Response to Reviewer 1

Abbreviations:

AR Author Response (Johannes Horak)

RC Reviewer Comment

RC: In their manuscript Horak et al. assess the skill of the Intermediate Complexity Atmospheric Research Model (ICAR) for downscaling mean precipitation amount, in a domain located over the South Island of New Zealand. Model evaluation is performed using established techniques, a range of observational datasets and two skill scores. Their main findings are: (a) ICAR provides additional skill over the main Alpine ridge, while results over coastal stations are deteriorated. (b) Added value is typically largest for stable upstream flow, impinging on the ridge at a 90◦ angle. These results seem related to the model's roots, which is built on linear theory of orographic precipitation. C1

The article is generally well written and suited for publication in HESS (also for GMD). I particularly appreciated its modest and plain language. All review criteria are met, and I did not detect major scientific flaws, considering the manuscripts scope.

AR:

We thank the reviewer for his effort, and are very appreciative of the detailed comments and criticism of our manuscript. We took every comment very seriously and adjusted the manuscript accordingly.

Our efforts to address one comment regarding the flow-linearity analysis led to the discovery of an error in the underlying data set. We redid the entire analysis with the correct data and updated the affected parts in the methods section and in the discussion. However, the essential characteristics of the results have not changed.

Please find a detailed response to every comment below.

Corrections 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."

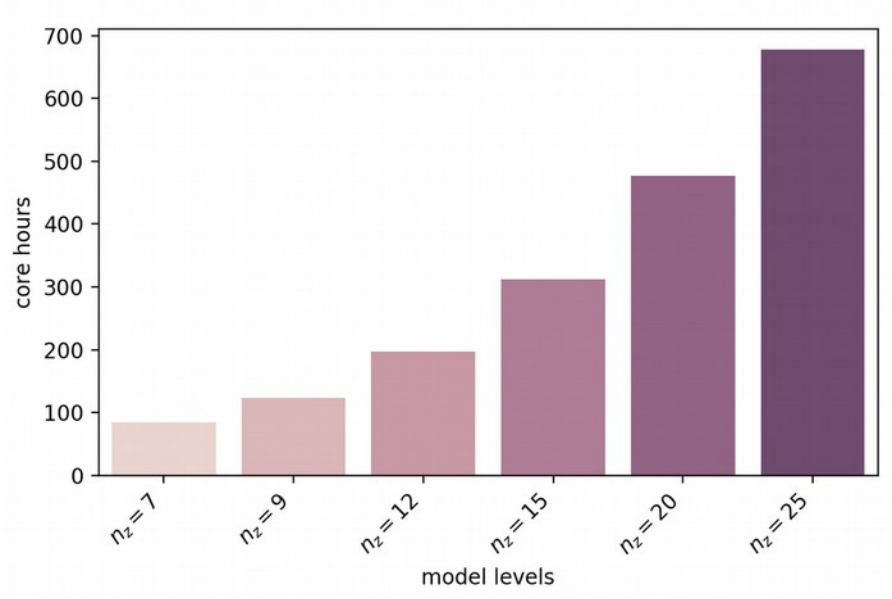
P32L14: The list of employed open-source libraries was incomplete. We added the missing library. The sentence now reads:

"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)**."

Minor Comments

RC: (1) P2L8: "While dynamic downscaling results in a self-consistent set of atmospheric fields, the computational cost required for the fine spatial and temporal grid spacing is high, especially for long-term simulations or sensitivity studies." This sentence would benefit from perspective. For example, for a similar computational domain we would achieve about 240 simulation days per day when running COSMO on a single node, equipped with a P100 GPU (Leutwyler et al., 2016; Fuhrer et al., 2018, I am not implying that you should cite my studies, but used then because I know the numbers by heart). How does ICAR compare these benchmarks?

AR:
The South Island of New Zealand ICAR simulations with 12 vertical levels, for instance, when run on one node of NCAR's Cheyenne cluster (with 36 2.3-GHz Intel Xeon E5-2697V4 Broadwell cores) have a ratio of about 10.5 simulated years per day (on average 200 core hours per simulated year). The following barplot shows the average number of core hours required by ICAR to simulate one year for the South Island of New Zealand domain in dependence of the number of vertical levels.



While we did not run WRF simulations for our study, Gutmann et al. 2016 did so. They found, that, depending on (but not only) the number of vertical levels and chosen microphysics parametrisation, ICAR speeds up simulations by a factor of 140. E.g. one simulated year for the Colorado domain as specified in Gutmann et al. (2016) and a WRF setup as given in Rasmussen (2014) required ~40,000 core hours (if the simulation were run on one CPU core only). ICAR, on the other hand, completes the simulation after ~300 core hours.

To clarify and lend perspective we added a sentence to P2L22-26 that references the findings of the Gutmann 2016 paper in the context of ICARs computational frugality. We also replaced the erroneously used term "linear theory of orographic precipitation" with the correct term "linear mountain wave theory". Please be aware that the updated paragraph includes changes made due to another comment as well (shown as non-bold, orange text):

"The Intermediate Complexity Atmospheric Research model (ICAR; Gutmann et al., 2016) offers a computationally frugal and physics-based alternative that does not rely on measurements with **linear mountain wave theory** as its theoretical foundation. In comparison to other downscaling

approaches of intermediate complexity (e.g. Sarker, 1966; Rhea, 1977; Smith and Barstad, 2004; Georgakakos et al., 2005), ICAR is a more general atmospheric model that requires fewer simplifying assumptions about the state of the atmosphere, such as spatial and temporal homogeneity of the background flow. Furthermore, in contrast to the linear theory of orography precipitation (LOP; Smith and Barstad, 2004), ICAR considers a detailed vertical structure of the atmosphere and employs a complex microphysics scheme as opposed to the characteristic timescales for cloud water conversion and hydrometeor fallout of the LOP. **With regards to dynamical downscaling, in particular the Weather Research and Forecasting model, Gutmann et al. (2016) have shown that ICAR may reduce the required computational time for one simulated year for a domain in the Western United States by a factor of at least 140.**"

RC: (2) P2L12: "to a lesser extent, to dynamic downscaling as well" I don't fully understand the statement in this fragment. Please elaborate on the stationary assumptions in dynamical downscaling, and how precisely this is overcome in ICAR.

