Anonymous Referee #2

This paper presents a modeling work on 22 small watersheds using WRF-hydro. The model is forced with modeled WRF data and observe precipitation for 2 periods in January 1989 and November 1994. The authors concluded that using WRF precipitation may not be suitable for hydrological studies in small mountain basins although they are still useful for long term studies.

General comment

1. I am not quite sure what is the research question that the authors are trying to answer. If it is WRF-hydro ability to simulate streamflow, I think that is widely covered in the literature review, but it may be important a benchmarking in this specific area. If it is the advantage of using observed precipitation, I think that it is not necessary to write a paper about it since it is well known that if there are observations available, it is better to use them over modeled data or to correct the modeled data. Therefore, it is not clear to me what is the actual contribution that the authors are trying to deliver. The problem with modeled precipitation is because the WRF model does not work? Or the modeler did not implement it correctly? When the model and observe precipitation does not compare well, why did you use it anyway? I think that using incorrect input will certainly result in poor performance. But jumping from there to conclude that we should not use modeled precipitation is a big stretch. I think that the paper should focus on the performance of WRF-hydro. The model performance is not affected if the precipitation is observed or modeled since you are using it uncoupled.

   We used the best possible approach to model convective rainfall events and the results obtained show a good agreement with the observed fields. They compare well. Therefore, we consider our modelled rainfall input as a sub-optimal input. Although the high quality of it, still the small errors in the rainfall, in such small watersheds, propagate in the streamflow deeply affecting the performance. We have modified the introduction and the conclusion to state in a clear way that the value of the results is limited to small watershed (below 100 km^2) and that model calibration carried out with modelled data is not optimal, not that it shouldn't be performed at all. Also, we suggest that WRF rainfall forecasts may not be sufficiently accurate for predicting the location and size of the floods of such watersheds thinking of implementing a similar system as an operational flood forecasting tool.

   Introduction, Line 87-89: "Model performance loss due to differences between observed and modelled rainfall is rarely discussed. Also, little attention has been given to small watersheds (area below 100 km^2), which are often ungauged and prone to flash floods. This study aims to address this gap".

   Conclusion, Line 523-527: "This study evaluates streamflow simulations of the one-way coupled atmospheric-hydrologic model WRF-Hydro, forced with observed and WRF-modeled rainfall, during two extreme events, over 22 small mountain watersheds in Cyprus (area below 100 km^2). Following model calibration and validation with observed rain, the model was run with WRF-downscaled (1 × 1 km^2) re-analysis precipitation data (ERA-Interim). These forcing data represent best-performing hindcasts of two extreme rainfall events, i.e. a model product that is as similar as possible to reality and considered sub-optimal".

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".

Specific Comments

1. First line abstract: "Few studies evaluate the hydrologic performance of coupled atmospheric-hydrologic models when forced with observed rainfall and even fewer when forced with modeled precipitation." This is not quite true, there is extensive literature on this topic. If you are specifically referring to WRF-hydro you should state that.

   We have modified the first sentence of the abstract 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".

2. The first paragraph of the introduction is related to the land-atmosphere feedbacks, which is not the case of this study, so I suggest eliminating it or to refocus it to the topic of the paper

   We have refocused the paragraph to the topic of the paper by adding a sentence at the end of it.

   Introduction, Line 39-40: "However, recently authors have started to see these systems as instruments for flood forecasting, making the performance of the hydrologic routines a key indicator of the model quality (Givati et al., 2016; Maidment 2017)".

3. What about the bucket model parameter sensitivity? There is no indication of the interaction of those in the uncertainty analysis.

   Yes, good point. We conducted the sensitivity analysis on the parameters mainly influencing the rainfall runoff-infiltration partitioning (including few sensitivity runs regarding the overland roughness coefficient that we added during the review), with the baseflow bucket model switched off. For the calibration, we found that we could improve the simulations with a straightforward tuning of the baseflow parameters based on its filling time and the fit of the pre-peak hydrograph. We agree that a sensitivity analysis of the baseflow parameters is useful, but it would be more suitable to do this for a continuous, long-term streamflow analysis. We have added a recommendation about this in the conclusions.

   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".

