Review of nhess-2019-174:

"Simulations of the 2005, 1910 and 1876 Vb cyclones over the Alps – Sensitivity to model physics and cyclonic moisture flux"
by
Peter Stucki, Paul Froidevaux, Marcelo Zamuriano, Francesco Alessandro Isotta, Martina Messmer, Andrey Martynov

Recommendation: major revisions

The authors present three cases of extreme precipitation along the Alps associated with the classical synoptic situation Vb. They utilize downscaling of different reanalysis products for their analysis, where they employ WRF in a nested method to go down to 3 km resolution in the innermost domain. The authors test several sensitivities of their results with respect to the simulation setup, addressing both numerical as well as physical changes. They find that the lead time is among the most crucial parameters. While I find the manuscript well written, I struggle to see a clear motivation and conclusion for the study. The motivation is probably also somewhat difficult, because the manuscript tries to address many different questions at the same time: (1) What is the best downscaling setup for the cases in question. (2) What lead to the extreme precipitation. (3) Description of
the three individual cases and comparison. The treatment of all these topics makes the paper sometimes difficult to follow. Regarding the conclusions, similar arguments apply and the authors sometimes appear to state opinions/speculations that are not necessarily solidly grounded in the material they presented. However, after responding to some of my major concerns, I believe this manuscript is acceptable for publication in NHESS.

We thank the reviewer for the comments and helpful, concrete suggestions for many aspects and important details. Many of the comments made us rethink our argumentation, and we found ways to better explain a topic or a reasoning based on this review. At times, this needed a new paragraph, at other times a couple of extra words were sufficient. With all these efforts taken, we think that the reviewer contributed largely to an enhanced, more specific and clearer manuscript. We hope that he / she will also more easily find our general line of thought in the revised manuscript - from first finding a setup, and then applying it to assess precipitation and moisture flux in the three cases.

General Comments:

In the title, I am not sure what the authors really mean with the “cyclonic moisture flux”. Is it a moisture flux going in a cyclonic direction or the moisture flux associated with a cyclone? Given this ambiguity, I encourage the authors to further clarify this aspect in their title as well as throughout the manuscript to help the reader making a clear link.
Indeed, the comment has made us re-think about what is meant by ‘cyclonic’ and ‘cyclonic moisture flux’. We come to the conclusion that both aspects (cyclonic rotation and association to a cyclone) are relevant in our case, although we mainly refer to it as ‘anti-clockwise’. To be clearer, we’ve added an explicit description of the specific dynamic mechanism of what we call a ‘cyclonic moisture flux’. This can be found in the Introduction section.
We also see that the term ‘cyclonic moisture flux’ is indeed not commonly used in publication titles or abstracts. However, we have not found a less specific, but still correct and concise term that would substitute it in the title. For this, and because it is also a relatively important term in our article, we would like to keep it there. We think that most readers may have some notion of what is meant when they read the title, and will be ready to read up about it in the introduction.

Given that there are only three cases, and the fact that a majority of the reasoning for the downscaling is based on the most recent case, the authors should be more cautious about general statements on downscaling procedures, as the results are highly sensitive to the case(s) at hand. While the technicalities that were overcome by the group are certainly impressive, it is still not clear to me how generic these results can be treated. In order to make a more general claim about the downscaling for Vb situation, one would need to explore many more cases to arrive at a firm conclusion. The authors should thus make it clear that this study can at most give an indication what one might need to test in order to arrive at a more general conclusion.
According to the reviewers’ suggestion, we have tried to avoid the impression of making general conclusions throughout the manuscript. This is mostly done by adding some words like ‘for our purposes’, ‘in our case’, ‘our experiments’, etc. We have also added a clear statement in the conclusions: ‘Although going back far in time, we have only analyzed a very small number of events – many more cases would be needed to reach robust recommendations on how to configure a model for Vb cases. Nevertheless, we have demonstrated that one can achieve a relatively best configuration for the desired application with a well-thought, logical series of experiments.’

What made the authors pick a 10-day spin-up time? It seems excessively long for the investigation of such a regional and meso-scale influenced precipitation event. At the end, the authors arrive at a 1-day spin-up time anyway, but the vastness of the parameter space is not sufficiently motivated, similar to some of the other sensitivity tests.
The list of sensitivities is extremely exhaustive, ranging from resolution to resolution ratios over spin-up time to parameterizations and model domains. The enormous parameter space is rather difficult to grasp and all results will primarily be in relation to

the 2005 case, with general deductions being rather limited due to the specifics of the case. In general, it would aid the reader if the authors more clearly state their working hypotheses as well as the reasoning for their choices and expectations. This will make it more straight forward to follow the ensuing arguments.

Based on these comments, see also the specific comment below, we now state more clearly in Section 3.1 what (i) our initial expectation was (10-day spin-up allows soil moisture and other variables to reach a partial equilibrium), (ii) that unsatisfactory first results led to experiments with reduced spin-up, and (iii) when the best setup was found, checking its robustness by testing additional changes of the model configuration. This reasoning is now also better underpinned with references.