AR:
If, for instance, a dynamical downscaling model is calibrated with measurements this indicates that not all parameters or variables may be inferred from theory or first principles. It follows that the parameters (or even a specific choice of a parametrization over another) determined by the calibration period may not necessarily apply to other periods with altered conditions equally as well. For global climate models, for instance, Maraun et al. (2017) note that "Often, a realistic behaviour is achieved only by tuning the model."
This applies to ICAR as well if empirical parameters of a physical process (i.e. parameters of the microphysics parametrization) are calibrated with measurements. Therefore, dynamical downscaling and intermediate complexity downscaling are both affected by the stationarity assumption if calibrated with measurements. We removed the part of the sentence to avoid insinuating that ICAR, when calibrated with measurements, somehow overcomes the stationarity assumption. The sentence now reads (removed text crossed out):

"Even more problematic, as soon as observation-based training or tuning is applied, the assumption of stationarity is introduced for statistical downscaling ~~**and, to a lesser extent, to dynamic downscaling as well,**~~ which may not hold under a changing climate (Maraun, 2013; Gutmann et al., 2012)."

RC: (3) Section 2.1: Adding a few plain language sentences how ICAR works and how the approach differs from dynamical downscaling would aid the wider audience. Additionally, a concise summary about linear theory of orographic precipitation and how it is incorporated into ICAR would help. I had to read Gutman et al. (2016) to understand this Section.

AR:
We rephrased the first and second paragraph of Section 2.1. (P3L11-16 and P3L17-21) to give a better overview of the basic functionality of ICAR, and its main difference from dynamical downscaling. The first paragraph (formerly at P3L11-16) now reads:

"**ICAR (Gutmann et al., 2016) is a three-dimensional atmospheric model based on linear mountain wave theory. As input ICAR requires a digital elevation model and a forcing dataset with 4-D atmospheric variables generated by, for instance, a coupled atmosphere-ocean general circulation model or an atmospheric reanalysis such as ERA-Interim. The forcing dataset should at least contain the horizontal wind components, pressure, temperature and water-vapor mixing ratio, with the possibility to additionally include**

hydrometeor fields, incoming long and short-wave radiation or the skin temperature of water bodies. ICAR employs linear mountain wave theory to calculate the wind field from the topography information and the horizontal wind components to avoid a numerical solution of the Navier-Stokes equations of motion, the core of dynamical downscaling models. With this wind field, ICAR advects atmospheric quantities, such as temperature and moisture as supplied by the forcing dataset at the domain boundaries. In its standard setup ICAR applies the Thompson microphysical scheme (Thompson et al., 2008), a double moment scheme in cloud ice and rain and a single moment scheme for the remaining quantities to compute the mixing ratios of water vapor, cloud water, rain, cloud ice, graupel and snow."

The second paragraph (formerly at P3L17-21) now reads:

"**The classic approach of linear mountain wave theory predicts the wind field based on the topography and the background state of the atmosphere. (Sawyer, 1962; Smith, 1979). With the background state known, its perturbation due to topography is given by a set of analytical equations (Barstad and Grønås, 2006). However, linear theory does not take into account interactions among waves or waves and turbulence, nor transient and non-linear phenomena such as time-varying wave amplitudes, gravity wave breaking or low-level blocking and flow splitting. A basic discussion of the limitations implicit to these assumptions can be found in Nappo (2012). In ICAR, the atmospheric background state is given by the forcing dataset. This yields a time sequence of steady state wind fields between which ICAR interpolates linearly. A detailed description of the model is given in Gutmann et al. (2016).**"

RC: (4) Section 4.1: Maybe it would be good to discuss the known biases for mean precipitation in ERAI and outline weather it is difficult to beat it.

AR:
A general statement about the performance of ERAI and how hard it is to beat is difficult to make since it depends, among other things, on the particular region of the world that is investigated and the specific factors that influence the local climate. Skill scores alone, in terms of percentage improvement, cannot fully account for how accurate a model is if nothing more is known about the reference model. For this reason we based our evaluation not on skill scores alone. We investigated the ICAR and ERAI precipitation time series at the weather stations as well and compared them directly to measurements. In our region ERAI simulates occurrence well and reproduces the measured time series but underestimates the precipitation magnitude (see Figure 4). This is in stark contrast to, for instance, the Peruvian Andes, another region we are currently investigating. Here ICAR skill scores are positive as well but precipitation occurrence and magnitude is not reproduced at all at some locations. The reason for the positive scores is that the performance of ERAI at these sites is worse. While ICAR is able to correct a little bit towards the measurements, this does not imply that the generated time series are realistic.

For a definitive assessment whether ERAI is difficult to beat, it is necessary to compare ERAI precipitation time series to those of measured at sites of interest. However, this was not the intended aim of the manuscript presented. Nonetheless we believe that Figure 4 gives a representative overview of the capabilities of ERAI with regards to modelling the 24h accumulated precipitation at the sites investigated within the study domain. While the timing of precipitation events is generally well captured by ERAI this is not the case for the magnitudes of the precipitation events.

RC: (5) Section 4.3 (a) Unfortunately, the chosen calibration period overlaps with the analysis period and employs the same stations. Cross-validation with other periods or station replacement would make the arguments more robust.

AR:
We agree with this assessment. Unfortunately even though ICAR is computationally more efficient than dynamic downscaling, performing, for instance, leave-p-out cross-validation would require extensive computational resources. However, the results suggest that the calibration period (2014-2015) is representative of the full study period (2007-2017) with regards to the presented calibration method. For the simulations with 12 vertical levels, the mean MSE of ICAR shows only little variation on whether the MSE is calculated for the calibration period, the entire study period or the study period excluding the calibration period.

To address this comment we added an additional paragraph to Section 4.3, an additional Panel to Figure 2 and an additional paragraph to the Discussion.

The new paragraph in Section 4.3 and the adapted Figure 2:

"The mean MSE over all alpine weather stations is almost constant when calculated either for the reference period (2014-2015), the full study period (2007-2017) or the reduced study period, where the reference period is excluded from the time series (2007-2013 and 2015-2017), see Fig. 4c. This result indicates that the reference period is representative of the full study period."

