Anonymous Referee # 1

The paper of Camera et al. presents a complete hydrometeorological reanalysis of two high impact events in Cyprus island (Eastern Mediterranean) addressing the challenge of effective reconstruction of such kind of events for small to very small catchments (ranging from 5 to less than 100 km^2 in this study). Overall, the paper presents a detailed and complete exercise, which adds another piece to the puzzle, benefiting from the availability of increasingly advanced modelling systems at all scales of analysis. Furthermore, the analysis is performed over an extraordinarily important area for Cyprus water resources, using a considerable set of discharge data and also (even though partially) with the challenging issue of hydrological modelling in a mountain environment with rock fractures. I suggest three main improvements to the paper, listed below, and have some other minor comments. I hope my comments are helpful to further enhance the quality of the paper.

1. My first main comment concerns the GCM data source (i.e., ERA-Interim). I acknowledge that this study inherits the work done by Zittis et al. (2017), but this global reanalysis is now replaced by the ERA5 reanalysis. This point is important, also given the fact that ERA5 offers ensemble members, which could be very usefully used exactly for the problem analysed (i.e., hydrometeorological chains targeted to small and very small catchments). I ask the authors to deal with this point, of course not requiring new simulations with ERA5, but discussing it.
   The decision to use ERA-Interim was driven by the previous work of Zittis et al. (2017) and also by the fact that we wanted to downscale a re-analysis dataset that was closer to the resolution of existing forecasting, decadal prediction, and global climate models in order to resemble a realistic modelling chain for forecasting applications. Moreover, ERA5 is not yet in a very mature stage, as evidenced from the emails alerting users from time to time to the presence of errors in the database. Also, in some cases re-runs are released for some years because of simulation errors (Simmons et al., 2020). However, we agree that ERA5 represents an opportunity for future improvement of the model skills. We have added few lines in the abstract and a discussion of the matter in Section 5.3.

   Abstract, Line 18-19: "This set up resembles a realistic modelling chain for forecasting applications and climate projections".

   Results, section 5.3 WRF-Hydro simulations with modeled precipitation, Line 481-486: "The rainfall fields modelled by Zittis et al. (2017) and used in this study were downscaled from the ERA-Interim re-analysis dataset. The decision to use these modelled data was driven by the fact that ERA-Interim presents a resolution closer to that of existing forecasting, decadal prediction, and global climate models, therefore it resembles a modelling chain for forecasting applications and climate change projections (e.g., Reyers et al., 2019; Saha et al., 2014). For future studies ERA5, thanks to its finer resolution and the availability of ensemble members for uncertainty estimates, will be a valuable data source for improving the modelling chain over small (< 100 km^2) catchments".

2. Furthermore, I have some concerns about the calibration and use of the bucket model. In general, my idea is that the baseflow bucket model could not be so important for such short-time events. Indeed, the case studies analysed are rather impulsive. Furthermore, I think that the effects of the bucket model are somehow misinterpreted (please refer to a specific comment below). I suggest the authors revise and comment on their choice of calibrating in detail the baseflow bucket model.

   The hydrograph recession is made up of delayed surface runoff, interflow (lateral subsurface flow from the soil) and baseflow (groundwater). As suggested by Reviewer 1 in specific comment 20, we investigated the possibility to fit it by calibrating the overland roughness routing factor (OVRGH). We tested the sensitivity of OVRGH and we noticed that the parameter wasn't helpful in redistributing discharge, it was just increasing or decreasing it without modifying the shape of the hydrograph (new Fig. 3 and Fig. 4). In addition, for our runs we had already set OVRGH=1, which according to many authors is the maximum possible value that can be assigned to the parameter (Yucel et al., 2015; Verri et al., 2017). Therefore, we tried to capture the hydrograph recession better by increasing baseflow through the calibration of the reservoir (bucket) maximum volume (Z_{max}) and exponent (α). For Z_{max}, we aimed to set its value so that the reservoir could be filled between 10 January at h. 00:00 and 11 January at h. 12:00, indicatively within few hours and 2 days after the peak rainfall. The model redistributes the deep percolation exceeding the reservoir volume between the channel cells of the corresponding watersheds. For those watersheds that highly overestimated the baseflow due to spilling out of the groundwater reservoir, we further increased Z_{max}. For the exponent, we calibrated it fitting the pre-peak hydrograph. Details regarding how we modified the manuscript to incorporate these analyses and their results are given in the answers to the specific comments.

