Reply to referee #1

Dear reviewer

We would like to thank you for the very thorough review and many valuable comments. Please find our detailed answers below (in blue).

I'd first like to say that I found your manuscript very interesting. You were thorough in your evaluation of your model. I also like that your GitLab page encourages reproducibility and people to use your model. I took the liberty of directly annotating your manuscript. I think the content and research are very good, but the form and presentation quality could be improved so that it's easier to understand.

- We thank RC1 for his/her very thorough review of the manuscript. There are many valuable comments in the attached PDF that we will go through and incorporate to improve the presentation quality and clarity of the manuscript.

But here are some other more general comments:

- ML level: because you're aiming for an EGU journal, I think there might be a majority of readers who are not experts in ML. Therefore, I think that at some points, it's necessary to explain some terms (I've pointed some out in the manuscript). Be careful of the line where you become unnecessarily too technical and where you might lose some of your readers. I'd also add a section in the discussion advising people (who are unfamiliar with ML and are set on their traditional numerical models) on how to use your model.
    - We are aware of GMD not being an ML journal. We will carefully revise the manuscript and provide further references or explanations for ML terms that might not be common across GMD community.

- Consistency of terms: Be careful not to use too many names to refer to your model, sometimes CleanSnow, sometimes TCN. I'd stick to CleanSnow everywhere and use TCN only when you refer to the architecture; otherwise, it becomes very confusing. You devised a nice name for your model, so use it :) Re-read the manuscript and change it where needed.
    - We agree and we will refer to our model only as CleanSnow where appropriate.

- Cross-validation: The part about hyperparameter tuning of your model is briefly mentioned but very important. Did you do any cross-validation for this (if not, why not?)? And which hyperparameters were tuned and came out as best?
    - We have used 5-fold cross-validation with random train/validation set splitting in order to select hyperparameters of our model architecture and optimization process. These were very early experiments which we did not consider the key contributions and therefore did not include in our manuscript.
      In particular, we have searched across a pre-defined set of values for each of the parameters listed below in order to find the one which yields best results:
        - TCN parameters num_res, dropout and output_activation (num_res = 1, dropout = 0.25 and no output_activation yield the best results).
        - MLP parameters normalization and activation ("batch_norm" has improved the results, "relu" has shown superior performance over other activation functions)
        - Loss function parameters gamma and alpha (gamma = 2.0 and alpha = [0.5, 1.5] came out the best)
        - Optimizer learning rate (0.001 came out as the best)
    - For all remaining experiments, we have set a fixed random seed for training/validation split in order to assure easy and full reproducibility of our results.
    - batch_size = 128 was selected so that it is sufficiently large while still fitting into the GPU memory we had available.
    - Due to limited compute resources we have not optimized the remaining parameters and selected them based on similar architectures available in other works and our experience with designing machine learning models. The architecture of the TCN (num_channels and kernel_size) was selected so that the TCN aggregates information from the whole input sequence into the final timestep, which is then selected as the information which is passed further through the model (predict_timestep = 47).
      The appropriate number of layers can be computed as:
      num_layers = log2((seq_len – 1) * (dilation_base – 1) / (kernel_size – 1) + 1)
    - Other parameters such as input features or sequence length are not considered model hyperparameters. Their selection is described in detail in our manuscript.

- Figures and their legend: your figure legends are generally concise and need more information. Although this is very tedious work, legends should respect a few things, such as acronyms that come up in the figure should be referred to (and

generally explained) in the legend, and a reader should be able to understand the figure on its own without having to go look things up in the text. Please go over your legends again and make them more descriptive.
    - Thanks for the suggestion. We will revise all of our figure legends so that they comply with the rules you mention above.

- Results: In your description of results (Section "Experiments"), when making statements that can be backed up by numbers in parenthesis, these numbers should be provided (such as F1 scores). There are a lot of F1 scores in your figures that can be easily used to back up your claims. Otherwise, the reader has to go look them up in the figures, and your statements seem empty. One good example where you do this is line 285, but this should be in all other results, too: "e.g., demonstrates that the model confidently classified snow (TPR = 99.4%) in contrast to the classification of snow-free ground with (TPR = 88.4%)".
    - Thanks for the suggestion. We will include numerical results in all relevant paragraphs, as done in the mentioned example on line 285.

- Grammatical tense: Be careful about mixing up too many tenses; sometimes, you switch from past to present without making too much sense. For the sake of consistency, try to keep the same when talking about the same things: for example, keep past tense when talking about your experiments and present tense for the results.
    - We will review the manuscript and remove inconsistencies in tense.

- Presentation of results and discussion: it seems to me that quite a lot of the results are simply repeated in the discussion, and that's not very interesting. I suggest that if a question comes up in the results, you discuss it immediately (for example, the negative effect of the solar variable). Otherwise, the reader doesn't get an explanation, reads on, forgets about it, and suddenly finds it again in the discussion. Instead of repeating results, the discussion, for example, also needs the limitations of CleanSnow.
    - We will review the Discussion section and move items to the Results section where appropriate.
    - We briefly touch upon limitations of CleanSnow in paragraph 5.4 in the Discussion section. We will expand the discussion of limitations in the revised manuscript.

- Repetitiveness: your text is quite long, and I think you can make it shorter by removing unnecessary repetitions. Some things to remove are repetitions of things said previously in other sections ("as previously described in ..."), which can

just be a reference to a section. "As shown in Figure ..." can just be a statement with a "(Figure number)". I tried to strike out some things that jumped up to me, but I'll let you have a look.
    - Thanks for the suggestion. We agree that several things are repetitive and will make changes according to your suggestions in the annotated PDF.

- The problem of generalization: this comes back to the limitations of CleanSnow. You've shown that it generalizes well to stations within its training range but performs less well to those outside it. This is a normal limitation in ML but should be presented as such. CleanSnow will struggle when applied to a new station that is not within the distribution it's been trained on (which is normal because it's not like you did any transfer learning or something), but that's not a good generalization. So, I think your text needs more transparency about this limitation.
    - As you point above, generalization within the elevation range included in the training set is good. This means generally good generalization across the Alps, which is the region of interest for us. Generalization to out-of-distribution samples (stations located at elevations which are not well represented in the training data) is not good. Out-of-distribution generalization, however, remains an open problem in the ML community. We will dedicate a separate section in the Discussion regarding generalization and describe in detail how well our method generalizes and propose some ideas, which could improve the generalization ability of the model further, e.g. by incorporating known physical constraints into the model.

- I have an open question for you: you briefly mentioned input anomalies in your discussion. Did you notice any particular behavior for 2022 and 2023 (seeing as they're strong temperature anomalies)?
    - We have not noticed any particular behavior for 2022 and 2023, but we have not looked into these two years in much detail.

Best regards,

Jan Svoboda, on behalf of all co-authors