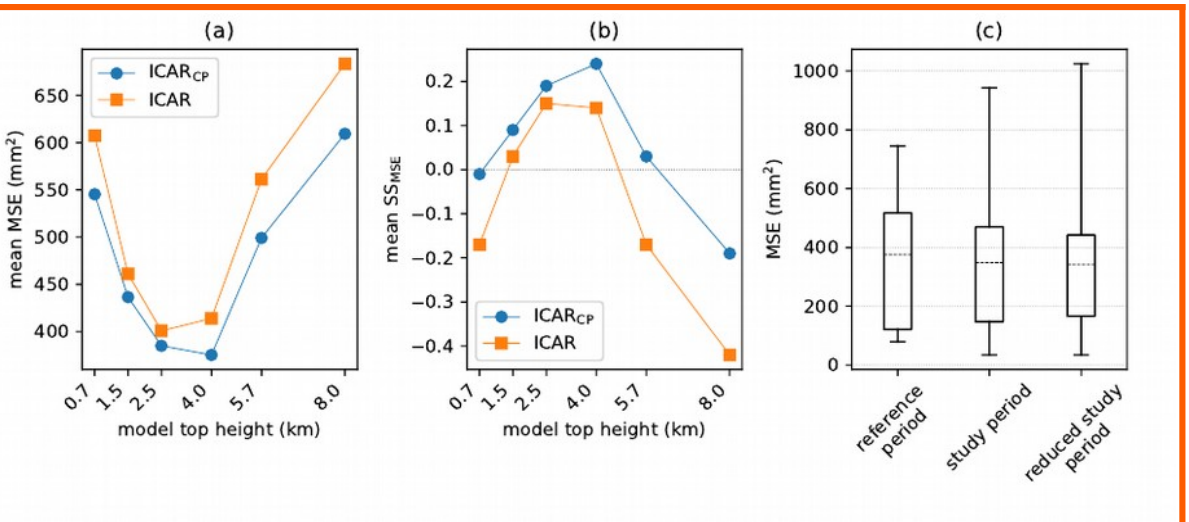


Figure 2. The average (a) MSE (b) SS_{MSE} of ICAR and $ICAR_{CP}$ time series from simulations for the reference period 2014-2015 at alpine weather stations as a function of the chosen model top (in km above topography). Connecting lines serve as guides to the eye. Panel (c) shows the distribution of 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.

The new paragraph that we added to the discussion:

"In this study, the chosen reference period (2014-2015) overlaps with the study period (2007-2017). While ICAR is computationally more efficient than dynamic downscaling, performing, for instance, leave-p-out cross-validation would require extensive computational resources. However, the results suggest that the reference period is representative of the full study period with regards to the presented calibration method: For simulations with the model top set at 4 km, the mean MSE over all alpine weather stations of ICAR shows only little variation on whether the MSE is calculated for the reference period, the study period or the study period excluding the reference period (see Fig. 2c). Furthermore, the variation between the mean MSEs for simulations with different model top settings (Fig. 2b) is larger than the variation between different evaluation periods (Fig. 2c)."

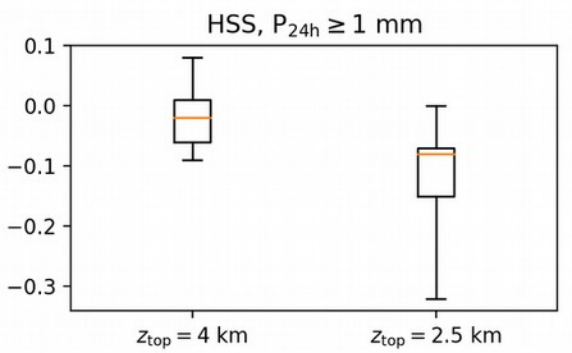
RC: (b) "Potential reasons for the observed behavior are discussed in Sect. 5." ! That statement is a bit misleading, since in Section 5 you only say that the question remains open.

AR:
Thank you for bringing this to our attention. The second part of the discussion that is concerned with this statement is located in a different paragraph in the discussion section (see P31L11-13). We rearranged the discussion section and included a only recently discovered potential cause (numerical artifact from model top treatment), the corresponding paragraph now reads:

"The sensitivity studies leading to the choice of the model top at 4 km have shown that the model top elevation greatly influences precipitation amounts and in turn the obtained mean squared errors, see Fig. 2. It is not immediately obvious though why precipitation amounts decrease (not shown) and the MSEs deteriorates for higher model tops. **Potential reasons are influences of divergences in the forcing wind field on the ICAR wind field or numerical artifacts arising from the treatment of the model top in ICAR. However, further research is necessary to develop a better understanding of this issue and its causes. Subsequently future studies could focus on finding** a method that allows the estimation of the model top elevation best suited for a domain without relying on measurements, as well as on **investigating** the influence of the choice of the forcing data type (i.e. global or regional reanalyses, GCMs, weather forecast models) and the spatial grid resolution thereof on ICAR dynamics and skill."

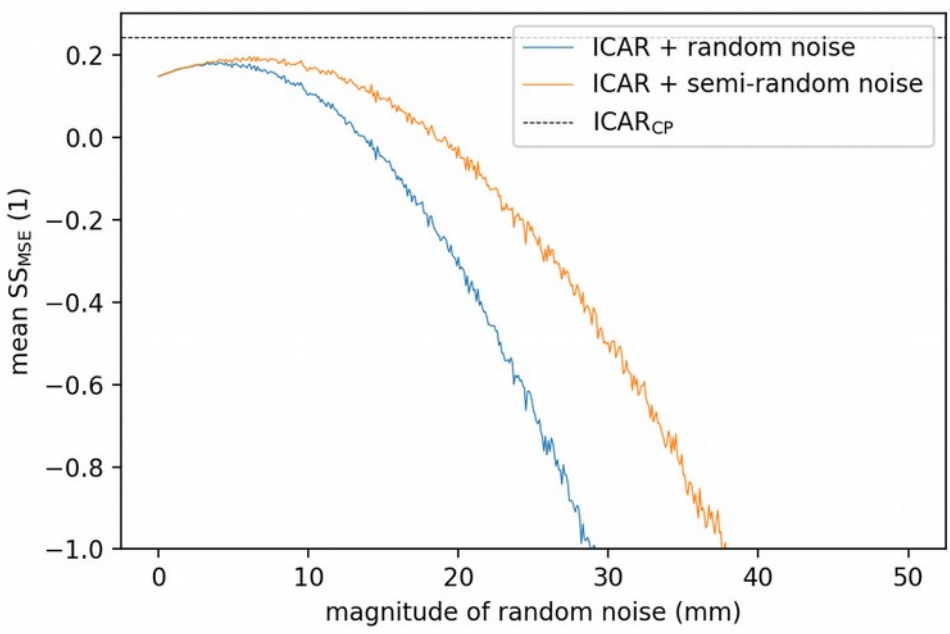
RC: (c) I am skeptic if the results at 2.5 km and 4 km are substantially different from each other.