3. Finally, I believe the authors can go more into details analysing the catchments with rock fractures, which show too low performances that should be increased somehow (please refer to specific comments below).

   We tackled the problem from two sides. First, we modified the terrain slope categories (SLOPECAT) map and consequently the SLOPE coefficients (controlling deep drainage) based on geology. For gabbro and ultramafic rock types we forced a SLOPECAT resulting in a SLOPE coefficient equal to 1 (i.e., the maximum possible value) and therefore in a maximization of the drainage from the soil column to the groundwater reservoir. Second, based on geology and field observations (Camera et al., 2018), we modified the soil type map as well. The MODIS database, which was used for soil characterization, attributes a uniform clay loam soil texture to the Troodos Mountains. However, we have observed that at the higher elevations, where predominant geology is gabbro and ultramafic rocks, soils show a gravelly sandy loam texture (Camera et al., 2018). Therefore, we modified the MODIS map, attributing a sandy loam soil type for cells characterized by gabbro and ultramafic rocks. In the WRF-Hydro model, soil properties are linked to soil type. For the cell involved, this change of soil type resulted in a modification (among other properties) of the saturated hydraulic conductivity from 2.45E-6 m/s to 5.32E-6 m/s. Before applying these changes, we investigated the sensitivity of the saturated hydraulic conductivity (Ks), which was found to be a sensitive parameter (see new Fig. 3).

Despite our efforts to maximize infiltration and deep drainage to reduce the hydrograph peak, the model still overestimated the observed flow in the high elevation watersheds. Looking at observed temperature time series, it is likely that part of the precipitation on the mountains occurred as snow during the January 1989 event. However, we do not have observed snow height data. The WRF atmospheric forcing data, which was used coupled with the observed precipitation, slightly underestimates the temperature on the top of the mountains (i.e., the model is colder than reality). Thus, it does not seem to be a modelled temperature issue. The land surface model converts precipitation into snow and snow into melt water through a radiation- and temperature-based routine. The simulated snow depth and snow-water equivalent during the event of January 1989 might be lower than expected. Another indication sustaining this hypothesis is that for the event of November 1994 the model slightly underestimates the hydrograph peak. Details regarding how we modified the manuscript to incorporate these analyses are given in the answers to the specific comments.

Minor/specific comments

1. Abstract: stating that "few studies evaluate the hydrologic performance etc...." is a bit debatable concept (e.g., few with respect to what?). This statement is different from a similar one on L81, where the authors specify that they are referring to WRF-Hydro. I would start the manuscript with a stronger sentence. Furthermore, in the Abstract the fact that 1989 events are used for calibration and 1994 events for validation should be stated more clearly.

   We have modified the first sentence of the abstract and have added the reference to calibration and validation for the two events of January 1989 and November 1994 as follows.

   Abstract, Line 12-13: "Coupled atmospheric-hydrologic systems are increasingly used as instruments for flood forecasting and water management purposes, making the performance of the hydrologic routines a key indicator of the model functionality".

   Abstract, Line 19-20: "Streamflow was modelled during extreme rainfall events that occurred in January 1989 (calibration) and November 1994 (validation) over 22 mountain watersheds".

2. L46 (and throughout the text): I would write "As summarized by Rummler et al. (2019)" rather than "As summarized by (Rummler et al., 2019)".

   Thanks for spotting it, we modified as suggested and we searched for similar occurrences throughout the manuscript.

3. L85: it looks like the events are much shorter. Including the spin-up period in this time interval could be misleading.

   To clarify this point, we modified the manuscript as follows.

   Introduction, Line 89-95: "The focus is on two extreme events that occurred over 22 small watersheds, located in the Troodos Mountains of Cyprus, between 8-10 January 1989 and 20-22

November 1994. The main objectives are: (i) to calibrate the uncoupled WRF-Hydro model for simulating extreme events in Cyprus with observed precipitation; and (ii) to evaluate the model performance when forced with WRF-downscaled (1 × 1 km^2) re-analysis precipitation data (ERA-Interim). The model runs covered two 15-day periods (1-16 January and 11-26 November) to include a short spin-up of the WRF-Hydro routines and the simulation and evaluation of the receding limb of the hydrograph".

4. Fig. 1: I suggest the authors focus more on the WRF-Hydro domain, which could be represented with a larger scale (so that also other information, e.g., location of raingauge stations and reservoirs, can be added). Location of the WRF-Hydro domain in Cyprus island could be shown with another small map in the figure.