Also, we state more clearly that 'The goal of these last experiments is not to achieve a thorough sensitivity assessment for each tuning option, but to make sure that we have not chosen a sub-optimal setup. This is checked by further modifying a number of configurations of the WRF model which may have an influence on the simulation performance according to literature and from our experience'**. In addition, we have introduced more literature and some short explanations of why we check against another (technical) configuration. In the Abstract, we focus on the result concerning the initialization time only because it is best documented in the manuscript.**

The authors often refer to reproducing "correct" precipitation amounts. What is meant by correct? Presumably compared to observations, though the authors list several observations that are used. In addition, all of these "observations" also rely on some sort of downscaling and gridding, as data voids need to be filled. The authors, however, do not provide a detailed analysis of the representativeness of these observations. They refer to other studies that addressed these to some extent, but given the specifics of the case studies, the authors should also comment on the validity of the observations before comparing the model simulations to the data in order to claim "correctness".

We think that the term 'correct' describes well what we are striving for in the experiments, and therefore prefer to keep it in the manuscript. To be more specific about what we mean by correct, we have tried to include a reference where possible throughout the manuscript. For instance, we insert 'that would ideally produce', 'i.e., in the control area', 'compared to CombiPrecip'.

We are aware that the interpolations (as well as the original measurements) are affected by uncertainties and errors. The uncertainties and errors were analyzed in detail as described in the publications mentioned in the manuscript as well as in many applications (e.g. at MeteoSwiss). In-depth knowledge about the strengths and deficiencies of each datasets allows MeteoSwiss to make comparisons always considering the performance of the datasets used as reference. A detailed description of the representativeness of the observation for each case would replicate already published results and go beyond the scope of the present study. In section 2.1, we add a comment about our awareness about the uncertainties and errors of our datasets (and the consequential appropriate use).

The fact that more than one dataset is used is due to their availability (e.g. CombiPrecip is not available for the 1910 case) and the differentiated comparison with grid data and time series.

For the validation of, for example, precipitation, it has proved useful to use feature-based detections that consider location, shape, and timing. Why have the authors not considered more such verification tools for the study at hand? It appears the method referred to as EMD is in fact such a measure, though it appears confusing why the authors use a visual inspection for a quantitative comparison. The reasoning for the choices and omission of other tools should be clearly motivated.

The same as the reviewer, we see EMD as a kind of feature-based measure. In fact, we tested other measures, mostly using the spatialVx package for R (https://ral.ucar.edu/projects/icp/references.html), including SAL by Wernli and Schwierz, 2006. Advantages of the EMS were in our view: smaller number of subjective decisions (thresholds, smoothing options etc.), one instead of three measures, no failures of actual feature detection, a measure for the relative distribution of precipitation totals. As the reviewer suggests, we have summarized these reasons in the text.
Regarding the 'eyeball inspection', it seems to be a common and valid option and complement to machine techniques for pattern recognition, see
https://www.cawcr.gov.au/projects/verification/#Standard_verification_methods
https://www.swpc.noaa.gov/sites/default/files/images/u30/Spatial%20Forecast%20Verification.pdf
We do agree on the point that more reasons are required to justify the selection of measures and scores, though. Therefore, we have introduced some short justifications in Section 3.2.

For the philosophical concluding paragraph on page 25 not much hard evidence has been provided in the manuscript for the claims put forward. It thus reads more like a written piece of opinion than a well and quantitatively justified conclusions.

We suppose that the reviewer alludes to the use of 'propositions of possible weather'. We see this rather as an intuitive and informative description of what the ensemble provides in the end, and think this could be appropriate in a final paragraph. Apart from that, we do not see any other philosophical conclusions on page 25. We rather think that we describe options that have been tested already, or paths that could be taken to balance computational costs and meaningfulness of downscaling an ensemble.

<u>Specific Comments:</u>

The page (P) and line (L) numbers refer to the ones in the manuscript.
P1 L23: "to the cyclonic" and see comment above about the ambiguity of "cyclonic moisture flux".

Please refer to our reply to the general comments above.

P1 L28: "accurate directions" with respect to what? What is the reference?

We agree that 'accurate' is too vague here and have changed it to 'northerly'.

P8 L1: The precipitation data is interpolated. Can the authors please clarify if the interpolation was carried out in such a way that the total precipitation was unaffected by the interpolation? Depending on the kind of interpolation, the results for the totals can deviate.

Please refer to our reply to the general comments above (fourth comment). We added some sentences in Section 2.1 to illustrate our awareness about the interpolated datasets.

P11 L12: The authors speculate on the differences between ERA-Interim and 20CR in terms of moisture distribution, though the authors could provide direct evidence for their claim by investigating differences between ERA-Interim and 20CR fields for the case at hand in more detail.