AR:
One major difference between the two runs is that the simulations with a model top at 2.5 km cut off the atmosphere within layers that transport a significant amount of moisture within the domain. This entails a less faithful representation of the moisture content of the atmosphere and may in part lead to unphysical artifacts in the moisture distribution due to the way the model top is treated by ICAR. However, more research is necessary to quantify and understand this effect and how it affects the distribution of precipitation and moisture throughout the domain. While MSEs at alpine sites are similar but lower for simulations with a model top at 4.0 km, a particularly adverse effect is observed with regards to precipitation occurrence (HSS scores with $P_{24h} > 1$ mm). Here, a distinct score decrease is observed at all except one weather station if the model top is set to 2.5 km or lower.



RC: (d) A devil's advocate could argue that ICARCP mainly improves skill over ICAR because the latter underestimate precipitation amount (see P30L27). I.e., could the same skill be achieved by adding random noise with the right magnitude?

AR:
We tested this hypothesis and found that the addition of random noise to ICAR precipitation time series is not able to achieve the same mean skill as $ICAR_{CP}$. Moreover, even adding random noise only to days where ICAR predicts non-zero precipitation (semi-random noise) does not lead to a higher skill than is achieved by $ICAR_{CP}$.



RC: Does ICAR beat ERAI too?

AR:
We added an additional panel to Figure 2 that indicates the mean SS_{MSE} at alpine weather stations achieved for each model top setting during the calibration period. ICAR is able to outperform ERAI for model tops at 1.5 km, 2.5 km and 4.0 km.

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

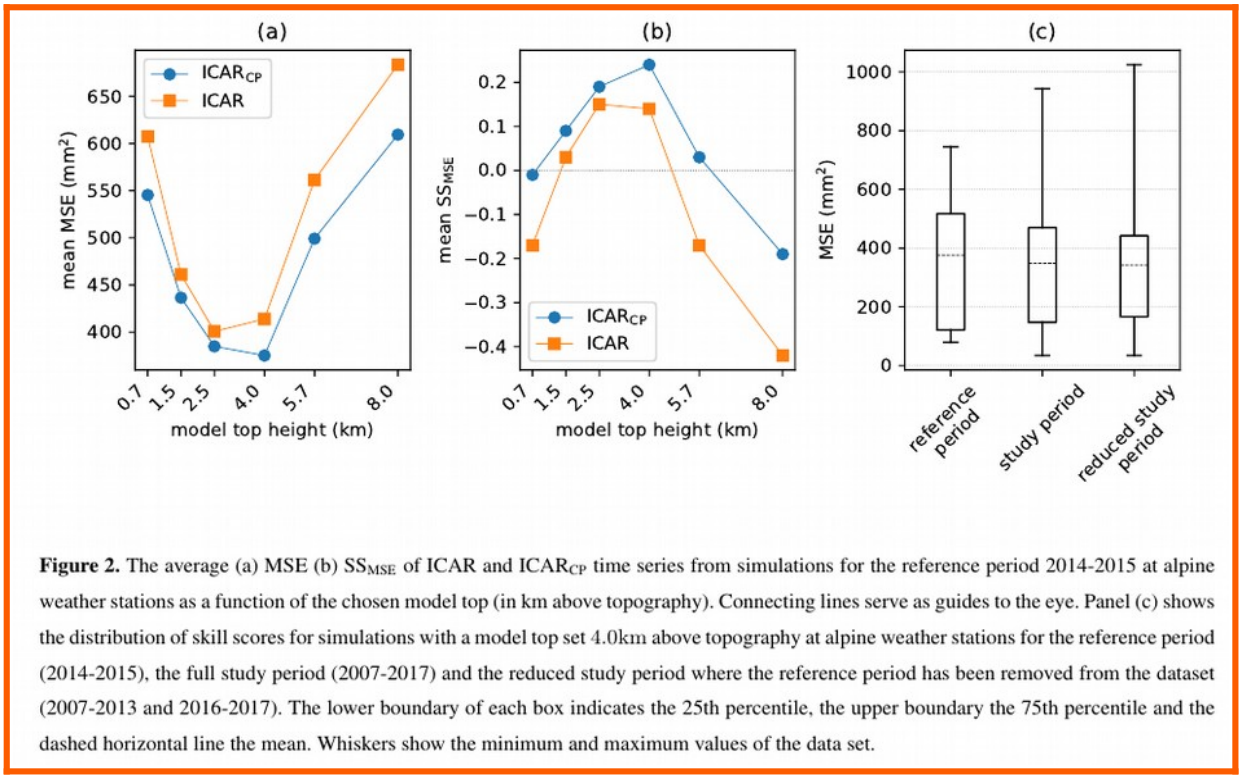


Figure 2. The average (a) MSE (b) SS_{MSE} of ICAR and $ICAR_{CP}$ time series from simulations for the reference period 2014-2015 at alpine weather stations as a function of the chosen model top (in km above topography). Connecting lines serve as guides to the eye. Panel (c) shows the distribution of 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: Please elaborate (here or in Section 2.6) to justify your choice to add interpolated conv. precip. from ERAI

AR:
ICAR is not able to model convective precipitation by itself in the setup used. Since ERAI does simulate convective precipitation and store the value in a separate field it seems a reasonable choice to use this additional information provided by the forcing dataset to improve the precipitation fields and time series simulated by ICAR. This is elaborated in Section 2.6, P5L18-20. Furthermore, it is a common technique to use convective or large scale precipitation from the forcing dataset this way, compare, for instance Roth et al. (2018) or other studies where the downscaled precipitation is a composite of precipitation generated by the downscaling model and the forcing for types of precipitation the applied model cannot account for (Jarosch et al. 2012; Weidemann et al. 2013 and Paeth et al. 2017).