   We have modified Fig. 1 according to the suggestions.

5. Table 1: A clear geological description is ok, but I would also highlight some essential geographical/morphological features, such as area, channel length, etc. Maybe authors can move some piece of information from Table 4 or just repeat it.

   We added area and channel length in Table 1 and left all the other variables in Table 4 as they were in the previous version of the manuscript.

6. L121: the problem of getting a reliable rating curve is rather common. More details about the "appropriate" rating curves used would be useful.

   We have added the following.

   Data, section 3.1 streamflow data, Line 124-130: "For the 22 watersheds, daily discharge data (m^3 s^{-1}) from streamflow stations of the Cyprus Water Development Department for the period 1980-2010 were analyzed. In addition, the original continuous hydrograph charts (water levels) of 16 of the 22 streamflow stations from the Water Development Department, for the Jan-1989 and Nov-1994 events, were scanned and manually digitized through the GetData Graph Digitizer software (http://getdata-graph-digitizer.com). The digitized water levels were interpolated to obtain values precisely every 15 minutes (00.00, 00.15, 00.30, 00.45, 01.00….) and converted to discharge with the appropriate rating curve of the station. The streamflow stations and rating curves are maintained by the Water Development Department through frequent observations".

7. Eq. 6: the variable Z should be explicitly defined

   We have added an equation (eq. 7) to define Z. The manuscript has been modified as follows.

   Modelling setup, section 4.1 WRF-Hydro model description, Line 208-218: "The second solution consists of calculating a baseflow discharge [m^3 s^{-1}] (Q_{bf}) by means of an exponential bucket model, described by the following equation:

$$Q_{bf} = C \cdot \left(e^{a \cdot \frac{z}{z_{max}}} - 1\right), \qquad (6)$$

where *C* is the bucket coefficient [m^3 s^{-1}], *a* is the bucket model exponent [-], Z_{max} is the maximum bucket level [m], and *Z* [m] is the bucket level at a certain time step. The user defines the *C*, *a* and Z_{max} parameters for each sub-watershed, together with a Z_{ini} [m] parameter to initialize the water storage in the bucket groundwater reservoir. At each time step the *Z* value is updated first adding the deep drainage contribution (*Perc*) and subsequently subtracting Q_{bf}:

$$Z_t = Z_{t-1} + \sum_{n=1}^{n=ncells} Perc_n - \frac{Q_{bf} \cdot DT \cdot 3600}{A} \qquad (7)$$

where A is the area of the sub-watershed [m^2], DT the model time step [day], n is the index for the sub-watershed cells, and ncells represents the number of cells of the sub-watershed. Similar to the first solution, Q_{bf} is equally redistributed to channel segments. If *Z* equals or exceeds Z_{max}, all deep drainage is transferred to the channel network”.

8. L218: information about average soil moisture would make more sense if information about soil type was provided

   We have modified the manuscript specifying the soil type as measured during experiments. Also, we added how we modified the original MODIS soil map to take into consideration the high permeable soils of the upper mountains (see also answer to general comment 3):

   Methods, section 4.2 WRF-Hydro Parameterization, Line 234-242: “Experimental data (Camera et al., 2018) show that in these conditions soil moisture for a gravelly sandy loam at 1300 m a.s.l. in the Troodos Mountains can vary between 0.10 and 0.15 m^3 m^{-3}. Therefore, the WRF-derived initial soil moisture values for November were halved.

   Land use and vegetation cover data were derived from the MODIS dataset through the WRF Pre-Processing System. According to the MODIS dataset, the Troodos Mountains has a uniform clay loam texture. However, field observations at higher elevation in the mountains, where the predominant lithologies consist of gabbro and ultramafic rocks, showed a gravelly sandy loam texture (Djuma et al., 2020; Camera et al., 2018; Cyprus Geological Survey Department, 1995). In addition, it is known that the Troodos gabbro is very weathered and therefore permeable (Christofi et al., 2020). Therefore, a sandy loam soil type was assigned to these areas.”.

9. L228: 1500 cells should be 1500 x 100 x 100 = 15M m^2, that is 15 km^2 (it should be better stated explicitly). However, in Table 4 there are some catchments with area lower than this threshold.

