Response to Reviewer 2

Abbreviations:

AR Author Response (Johannes Horak)

RC Reviewer Comment

RC: This manuscript assesses the added value of ICAR relative to coarse reanalysis for estimating precipitation in complex topography. Not yet widely evaluated, ICAR is a promising tool for a range of applications. The methods in this study are robust and the conclusions are important. The manuscript should be a valuable contribution to the literature.

AR:

We thank the reviewer for his or her time spent on providing valuable and important feedback to the manuscript. We carefully considered all points that were brought up and incorporated the appropriate changes to the manuscript. Please find our detailed responses below.

Correction to the manuscript independent of the RCs:

P5L8: We found that the list of fields contained in the forcing file was incomplete. We added the two missing fields, the sentence now reads:

"The assembled ICAR forcing file contains ERAI zonal and meridional winds U and V, potential temperature Θ, pressure p, specific humidity q_v , **cloud liquid water mixing ratio q_c , cloud ice water mixing ratio q_i** and surface pressure p_0 at each 6 h forcing time step and every grid point within the domain."

P32L14: The list of employed open-source libraries was incomplete. We added the missing library. The sentence now reads:

"numpy (van der Walt et al., 2011), pandas (McKinney et al., 2010), xarray (Hoyer and Hamman, 2017), matplotlib (Hunter, 2007), cartopy (Met Office, 2010) **and salem (Maussion et al., 2019)**."

Comments

RC: 1. In terms of the manuscript structure, it seems a bit unusual to have a combined "Methods and Results" section (4). I can see why the manuscript was structured as it is, but I wonder if it could be rationalised at all. Could there be benefits from a more "traditional" separation of methods and results? For example, the major sections could go something like: 1. Introduction 2. Study Area and Data 3. Methods 3.1 ICAR Overview and Setup 3.2 Evaluation Strategy 3.3 Skill Scores and Significance Tests 3.4 Flow Linearity (explaining how flow linearity and stability are calculated) 3.5 Weather Patterns (explaining dataset with figure of weather patterns) 4. Results (as currently structured but removing the methods now described in the previous section) 5. Discussion 6. Conclusions. This is just a possibility; there could be a better way.

AR:

We agree that in this sense the structuring of the manuscript follows a more non-traditional approach. A separation of methods and results has its advantages, however, we are of the opinion that this format lends itself better to manuscripts that focus on one or two central methods. The manuscript in discussion, in comparison, introduces six different methods employed to investigate different aspects of the ICAR simulations. Here, combining methods and results allows for a more fluid reading experience while still enabling non-linear reading where readers may jump from method to method or result to result. This optimizes the logical flow by avoiding zig-zagging as suggested by Mensh, 2017.

RC: I would also suggest checking the manuscript for repetition and trying to minimise the amount of "referencing forward" (i.e. sometimes it is not necessary to say "X will be discussed in Section Y").

AR: We checked the manuscript for the aforementioned repetitions and removed the forward referencing where possible.

P5L8: We removed the forward reference "~~**(defined in Sect. 3)**~~"

P7L16-17: We removed the forward reference "~~**see Sections 4.7 and 4.8**~~."

P12L5-6: We removed the forward reference "~~**The performance of individual stations is discussed separately in Sect. 5.**~~"

P13L15-16: We removed the forward reference "~~**The potential factors contributing to the observed underestimations are discussed in Sect. 5.**~~"

P27L12: We removed the forward reference "~~**The performance of individual stations is discussed in Sect. 5.**~~"

RC: 2. One specific point on structure is that the optimal model top height is stated in Section 2.3, before the results from the sensitivity test are presented. This should be avoided I think. It is already stated in the relevant part of the results section (i.e. on model top sensitivity) that the remainder of the evaluation uses the optimal model top.

AR:

We agree and removed this forward reference. The last line of the paragraph now reads:

"**Therefore, a sensitivity analysis was conducted to identify the optimal elevation of the model top for this study.**"

RC: 3. Also regarding the model top sensitivity, it would be interesting to contextualise the variation in ICAR performance shown in Figure 2 by providing the equivalent MSE for ERAI. This could be as a horizontal line on Figure 2 or just stated in the text. I.e., even for the model tops leading to larger errors, do they still outperform ERAI overall?

AR:

Figure 2 shows the mean MSE of ICAR and $ICAR_{CP}$ over all alpine weather stations. Indicating the mean MSE of ERAI over all alpine weather stations does not necessarily indicate whether ICAR, on average, outperforms ERAI. To highlight this we added an additional panel to Figure 2 that shows the SS_{MSE} averaged over all alpine weather stations for ICAR and $ICAR_{CP}$.

ICAR is able to outperform ERAI at model top settings of 1.5 km, 2.5 km and 4.0 km, while $ICAR_{CP}$ additional shows added value over ERAI for a model top set at 5.7 km.