To clarify we added these references to the manuscript. Please note that the updated paragraph contains an additional sentence added due to another RC (orange, non-bold text). Section 2.6 P5L24 now reads:

"where in the following the P(t) time series is referred to as ICARCP and P_I(t) as ICAR. **This is a common technique that allows to include types of precipitation not accounted for by the downscaling model (e.g. Jarosch et al., 2012; Weidemann et al., 2013; Paeth et al., 2017; Roth et al., 2018).** Table 1 shows the mean annual precipitation at each site for $ICAR_{CP}$ and ERAI, as well as the ratio of ERAI convective precipitation to ERAI total precipitation."

RC: (6) Section 4.7 Why is the underlying dataset changed to NCEP/NCAR?

AR:

The cluster analysis yielding the weather-patterns was not performed by us but by Kidson (2000), who employed the NCEP/NCAR dataset. We rephrased for clarity, it now reads:

"For the underlying cluster analysis, **Kidson (1994a) employed** the NCEP/NCAR 40-year reanalysis dataset (Kalnay et al., 1996) between January 1958 and June 1997 ~~**was employed**~~."

RC: (7) Section 5 (a) 1st paragraph: It might be worthwhile to elaborate on how these results relate to linear theory of orographic precipitation.

AR:
The linear theory of orographic precipitation (LOP) is, while connected to ICAR via the common basis of linear mountain-wave theory, not directly related to the results presented here. One fundamental difference is that the LOP, unless adapted as in, for instance, Jarosch (2012), is only able to consider a homogeneous background state across the entire domain. Similarly, unless adapted as in Barstad and Schüller (2011), information about the vertical structure of the atmosphere is, compared to ICAR, very basic. Another key difference is the use of a complex microphysics scheme (Thompson, 2008) in ICAR, while the LOP considers characteristic timescales for cloud water conversion and hydrometeor fallout. A comparison between the LOP and ICAR would be of interest, but outside of the scope of our manuscript.

To highlight the differences between the two models we modified P2L22-26 in the introduction. We also replaced the erroneously used term "linear theory of orographic precipitation" with the correct term "linear mountain wave theory". Please be aware that the updated paragraph includes changes made due to another comment as well (orange, non-bold text). The updated paragraph now reads:

"The Intermediate Complexity Atmospheric Research model (ICAR; Gutmann et al., 2016) offers a computationally frugal and physics-based alternative that does not rely on measurements with **linear mountain wave theory** as its theoretical foundation. In comparison to other downscaling approaches of intermediate complexity (e.g. Sarker, 1966; Rhea, 1977; Smith and Barstad, 2004; Georgakakos et al., 2005), ICAR is a more general atmospheric model that requires fewer simplifying assumptions about the state of the atmosphere, such as spatial and temporal homogeneity of the background flow. **Furthermore, in contrast to the linear theory of orography precipitation (LOP; Smith and Barstad, 2004), ICAR considers a detailed vertical structure of the atmosphere and employs a complex microphysics scheme as opposed to the characteristic timescales for cloud water conversion and hydrometeor fallout of the LOP.** With regards to dynamical downscaling, in particular the Weather Research and Forecasting model, Gutmann et al. (2016) have shown that ICAR may reduce the required computational time for one simulated year for a domain in the Western United States by a factor of at least 140."

RC: (b) P30L6: "Therefore these two instances are considered as outliers." I think there is a problem here

AR:
Following up other suggestions of the reviewer led us to discover that some ERAI grid points used for the flow linearity analysis were at the wrong locations (too close to the coast). We corrected this and redid the entire analysis. With the updated plots the added value of ICAR over ERAI for higher flow linearity and atmospheric stability is now more evident and the corresponding outliers have vanished. For more details see comment "P19L15: Cloud you add these regions to Fig. 1?" farther below.

Suggestions for optional extensions

RC: (1) Downscaling low-resolution global climate simulations (rather than re-analysis), along major mountain ridges could more evidently illustrate the added value of the approach.

> **AR:**
> We agree that this would indeed be a worthwhile analysis, it is outside of the scope of the presented manuscript. Additionally, some of the presented methods appear to be difficult to apply to global climate simulations, in particular the weather pattern analysis and the dependency of model performance on flow linearity.

RC: (2) From an application/user point of view, employing the outlined techniques to obtain higher-resolution fields is still a somewhat cumbersome procedure. It will therefore only be performed operationally if the added value is rather substantial. Therefore it would be interesting to see the added value over low-resolution precipitation climatologies such as, e.g., GPCC or GPCP..

> **AR:**
> We agree that this is a potentially fruitful avenue for further investigations. However, generally dynamic and statistical downscaling methods alike are generally tested for whether they actually improve over the employed forcing dataset (e.g. Jarosch 2012, or, for a review, Torma et al. 2015). ICAR is a relatively new model and, as mentioned in the introduction, this has not been established yet at the weather station level.

Technical Comments

Technical comments

RC: P1L1: climate downscaling => downscaling techniques

> **AR:** rephrased as suggested.

RC: P1L7: the eleven-year period from 2007 to 2017 => an eleven-year period, ranging from 2007 until 2017

> **AR:** rephrased as suggested.

RC: P1L9: diagnosed=> assessed

> **AR:** rephrased as suggested.

RC: P1L14: In the abstract, I would use a more general term for “flow of higher linearity”

> **AR:** Exchanged “flow of higher linearity” for “flow linearity”.

RC: P1L17: tuning => calibration (tuning has a negative connotation). Same applies to the rest of the manuscript.

> **AR:** Exchanged tuning for fitting variations of calibration throughout the manuscript.

RC: P2L21: Maybe add weather generators to the discussion?

> **AR:** While weather generators are functionally different from regression models, they do fall in the statistical downscaling category.