   Right, that data was wrongly reported. The threshold is 250 cells (2.5 km^2). It is now clearly stated in the manuscript.

   Methods, section 4.2 WRF-Hydro Parameterization, Line 245-246: “For the channel grid, a flow accumulation threshold of 1500 250 cells (2.5 km^2) was adopted”.

10. L267: at a time

    Modified as:

Methods, section 4.2 WRF-Hydro Parameterization, Line 265-267: “The initial level of the conceptual reservoir (Z_{ini}) was set as a fraction of the maximum level (Z_{max}), based on the saturation degree of the deepest soil layer at the end of the 15-day WRF spin-up period”.

11. Fig. 4 and elsewhere: to compare the performances of the model system for the two events, probably percent bias and MAE are more appropriate indices

    We have modified Fig. 4, Fig. 7, and Fig. 8 substituting BIAS with percent bias (PBIAS). Figure numbering changed because we added a new Fig. 4 for the sensitivity analysis results, so they are now Fig. 5, Fig. 8, and Fig. 9.

12. LL320-333: [this comment refers to the main comment about dealing with rock fractures] from this paragraph, it’s not clear if the problem is mainly related to the snow model in the LSM or the not good representation of the geological features. I would favour the second hypothesis, and I think that some test should be performed (and shown) by the authors increasing drainage.

    As explained in the answer to the general comment 3, we have modified the parameter controlling deep drainage and the soil type based on geology (increased deep drainage and coarser soil for areas with gabbro and ultramafic rocks). We have incorporated in the sensitivity analysis one run with the modified deep drainage and three runs with different saturated hydraulic conductivity values, relative to different soil textures in the soil parameter tables. We have noticed a high sensitivity of saturated hydraulic conductivity and a rather low sensitivity of the deep drainage parameter. In the final model parameterization, we considered the results of the sensitivity analysis. In detail, we modified the manuscript as follows.

    Methods, section 4.2 WRF-Hydro Parameterization, Line 237-254: “Land use and vegetation cover data were derived from the MODIS dataset through the WRF Pre-Processing System. According to the MODIS dataset, the Troodos Mountains has a uniform clay loam texture. However, field observations at higher elevation in the mountains, where the predominant lithologies consist of gabbro and ultramafic rocks, showed a gravelly sandy loam texture (Djuma et al., 2020; Camera et al., 2018; Cyprus Geological Survey Department, 1995). In addition, it is known that the Troodos gabbro is very weathered and therefore permeable (Christofi et al., 2020). Therefore, a sandy loam soil type was assigned to these areas. The related properties were attributed through the default table values implemented in WRF-Hydro (see Gochis et al., 2015). The hydrologic input layers (latitude, longitude, topography, flow direction, channel grid, lake grid, stream order, watersheds) were all calculated in ArcGIS® 10.2.2 starting from a 25 × 25 m^2 Digital Elevation Model (see Camera et al., 2017), resampled on the 100 × 100 m^2 grid, and the known locations of stream gauges and lakes. For the channel grid, a flow accumulation threshold of 250 cells (2.5 km^2) was adopted.

    For the definition of the deep drainage related parameter, two approaches were tested. First, nine slope terrain classes were derived following Silver et al. (2017). In the second case, for cells where the bedrock consists of gabbro or ultramafic rocks (Cyprus Geological Survey Department, 1995), the slope terrain class (3) that maximizes drainage (representing a highly fractured system) was assigned. In both cases, for each slope terrain class, the related default SLOPE value listed in the WRF-hydro general parameters table was given. These changes in soil type and deep drainage based on geology affected mainly watersheds Ma, An, Pl, Ka, and At, where 70% or more of the surface bedrock is made up of gabbro and ultramafic rocks (Table 1)".

Methods, section 4.3 WRF-Hydro Sensitivity Analysis, Line 269-277: "A sensitivity analysis of the LSM parameters REFKDT, SLOPE, and soil depth (SD), which have been identified as sensitive parameters in previous studies (e.g., Fersch et al., 2019; Senatore et al., 2015), was performed for the Jan-1989 event. In addition, sensitivity runs for the OVRGH parameter and the saturated hydraulic conductivity (K_S) were performed, too. For these simulations, the baseflow routine was switched off. A reference scenario was set, with REFKDT and OVRGH equal to 1, SD equal to 1.0 m, K_S equal to 2.45E-6 m s^{-1} (value attributed to clay loam soils in the soil parameter table), and the deep drainage parameter (SLOPE) assigned based on terrain slope, as in Silver et al. (2017). Parameters were changed one at a time. Eight values were tested for REFKDT (0.3, 0.5, 3.0, 5.0, 8.0, 10.0, 100.0, 1000.0), two for SD (0.5 and 2.0 m), two for OVRGH (0.1, 0.5), three for K_S (3.38E-6 m s^{-1} as for loam, 5.23E-6 m s^{-1} as for sandy loam, 1.41E-5 m s^{-1} as for loamy sand), and a different set of SLOPE values was assigned based on terrain slope and geology".