The updated Figure 2, the additional panel referenced above is panel b:

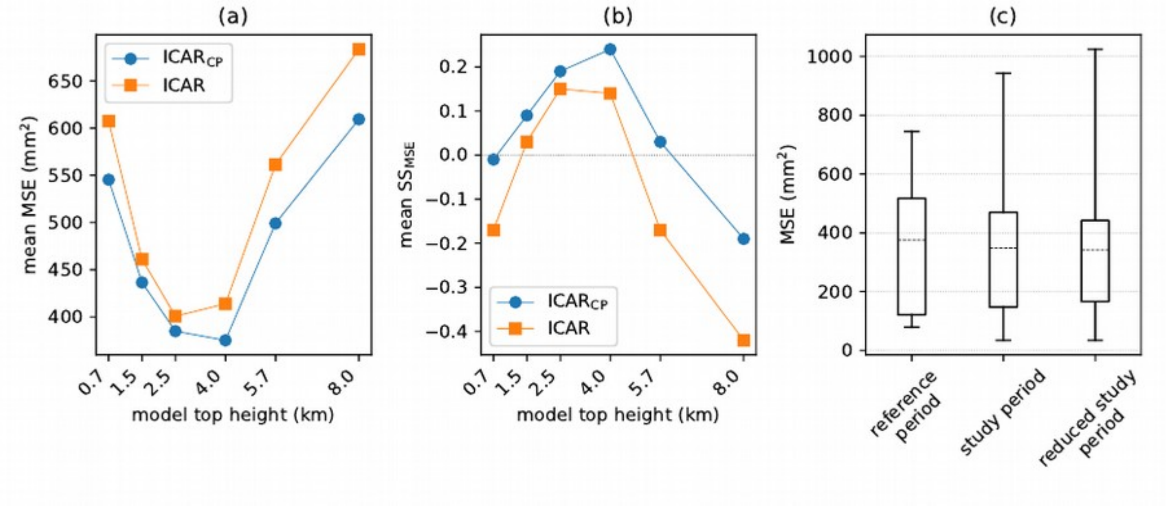

Figure 2. The average (a) MSE (b) SS_{MSE} of ICAR and $ICAR_{CP}$ time series from simulations for the reference period 2014-2015 at alpine weather stations as a function of the chosen model top (in km above topography). Connecting lines serve as guides to the eye. Panel (c) shows the distribution of skill scores for simulations with a model top set 4.0km above topography at alpine weather stations for the reference period (2014-2015), the full study period (2007-2017) and the reduced study period where the reference period has been removed from the dataset (2007-2013 and 2016-2017). The lower boundary of each box indicates the 25th percentile, the upper boundary the 75th percentile and the dashed horizontal line the mean. Whiskers show the minimum and maximum values of the data set.

RC: 4. It is mentioned in the discussion (P29 L19-20) that higher model tops lead to lower precipitation. Does this apply across the full range of model tops tested?

AR:

The statement refers to the mean bias observed for simulations with a given model top at alpine weather stations and applies to model tops greater or equal to 1.5 km. The simulations with model tops at 2.5 km and 1.5 km are, on average, wetter than those with a model top at 4 km, while the 0.7 km model top runs are, in that regard, similarly wet with respect to the 4.0 km simulations.

However, due to preliminary work outside of the scope of this article we suspect that the increased wetness associated with lower model tops is due to numerical artifacts. Further research is necessary to converge towards a definitive answer.

RC: Would there be any value in adding a second panel to Figure 2 showing mean bias for the different model tops? I.e. given that ICAR is generally low-biased for the Alpine stations, does a 2.5 km model top lead to reduced bias (even if the MSE is little different from 4 km)? Or does ICAR become high-biased with a 2.5 km top?

AR:

ICAR is, on average, too dry for all model top settings tested with respect to alpine weather stations. While simulations with 1.7 km and 2.5 km are wettest, preliminary work shows a strong indication that this behavior is due to numerical artifacts associated with the treatment of the model top. Furthermore, as indicated in the discussion (P29L26-27), it is currently not well understood why the precipitation decreases for higher model tops. Therefore, we are of the opinion that an additional panel showing the mean biases for different model tops would (currently) not be of added value.

RC: 5. In several places in the manuscript, results are discussed but the corresponding figures/tables are "not shown". This includes the relationship between model top and precipitation magnitude mentioned above, as well as seasonal averages at Alpine station locations (P16 L13-18). The latter case I found confusing initially, as the ERAI seasonality is reasonable-ish in Figure 5 but criticised in the text based on station locations (P16 L17-18). I would suggest considering whether some of the "not shown" figures/tables should be put into a supplement or whether references to them are necessary.

AR: While revising the manuscript we greatly reduced the amount of times secondary results are not shown. We added the corresponding Figures to the appendix and added references to them instead of the previously used "not shown". The locations in the manuscript are:

P10L27, P16L14, P16L17, P27L7, P27L10, P30L2, P30L9

RC: 6. It may be useful to provide the season definitions used in Section 4.5 (i.e. which months).