Following this comment, we have added some more information about a short investigation of specific moisture at 1000 hPa.

P11 L22: Why did the authors chose to identify and track cyclones using geopotential at 500 hPa? This seems rather unconventional and needs further motivation.

We agree that it is a good idea to extend the explanation of our choices. Concretely, we introduce the tracking paragraph with: 'For the analyses, we use both sea level pressure (SLP) and mid-tropospheric pressure fields. SLP fields inform about the quality of the assimilation process in 20CR, and the isobaric pressure fields (at 500 hPa here) tell about the derivation of upper-air variables from the SLP information in 20CR. Combining SLP and isobaric levels has been found useful for cyclone tracking (Hofstätter and Blöschl, 2019).**'**

Fig. 4: The cyclone tracks for ERA-Interim look very edgy. In order to compare them better to the other plots, a grid that is not fixed to the grid spacing of the data could be beneficial, which is most often done in other cyclone track algorithms, see also comment above about cyclone track determination in this manuscript.

We agree on the 'edginess', but would prefer to leave the visualization as is in Figure 4d. In our view, it reflects the nature (the coarse resolution) of the underlying data; a smoothed version would rather mask it. In addition, we see the ERAI line as a rather complementary information in the plot. Note also that for the same reason, we did not interpolate the grid points in Fig. g – i, and included the original grid points in the new Figure 11b, d and f. For other variables however, we see that smoothing is necessary, e.g. for cyclone tracks or SLP.

Fig. 7, 8, and 9: I find these figures not very legible. Maybe this is due to the downgrading of the figure quality for the review process, but otherwise the readability of the information of these figures needs to be significantly improved. In particular the arrows are not very visible.

We fully agree on that, and all reviewers are of the same opinion. Unfortunately, the quality loss was introduced by the conversion to PNG format. This should mostly be enhanced with the current PNG, and the PDF for print is fine as well, in our view. Note also that the color key had been missing.

P22 L21: The authors should explain how PV is produced in the downscaling process, as this appears to be crucial in their arguments.

In accordance to a similar comment by Reviewer 2, we have rephrased the discussion without PV streamers. The aim of the paragraph is to mention the underlying physical processes and interaction between the Alpine orography and upper-level troughs. These physical processes can be referred to without introducing the notion of PV, which is indeed not introduced in our study. In fact, this has allowed to focus more on explaining the untypical results from the 1876 case. The new text reads as follows:

"This behavior of the model can be explained by the interaction of the Alpine orography with atmospheric circulation. Indeed, Vb cyclone trajectories are typically initiated by deepening upper-level troughs, which finally cut off from the westerly flow when passing over the Alps (e.g. Awan and Formayer, 2017). The interaction of upper-level troughs with the Alpine orography have been described in detail (Buzzi and Tibaldi 1978; Aebischer and Schär 1998; Kljun et al. 2001); the underlying processes include flow splitting and lee cyclogenesis, with further amplifications of the cyclone formation by frontal retardation and latent heat release due to orographic lifting. The combination of these processes implies that the cyclones are formed on the lee of the right side of the Alps, typically over the Ligurian Sea. In 20CR, however, the Alpine orography is very coarse, smoothed and reaches only about 1000 m a.s.l. (cf. Stucki et al., 2012). Hence, the influence of the Alps on the large-scale flow is limited in 20CR. Given also that the 1876 case is least confined by pressure observations, this allows untypical cyclone tracks in many 20CR members. Once accounting for a more and more realistic orography throughout the downscaling steps with WRF, the high-resolution runs may thus end up in a compromise simulation - driven both by the WRF model physics and by the 20CR input flow. In other terms, the large-scale flow forced from 20CR might not be compatible with the orography of the high-resolution domains."

P23 L24: How can the authors conclude that "nudging smaller domains can still be beneficial"? Has any evidence been provided in this study to support such a claim?

We agree that this could be understood as a claim. Therefore, we rephrased the sentence to: 'Although we find no relevant enhancements from nudging in smaller domains in our test experiments, nudging smaller domains could still be beneficial for other specific studies.**'**

P24 L1: The authors should be more specific what they are referring to with "traditional spatial verification scores".

Regarding traditional (or standard) methods, we referred to https://www.wmo.int/pages/prog/arep/wwrp/new/jwgfvr.html, among others. However, we see that the term might be confusing, and substituted it by 'more common', and we name our measures again.

Further changes made

- **Additional references: Hofstätter and Blöschl, 2019; Zbinden, 2005 (Annals of MeteoSwiss); Cioni and Hohenegger, 2019; Coppola et al., 2018; Compo et al, 2015; Joliffe and Stephenson, 2012; Wernli et al., 2008.**
- **Figure A2 redrawn with the maximum- and minimum-precipitation members. This makes more sense since we focus on these in the text.**