Methods, section 4.4 WRF-Hydro calibration and validation with observed precipitation, Line 296-297: "SLOPE parameters were assigned using the slope terrain class map allowing the best performance during sensitivity."

Results, section 5.1 sensitivity analysis, Line 344-349: "More sensitive than OVRGH is Ks, suggesting a possible important impact of the soil type and property definitions on the model output. Senatore et al. (2015) presented one of the few WRF-Hydro studies that calibrated a hydraulic conductivity related parameter, although they focused on the saturated soil lateral conductivity. SLOPE appeared to have a low sensitivity, although in the mountain watersheds, where it changed, a small reduction in the total discharged volume was observed".

Results, section 5.2 WRF-Hydro calibration and validation, Line 363-365: "SLOPE attributed based on both terrain slope and geology resulted in slightly better performance indices in the mountain watersheds than SLOPE attributed through terrain slope only. Therefore, it was selected for the final parameterization".

Results, section 5.2 WRF-Hydro calibration and validation, Line 380-389: "The parameterization of watersheds Ma, An, Pl, Ka, and At is peculiar. These watersheds are mainly characterized by sandy loam texture (i.e., higher Ks than the other watersheds), maximum deep drainage obtained by using the SLOPE parameters based on slope terrain and geology, very high REFKDT values, and very large groundwater storage. However, poor model fit indices (for some watersheds even negative) were obtained for the calibration period (Fig. 5). Conversely, the same watersheds show positive *NSE* values and negative *PBIAS* (i.e., slight underestimation of the peak discharge), for the validation event. Overestimation of runoff in Jan 1989 could have been related to the modeling of snow and snowmelt in the LSM. Both observed and modeled temperature values for the upstream areas of these watersheds showed negative values, indicating that part of the precipitation was snow".

Conclusion, Line 533-534: “Modifications of deep drainage coefficients and MODIS soil types based on geology reduced the peak flow overestimation by up to 40% in watersheds characterized by a fractured and very permeable bedrock”.

Conclusion, Line 539-544: “Negative *NSE* values were found in three watersheds located at high elevation where an underestimation of the snow fraction, computed by the LSM, may have occurred. Modelled snow height, and possible improvements deriving from the use of alternatives routines (e.g. Noah MP), should be checked with observed snow depth data, which were not available for this study”.

Conclusion, Line 559: “Soil properties could be specifically calibrated for the study area”.

13. LL335-339 and Figs. 5-6: the Y scale for watershed Mk is not appropriate (much higher maximum value than needed). The comment about watershed St does not correspond to what I can see in the Figures.

    We modified the Y-scale of Mk in all figures and the comments related to both watersheds as follows.

    Results, section 5.2 WRF-Hydro calibration and validation, Line 404-409: “Mk is the only watershed showing higher rainfall and flow peaks towards the end of the Jan-1989 event rather than in the middle. The model slightly underestimates the flow peak occurred on January 9th and overestimates the flow at the end of the simulation period. For St, the model reacts sharply to precipitation input, simulating well the flow peak occurred on January 9th but overestimating the flow at end of the simulation period of the Jan-1989 event and above all the peak of the Nov-1994 event, therefore affecting the performance scores”.

14. L343: [this comment refers to the main comment about the groundwater bucket model] For Ak, it’s not a problem of baseflow, but of recession, which is typically a problem concerning especially interflow (i.e., quicker contribution than baseflow).

