2021.

Boehmke, B. and Greenwell, B.: Hands-On Machine Learning with R, 1st ed., Chapman and Hall/CRC, https://doi.org/10.1201/9780367816377, 2019.

Bouallègue, Z. B., Cooper, F., Chantry, M., Düben, P., Bechtold, P., and Sandu, I.: Statistical Modelling of 2m Temperature and 10m Wind Speed Forecast Errors, Mon. Wea. Rev., 1, https://doi.org/10.1175/MWR-D-22-0107.1, 2023.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cui, G., Anderson, M., and Bales, R.: Mapping of snow water equivalent by a deep-learning model assimilating snow observations, Journal of Hydrology, 616, 128835, https://doi.org/10.1016/j.jhydrol.2022.128835, 2023.

Dalla Torre, D., Di Marco, N., Menapace, A., Avesani, D., Righetti, M., and Majone, B.: Suitability of ERA5-Land reanalysis dataset for hydrological modelling in the Alpine region, Journal of Hydrology: Regional Studies, 52, 101718, https://doi.org/10.1016/j.ejrh.2024.101718, 2024.

Keller, R., Rajczak, J., Bhend, J., Spirig, C., Hemri, S., Liniger, M. A., and Wernli, H.: Seamless Multimodel Postprocessing for Air Temperature Forecasts in Complex Topography, Wea. Forecasting, 36, 1031–1042, https://doi.org/10.1175/WAF-D-20-0141.1, 2021.

King, F., Erler, A. R., Frey, S. K., and Fletcher, C. G.: Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada, Hydrology and Earth System Sciences, 24, 4887–4902, https://doi.org/10.5194/hess-24-4887-2020, 2020.

Ripley, B. D.: Pattern Recognition and Neural Networks, 1st ed., Cambridge University Press, https://doi.org/10.1017/CBO9780511812651, 1996.

Tanniru, S., Singh, K., Singh, K., and Ramsankaran, R.: Exploring Machine Learning's Potential for Estimating High Resolution Daily Snow Depth in Western Himalaya Using Passive Microwave Remote Sensing Data Sets, 2025.