RC: P2L31: due => emerging from

AR: Rephrased accordingly

RC: P3L23: storing => stores

AR: Corrected accordingly

RC: P4L5: no data are => no observations are

AR: Rephrased accordingly

RC: P5L7: ERAI have => ERAI employs (I think ERA-Interim reanalysis is singular).

AR: Corrected accordingly

RC: P4L11 P5L19: "convective precipitation from the ERAI" Add the name of the field. Also, add a reference to your Table 1.

AR: Name and ID of the ERAI field was added and we referenced Table 1 at the end of the paragraph. Please note that the updated text as shown below includes an additional sentence due to another RC (orange, non-bold text).

Section 2.6 now reads:

The ICAR configuration for this study, as described in Sect. 2.2, is able to model orographic precipitation and, at least in part, precipitation driven by the synoptic scale. To account for convective precipitation, convective precipitation from ERAI **(field name: cp, parameter ID: 143)**, P_{CP}, is resampled to the ICAR timestep and bilinearly interpolated in space to the sites of interest and then added to the ICAR precipitation time series P_I:

$$P(t) = P_I(t) + P_{CP}(t), \quad (1)$$

where in the following the P(t) time series is referred to as $ICAR_{CP}$ and P_I (t) as ICAR. This is a common technique that allows to include types of precipitation not accounted for by the downscaling model (e.g. Jarosch et al., 2012; Weidemann et al., 2013; Paeth et al., 2017; Roth et al., 2018). **Table 1 shows the mean annual precipitation at each site for $ICAR_{CP}$ and ERAI, as well as the ratio of ERAI convective precipitation to ERAI total precipitation.**

RC: P6L5: New Zealand

AR: We rephrased the first sentence.

RC: P6L7: ranges => maybe "ridges"?

AR: We rephrased the paragraph, it now reads:

"This study focuses on the Southern Alps **of New Zealand** located in the southwestern Pacific Ocean. The **Southern Alps are** oriented southwest-northeast and run almost parallel to the western coast of the South Island. **They are** approximately 800 km long and 60 km wide, **extend** across a latitude range from 41° S to 46° S and consist of a series of ranges and basins (Barrell et al., 2011)."

RC: P7L12: In case of => For

AR: Rephrased as suggested.

RC: P9L10: Move sentence " The aim is not a downscaling ..." to end of paragraph

AR: Moved to the end of the paragraph.

RC: P9L27: HSS is defined as The HSS

AR: Rephrased as suggested

RC: P12L4: I relate "occurrence" to precipitation frequency. Maybe better use magnitude?

AR: The HSS for thresholds of P24h > 1mm may be seen as an indicator of whether $ICAR_{CP}$ is better able to model the frequency/occurrence of wet or dry days in comparison to ERAI. Higher tresholds, on the other hand, are more indicative of whether $ICAR_{CP}$ improves the frequency of larger precipitation events over ERAI. We exchanged occurrence for frequency for better clarity.

RC: P15L5: For lazy or tiered readers it might be helpful to re-state that VCSR are the observations.

AR: Rephrased to "The **observation and expert-judgment based** VCSR, ICAR, $ICAR_{CP}$ and ERAI"

RC: P16L6ff: Maybe indicate which months these seasons are (DJF..)?

AR: We added abbreviations of the months that are associated with each season to the second paragraph of Section 4.5 and the caption of Figure 5:

"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.

RC: P19L15: Cloud you add these regions to Fig. 1?

AR: We want to explicitly thank the reviewer for this comment as it revealed that some ERAI gridpoints used to determine the flow linearity were not within the test region (they were closer to the coast) and that the length of the test regions was erroneously stated as 1000 km when it should be 500 km. With the new test regions, furthermore, the maximum value of κ where still enough data points remained in the near stable category to calculate SS_{MSE} is $375 \cdot 10^{-5}$ s^{-1}.

However, the characteristics of the results remain essentially the same, with only minor effects on their discussion and presentation.

This entails the following changes in Section 4.6 and in the discussion:

P19L13: testregion dimensions corrected:
"... and is about 200 km wide, **500** km long and 1500 m high"

P19L6-7: the upper limit of κ has changed:
"…the value of κ is varied between $25 \cdot 10^{-5}$ s^{-1} and **375** $\cdot 10^{-5}$ s^{-1} in steps of $25 \cdot 10^{-5}$ s^{-1}."

P20L15: the number of days that fulfill the defined criteria has changed and we adjusted the text according to a criticism of reviewer 2:
"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."

P20L17-23: we updated the description of the results:
"The results from Table 3 summarized in Fig. 7 show, that stable atmospheric conditions and Froude numbers larger or equal to unity lead to an increase in median scores for sites in complex topography. This behavior is observed for SS_{MSE} where the score median increases from **0.33** to **0.58** and, for $P_{24h} > 25$ mm and $P_{24h} > 50$ mm in case of HSS. For $P_{24h} > 1$ mm the **maximum median score is found for stable conditions and F < 1, with the F ≥ 1 regime even yielding a negative median score.**"

P21: Table 3 was updated and filled with the correct values.

P22: Figure 7 was updated and now shows the correct values.

