The discussion paper "Seasonal forecasting of snow resources at Alpine sites" by Terzago et al. comprises a thorough analysis of the capabilities of a new framework to predict snow depths over lead times covering the entire winter season. To this end, the authors ran the physically based snow model SNOWPACK using input data from seasonal weather forecasting systems by ECMWF and Météo France that were downscaled and debiased to the location of three stations in the Western Italian Alps. The analysis shows some prognostic skill of the framework relative to a climatology-based reference guess.

While an interesting topic and a well-prepared manuscript, my main concern arises from the fact that no (trusted) local precipitation measurements were available. This necessitated the use of ERA5 data as ground truth or target, which likely derailed some of the findings, in particular those that deal with the effect of downscaling and debiasing. I acknowledge that not having accurate precipitation data is the norm, in particular in such a use-case driven study. Yet, the absence of local precipitation data is critical to a point, where I wonder how useful certain sections of this study are at all (4.2 - 4.4). In this situation, I recommend to infer local (solid) precipitation from snow depth using data assimilation and use this in lieu of measurements. This approach would be in line with the authors’ underlying premise that the snow model is the strongest members in their system (see line 582). I think SNOWPACK even has an in-built data assimilation mode that allows to use snow depth instead of precipitation as input.

Another concern is the lack of discussion about the strength of the forecasting skills. In particular for readers who do not juggle with BBS, CRPSS, or AUCSS on a daily basis, wording such as "demonstrates skill", "show skill", or "surprisingly good skill" is not really meaningful. Is "skill" just slightly better than guessing? The authors have quantitative data, but need to put them in context. I found an interesting statement in line 490: "seasonal forecasts of snow depth appear more robust than streamflow forecasts". This is where quantitative reference should be made to corresponding score data from the evaluation of seasonal streamflow forecasts.
Further specific comments:

Line 86 / 603: In the context of this study "multi-model" arguably suggests that there is more than one snow model involved, which is not the case. Please revise.

Section 2.3.3: Was local terrain accounted for in the downscaling (sky view for longwave, and terrain shading for shortwave)?

Section 2.4: Even if the authors use soil temperature boundary conditions, I would recommend a spin-up to allow for a realistic initial soil temperature profile for all simulations, i.e. also if snow depth was zero on Nov-1.

Section 4.5: I generally appreciate the approach of using a complex and trusted snow model to focus the uncertainty analysis to the forcing data and the models / methods use to derive them. However, at the same time I wonder, if a temperature index model would actually provide better results, because a) it only uses two input parameters and avoids deteriorated performance due to uncertainties associated with the other forcing fields (wind, radiation, ...); b) local calibration of a temperature index model is simple and fast, and can compensate for remaining systematic biases, e.g. arising from the lack of local precipitation data. It's probably not realistic to expect the authors to perform such a comparison as part of this paper, yet, the above consideration would make a useful complement to section 4.5.

Line 627: Being able to predict that snow will melt at a certain location in, say, May is fairly easy even if the forecasted weather data is uncertain. Having said that, the increase of some of the skill scores towards the end of the season is no real surprise.