AR: We added abbreviations of the months that are associated with each season to the second paragraph of Section 4.5 (P16L6-10) and the caption of Figure 5. Please note that the updated caption of Figure 5 includes an additional sentence due to another RC (orange, non-bold text):

"The seasonal variations of precipitation as derived from the VCSR data set (Fig. 5b-e) are best reproduced by ICARCP (Fig. 5l-o). However, the improvements over the corresponding ICAR patterns 5g-j) are small and the remainder of this paragraph applies to ICAR and ICARCP alike. When comparing VCSR and ICARCP the similarities are largest for winter (**JJA**, Fig. 5h and 5m) and summer (**DJF**, Fig. 5e and 5o). The differences increase for the remaining seasons, with the Southern Alps being particularly affected. For autumn (**MAM**)**,** VCSR shows the precipitation as below average (Fig. 5b) while ICARCP indicates above average precipitation (Fig. 5l). For spring (**SON**), on the other hand, VCSR shows an increase in precipitation throughout the Southern Alps (Fig. 5d) but ICARCP shows the central part of the Southern Alps as drier than on average (Fig. 5n)."

Figure 5. The top four panels show patterns of P24h averaged over 2007–2016 for VCSR (left), ICAR (second column), ICARCP (third column) and ERAI (right) over the South Island of New Zealand and surrounding ocean. Rows two to five show seasonal deviations of the all-year average patterns, for autumn (**MAM**, second row), winter (**JJA**, third row), spring (**SON**, fourth row) and summer (**DJF**, bottom). Each panel shows the coastline and the 1000 m MSL contour line of the topography. High resolution plots are available in Horak et al. (2018).

RC: 7. The panels in Figure 5 are quite small so it is difficult to make out much of the detail. The overall improvement of ICAR over ERAI is clear though. It would be interesting to see a version of the figure zoomed in on the Alpine range, but perhaps this could be in future work.

AR: We added higher resolution plots of all the patterns to the data repository (https://doi.org/10.5281/zenodo.1135131) and indicated this in the caption of Figure 5 by adding the sentence:

"**High resolution plots are available in Horak et al. (2018).**"

RC: 8. Figure 3b has a spelling error - "coastal".

AR: We corrected the spelling error.

RC: 9. The boxplots for near-stable conditions in Figure 7c and 7d are quite different. What could be the reason for this?

AR: A question of Reviewer 1 led us to the discovery that some gridpoints were erroneously used for the flow linearity analysis. After redoing the analysis Figure 7c and 7d are more similar to each other. However, even in the updated version it is evident that the spread of scores in the "near stable" category is much larger than that in "stable" conditions. This is potentially attributable to the kappa threshold employed to distinguish between near stable and stable atmospheric conditions. If the threshold is set low, days that could reasonably be classified as stable (by investigating potential temperature profiles, for instance) are moved to the near stable category, leading to higher scores there. This behavior is observed for all skill measures employed.

RC: 10. There are a few places where the wording and grammar could be a little bit tighter. For example, sometimes "trend" is used when something like "pattern" might be better. There are other examples too, such as the first three sentences of the paragraph beginning on L15 on P20. The manuscript is generally fairly well written, but I would suggest that the authors check the wording and grammar throughout when making revisions.

AR: We reread the manuscript and, while making revisions, adjusted the text in cases where wording and grammar seemed problematic. See, for instance:

P3L27: We fixed a spelling error: "To avoid unstable atmospheric condition**s** present in the..."

P12L4: We fixed the grammar and spelling: "at this threshold $ICAR_{CP}$ performs very similar**ly** to ERAI, and that $ICAR_{CP}$ does not improve on modeling the frequency of precipitat**i**on."

P26L12: We exchanged "trend" for "**behavior**"

The paragraph starting at P20L15 now reads: "Of the 4018 days in the eleven-year study period, **1847 fulfill the criteria stated above**. A detailed overview **of** the distribution of these days among the three categories in dependence of κ is given in Table 3. The results from Table 3 ..."

RC: 11. In the abstract and discussion it is mentioned that ICAR can reduce MSE by up to 53%. If this is the maximum reduction, what is the mean/median? This may be worth including to give the "overall" picture.

AR: We rephrased the sentence in the abstract and discussion. In the abstract the rephrased sentence now reads:

"Furthermore, ICAR is found to provide added value over its ERA-Interim reanalysis forcing data set for alpine weather stations, improving mean squared errors (MSE) **by up to 53 % and 30 % on median.**"

In the discussion the updated sentence now is:

"At alpine sites in complex topography $ICAR_{CP}$ is then able to reduce mean squared errors in comparison to its ERAI forcing dataset by up to 53 % **and 30 % on median.**"

RC: 12. It could be mentioned again in the discussion/conclusion that a comparison of ICAR and WRF (or a similar model) might also be interesting for this study area. This might help us to understand some of the possible factors limiting ICAR performance discussed in Section 5. It would also give an idea of the relative performance gain from using WRF (if any) in a different climatic context to that tested in Gutmann et al. (2016).

AR: We agree that this might be of interest. However, ICAR is still a relatively new model. Preliminary work outside of the scope of this article gives us reason to suspect that other factors, such as numerical artifacts at the model top, might currently limit or influence the performance of ICAR. Before a meaningful comparison to a dynamic downscaling model can be made, it is necessary to develop a better understanding of these issues and how they can be overcome (or avoided) in future versions of ICAR.

References

Mensh, B., & Kording, K. (2017). Ten simple rules for structuring papers. *PLoS computational biology*, *13*(9), e1005619.