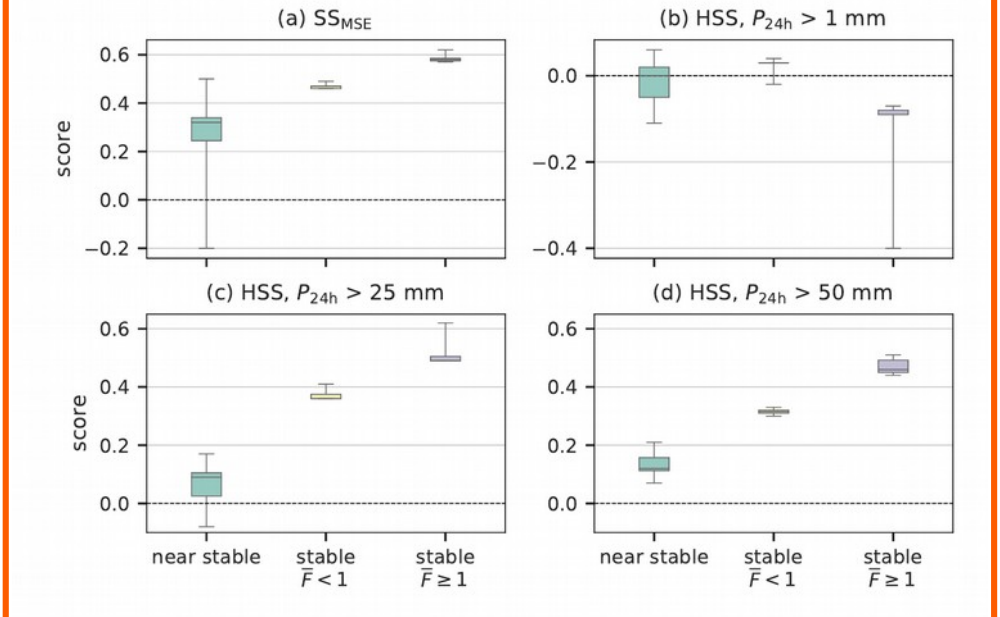


Figure 7. Dependence of SS_{MSE} and HSS at alpine stations on atmospheric stability and the Froude number regime, calculated for all available data for each value of κ (see Table 3). SS_{MSE} is shown in (a) and HSS for thresholds (b) $P_{24h} > 1$mm, (c) $P_{24h} > 25$mm and (d) $P_{24h} > 50$mm. The x-axis indicates atmospheric stability and Froude number regime. The lower boundary of each boxplot indicates the 25th percentile, its upper boundary the 75th percentile and a black line the median. Whiskers show the minimum and maximum values of the data set.

P29-30L34-10: We updated the discussion and included additional Figures to reduce the amount of times secondary results are not shown (as per request of Reviewer 2):

ICAR was found to perform better for upstream flows with Froude numbers larger than unity. This result is not unexpected, since linear theory is the theoretical foundation for ICAR. Therefore, flows of higher linearity lead to increased SSMSE and HSS for thresholds of 25 mm and 50 mm. These results hold even if the method for classifying near-stable or stable days is changed. For instance, using $N^2 \leq 0$ as classification criterion for near-stable days and $N^2 > 0$ for stable days leads to similar results (**see Fig. A2**). **For SS_{MSE} (see Fig. 7a) the spread of scores derived from varying κ for near-stable days is large enough to include the median score of the stable days with $F < 1$. Nonetheless, this is only true for $\kappa = 200 \cdot 10^{-5}$ s, in all other cases stable days with $F < 1$ always score higher than near stable days. Stable days with $F \geq 1$, in comparison, always achieve a higher score than the other two categories.** A potential issue with the methodology is the small number of cases in the stable regime with $F \geq 1$ compared to the two other classes (see Table 3). However, P_{24h} on stable days with $F \geq 1$ is three to seven times as high as P_{24h} during the other two classes (**see Fig. A3**). Therefore, while comparably small in number, stable days with $F \geq 1$ contribute above-average amounts of precipitation to the climatology, highlighting the importance of the improvement in skill for this category.

P6: Figure 1 with the test regions included now is:

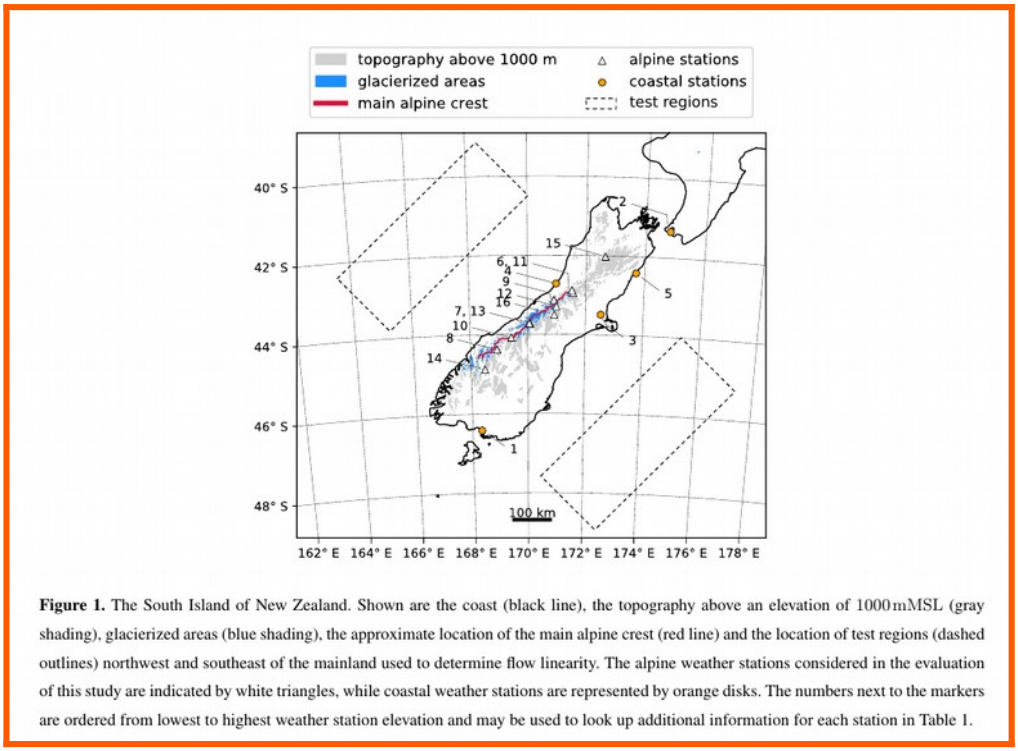


Figure 1. The South Island of New Zealand. Shown are the coast (black line), the topography above an elevation of 1000 mMSL (gray shading), glacierized areas (blue shading), the approximate location of the main alpine crest (red line) and the location of test regions (dashed outlines) northwest and southeast of the mainland used to determine flow linearity. The alpine weather stations considered in the evaluation of this study are indicated by white triangles, while coastal weather stations are represented by orange disks. The numbers next to the markers are ordered from lowest to highest weather station elevation and may be used to look up additional information for each station in Table 1.

RC: P29L9-17: (a) Maybe move this Paragraph to Section 3.2?