    Our bedrock is very fractured without a continuous groundwater table and we have predominantly shallow soils. It is difficult to distinguish between interflow and baseflow. We have observed slow dripping from the bedrock into upstream channels after large rainfall events. We also have streams that discharge to the bedrock with streamflow again recurring further downstream. Thus, we do have a streamflow recession made up of a combination of processes. As noted in general comment nr. 2, the OVRGH parameter influences the total discharged volume but not the shape of the hydrograph. Therefore, to better fit the post-peak shape of the hydrograph, we focused on baseflow calibration. To monitor the baseflow effect, we added four figures as supplementary material (Fig. S1 – S4), in which we showed the hydrographs for all watersheds together with the baseflow contribution, for both events and both observed and modelled rainfall as forcing. Fig. S1 and Fig. S3 show hydrographs for Jan-1989 event forced with observed and modelled rainfall, respectively. Fig. S2 and Fig. S4 show hydrographs for Nov-1994 event forced with observed and modelled rainfall, respectively. To incorporate these analyses, we modified the manuscript as follows.

Methods, section 4.2 WRF-Hydro Parameterization, Line 255-258: "Other general parameters are REFKDT and soil depth (SD), which were calibrated. REFDK was left to its default value (2.00E-6 m s^{-1}). The WRF-Hydro parameter OVRGH was tested and values were assigned based on the sensitivity analysis, whereas RTDPT was kept constant all over the study area and a value of 1, consistent with a steep mountainous terrain, was assigned".

Methods, section 4.3 WRF-Hydro Sensitivity analysis, Line 271-277: "In addition, sensitivity runs for the OVRGH parameter and the saturated hydraulic conductivity (K_S) were performed, too. For these simulations, the baseflow routine was switched off. A reference scenario was set, with REFKDT and OVRGH equal to 1, SD equal to 1.0 m, K_S equal to 2.45E-6 m s^{-1} (value attributed to clay loam soils in the soil parameter table), and the deep drainage parameter (SLOPE) assigned based on terrain slope, as in Silver et al. (2017). Parameters were changed one at a time. Eight values were tested for REFKDT (0.3, 0.5, 3.0, 5.0, 8.0, 10.0, 100.0, 1000.0), two for SD (0.5 and 2.0 m), two for OVRGH (0.1, 0.5), three for K_S (3.38E-6 m s^{-1} as for loam, 5.23E-6 m s^{-1} as for sandy loam, 1.41E-5 m s^{-1} as for loamy sand), and a different set of SLOPE values was assigned based on terrain slope and geology".

Methods, section 4.4 WRF-Hydro calibration and validation with observed precipitation, Line 297-309: "REFKDT and OVRGH were initialized, in each watershed, based on the evaluation of the sensitivity runs through performance indices, as for SD. For the baseflow bucket routine, initial values of a and Z_{max} were set to the default. Next, the initialized parameters were fine-tuned based on a trial and error procedure for all watersheds. Modifications were applied to a single parameter at the time and if changes could not improve the model performance according to three indices out of five after five attempts, the parameters were retained. Commonly applied changes were ±1 for REFKDT, ±0.1 for OVRGH, ±0.5 for a, and ±10% of the actual value for Z_{max}. Smaller (larger) changes were applied only in watersheds where the response of streamflow was (not) particularly sensitive to specific parameters. The parameterization of Z_{max} was aimed at filling the reservoir after the rainfall peak, between 10 January at midnight and 11 January at noon, to simulate the observed recession of the hydrograph. For those watersheds that highly overestimated the baseflow due to spilling out of the groundwater reservoir, Z_{max} was further increased. A good fit between observed and simulated flow before the peak was the target for the calibration of the exponent α".

Results, section 5.1 sensitivity analysis, Line 344-348: "Regarding OVRGH, results show that it has a slight control on the total volume discharge, as also presented in Yucel et al. (2015), while it has almost no effect on delaying the peak (Fig. 4). More sensitive than OVRGH is Ks, suggesting a possible important impact of the soil type and property definitions on the model output. Senatore et al. (2015) presented one of the few WRF-Hydro studies that calibrated a hydraulic conductivity related parameter, although they focused on the saturated soil lateral conductivity".

Results, section 5.2 WRF-Hydro calibration and validation, Line 365-368: "Also, for all watersheds OVRGH was set equal to 1 because it was the value returning the best performance indices in 19 out of 22 watersheds. Furthermore, considering that OVRGH effects total discharge volume and not hydrograph shape, its calibration would have been equifinal to REFKDT".

Results, section 5.2 WRF-Hydro calibration and validation, Line 418-423: "As it is visible in Fig S1 and Fig S2, flow in the receding limb of the hydrograph is mainly made up of baseflow. For Jan-1989 event, in all these watersheds the groundwater reservoir is filled up on January 10th and baseflow consists of the water spilling out from it. This water volume, redistributed along the channel network, is generally able to reproduce the hydrograph shape, except in Ak. In Nov 1994, no groundwater spilling is observed during the simulation and the receding limb is underestimated. Therefore, this could be partly due to a non-perfect reproduction of the model initial conditions and partly related to an underestimation of interflow and baseflow".

Conclusion, Line 535-537: "The overland roughness routing factor reduced the streamflow but showed a very limited effect on delaying flow. A straightforward calibration of the baseflow reservoir based on low flow fitting (exponent) and reservoir filling time (maximum capacity) was a good mean for obtaining a reasonable simulation of the hydrograph recession in most watersheds".

Conclusion, Line 560-561: "For a continuous, long-term streamflow analysis, an evaluation of the sensitivity of the baseflow reservoir parameters could be carried out.".

15. L344: the peak looks not so well simulated in Ak

We agree. We have modified the manuscript as follows.

Results, section 5.2 WRF-Hydro calibration and validation, Line 410-418: "In the eastern part of the modelling domain (La to Ni), for the calibration event both initial baseflow and the discharge peak are well modelled in all watersheds (Fig. 6). Differences between observed and simulated hydrographs can be observed in the post-peak, for watersheds Ak, Pe (Fig. S1), Ko and Ni. Ak and Pe present a very high peak flow (> 50 m^3 s^{-1}) and an underestimation of the receding limb of the hydrograph in the following days, which causes the negative *PBIAS* and high *MAE* values visible in Fig 5. In the case of Ko and Ni, the receding limb shows a little overestimation. For the validation event (Fig. 7), the peak is well simulated in Pe and Ao, slightly overestimated in Ak and Pd, underestimated in La, Vy, Ko, and Ni (Pe and Pd, Fig. S2). In the post peak phase, the simulated hydrographs show negative biases in comparison to the observed ones in all watersheds".

16. L368: passing -> moving?

Modified as suggested.

17. LL372-374: these sentences are confusing, especially if compared with LL351-353, which seem to refer to the same comparison. Not clear what the authors mean when they state that bias "on average increased by 8.6 times"

The first lines described rainfall, while the second group described streamflow. Throughout section 5.3 we have now modified the text so that it is explicitly said if the performance indices refer to precipitation or streamflow. We introduced PBIAS as a replacement of BIAS so we modified the unclear sentence.

Line 461-463: "The absolute value of flow *PBIAS* decreased in seven watersheds (Af, Li, Pl, Vy, Ak, Ko, Ni) but on average increased by 21.5% (96.6% in Pg and 120.3% in Le)".

18. L395: the three watersheds

Thanks for spotting it. Changing some parameter during calibration the watersheds became four (**Line 495** in the manuscript with track changes).

19. L400: decent -> reasonable? Besides, again I don't think it's a matter of baseflow

We modified the text discussing the receding limb of the hydrograph in general and not baseflow only.

Results, section 5.4 WRF-Hydro with observed and modeled precipitation evaluation at hourly scale, Line 500-503: "In addition, the receding hydrograph is well modelled for the calibration event but not so well for the validation event. This result is similar to what was observed for daily streamflow and was attributed to the possible non-perfect reproduction of the model initial conditions and underestimation of interflow. The fairly good post-peak simulations lead to reasonable hourly performance indices for the Jan-1989 event.".

20. L412: probably, increasing overland roughness coefficient could be also a way for improving interflow and, therefore, the simulation of the falling limb of the hydrograph

Please refer to answer to previous comments regarding overland roughness and interflow (general comment 2, minor comments 14, 15, 19).

21. LL442-443: please contextualize better this sentence

We have modified the sentence.

Conclusion, Line 551-557: "This suggests that model calibration with modelled rainfall forcing is not optimal for small mountain watersheds and should be carefully evaluated if no other options are available. As a consequence, WRF rainfall forecasts may not be sufficiently accurate for predicting the location and size of specific floods of such small mountain watersheds. However, due to the relatively small errors in total precipitation (average relative difference over the 22 watersheds of 17% and for 20% Jan 1989 and Nov-1994 events, respectively) and simulated daily maxima (average relative difference over the 22 watersheds of 22% and 18% for Jan 1989 and Nov-1994 events, respectively), modelled rainfall data could be suitable for investigating the effect of climate change on extreme rainfall and flood events".