AR: While we agree that Section 3.2. would be a fitting place for paragraph P29L9-L17 as well, in this manuscript the uncertainties associated with precipitation measurements are only brought up in the Discussion section. To void unnecessary zig-zag and keep the logical flow of the discussion intact, as proposed by Mensh (2017), we decided to leave the paragraph at its current location.

RC: (b) Does undercatch not affect HSS(P>50)?

AR: Undercatch does affect HSS(P>50) as well and, as detailed in paragraph P29L9-17, is expected to affect both, the performance of $ICAR_{CP}$ and ERAI.

RC: P29L9-33: I would move the caveats to another place such that the paragraph currently starting at L34 follows after the current L8.

AR: We agree that the discussion would benefit from a more rigid structure. We therefore moved the general discussion of results up so that it now begins after L8. The caveats are now discussed subsequent to the general discussion of results.

RC: P30L21-24: Could you elaborate why you think this issue is a likely candidate?

AR: We expanded the corresponding paragraph to elaborate further.

It now reads: "**A potential cause for the observed negative correlation is, that the reflection of mountain waves at the interfaces between atmospheric layers can impact the distribution of orographic precipitation (Barstad and Schüller, 2011). Siler and Durran (2015) found, for instance, that wave reflection at the tropopause may either strengthen or weaken low-level windward ascent, which in turn affects the amount and distribution of orographic precipitation. The outcome was found to depend on the ratio of the tropopause height to the vertical wavelength of the mountain waves. Since ICAR currently does not account for wave reflection, its implementation could therefore lead to improvements in this regard.**"

RC: Table 1: Outline in caption where the uncertainty estimates come from (+/- 0.1).

AR: We added a short outline to the caption:

"List of weather stations used in this study sorted by their elevation. The table lists station number, elevation z, latitude (lat), longitude (lon), name, average distance downwind of the main crest of the Southern Alps (Δ) based on westerly and northwesterly flow, mean annual precipitation $\overline{P}$ **with the standard deviation both calculated for the years where data was available at the respective weather station**, fraction of convective precipitation in ERAI annual sum f_{cp}, length of the time series (l) and number of days removed due to missing entries or failed quality checks (d_m). The superscript following the station name indicates the data provider: NCD (1), NIWA (2) and University of Otago (3). Precipitation data for Larkins and Potts were lineary extrapolated to a full year. Δ was not considered for coastal weathers stations and no values were assigned for Mahanga and Larkins since they lie north and south, respectively, of the main alpine crest."

RC: Figure 2: Are these MSE of the annual sums (Add to the caption)? Maybe add the mean values so the results can be put into perspective.

AR: Figure 2 shows the average over all the MSE of P_{24h} calculated at each alpine weather stations.

RC: Table 2: These are mm/day (e.g. RMSE (mm)), correct?

AR: Correct, we adjusted the units in the column headers for clarity, the header now looks like this:

No	Name	length (yr)	days with P_{24h} above (%)			SS_{MSE} (1)	RMSE (mm day $^{-1}$)		bias (mm day $^{-1}$)		HSS (1)		
			1mm	25mm	50mm		$ICAR_{CP}$	ERAI	$ICAR_{CP}$	ERAI	1mm	25mm	50mm

Figure 5:

RC: (a) NIWA (top-left) -> VCSR

AR: We exchanged the column header "NIWA" for "**VCSR**"

RC: (b) Maybe mean magnitude over land to panels?

AR: We considered the suggestion and decided not to add the mean magnitude over land to the panels. The reasons are that the mean magnitude over land is never specifically referenced or discussed in the text and that the panels mainly showcase the high resolution precipitation patterns. Adding text would, furthermore, conceal part of the patterns.

RC: Figure 6: Why do the no. samples (circles) differ among the various thresholds in HSS? Explain in the caption.

AR: We added the following sentence to the caption of Figure 6 to explain the reason:

"**At some weather stations no days with P_{24h} > 25 mm and P_{24h} > 50 mm were observed or simulated during certain seasons, therefore no HSS scores could be calculated.**"

References

Barstad, I., & Schüller, F. (2011). An extension of Smith's linear theory of orographic precipitation: Introduction of vertical layers. *Journal of the Atmospheric Sciences*, *68*(11), 2695-2709.

Gutmann, E., Barstad, I., Clark, M., Arnold, J., & Rasmussen, R. (2016). The intermediate complexity atmospheric research model (ICAR). *Journal of Hydrometeorology*, *17*(3), 957-973.

Roth, A., Hock, R., Schuler, T. V., Bieniek, P. A., Pelto, M., & Aschwanden, A. (2018). Modeling winter precipitation over the Juneau Icefield, Alaska, using a linear model of orographic precipitation. *Frontiers in Earth Science*, *6*, 20.

Jarosch, A. H., Anslow, F. S., & Clarke, G. K. (2012). High-resolution precipitation and temperature downscaling for glacier models. *Climate Dynamics*, *38*(1-2), 391-409.

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

Maraun, D. et al. (2017). Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, *7*(11), 764.

Paeth, H., Pollinger, F., Mächel, H., Figura, C., Wahl, S., Ohlwein, C., & Hense, A. (2017). An efficient model approach for very high resolution orographic precipitation. *Quarterly Journal of the Royal Meteorological Society*, *143*(706), 2221-2234.

Rasmussen, R. et al. (2014). Climate change impacts on the water balance of the Colorado headwaters: high-resolution regional climate model simulations. *Journal of Hydrometeorology*, *15*(3), 1091-1116.

Siler, N., & Durran, D. (2015). Assessing the impact of the tropopause on mountain waves and orographic precipitation using linear theory and numerical simulations. *Journal of the Atmospheric Sciences*, *72*(2), 803-820.

Torma, C., Giorgi, F., & Coppola, E. (2015). Added value of regional climate modeling over areas characterized by complex terrain—Precipitation over the Alps. *Journal of Geophysical Research: Atmospheres*, *120*(9), 3957-3972.

Weidemann, S., Sauter, T., Schneider, L., & Schneider, C. (2013). Impact of two conceptual precipitation downscaling schemes on mass-balance modeling of Gran Campo Nevado ice cap, Patagonia. *J. Glaciol*, *59*(218), 1106-1116.