4. Is there any evaluation of precipitation disaggregation eq 1 and 2?

We derived hourly fields with the presented disaggregation method and with simple IDW interpolation, since the method worked best for daily local events (Camera et al., 2014). We forced WRF-Hydro with both datasets and obtained a better fit of streamflow with the disaggregated one. In addition, looking at the 5-day cumulated rainfall fields calculated from the two hourly datasets, for the days around the peak precipitation for the two events of interest, we noticed a more plausible areal distribution for the disaggregation method (Fig. 1R). In addition, this method allows to preserve the mass balance between the daily and the hourly dataset. We did not include this explanation in the manuscript for conciseness.

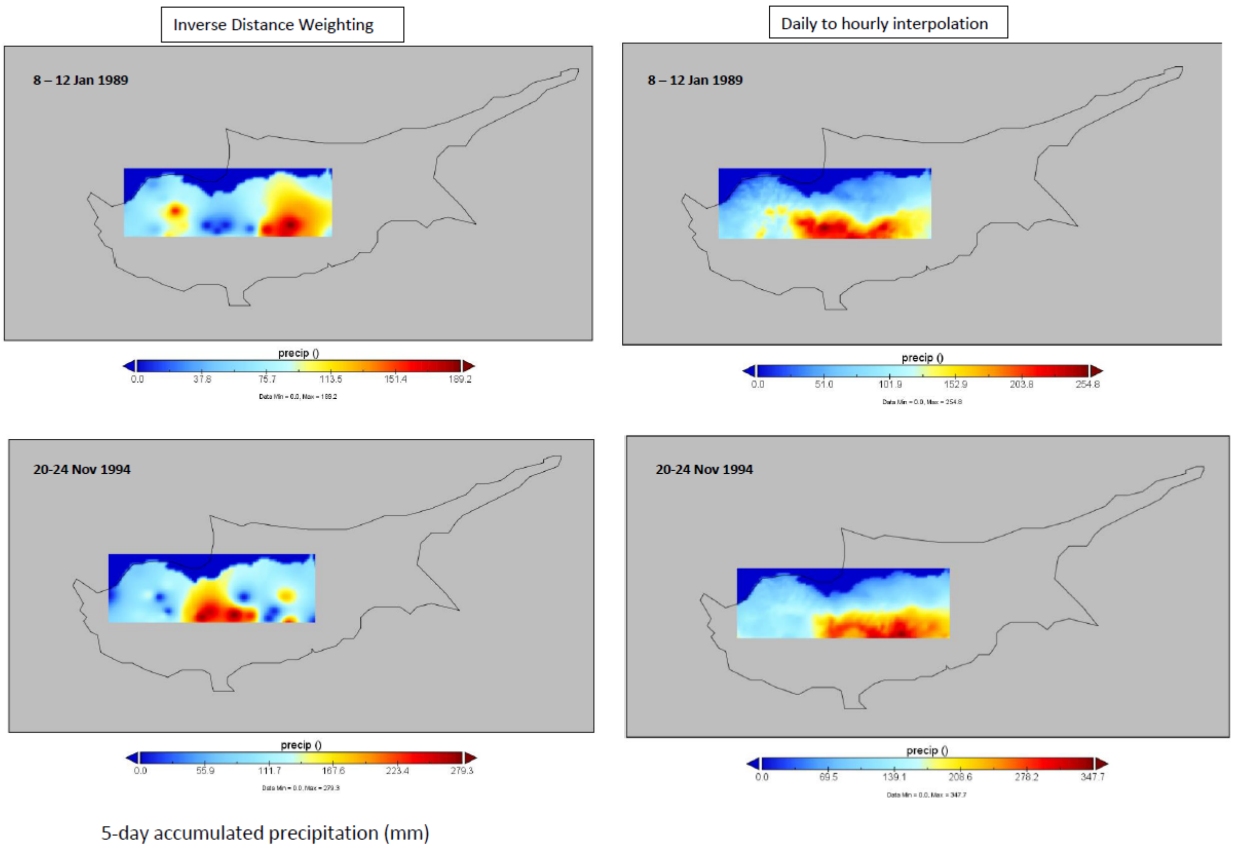

Fig. 1R: comparison of 5-day cumulated rainfall around the precipitation peak of the two events of interest (Jan 1989 and Nov 1994) obtained with two different interpolation methods, Inverse Distance Weighting (IDW) and disaggregation of daily to hourly values.

5. In Figure 4, there a series of inconsistencies between the performance of the model between the calibration and evaluation period. Is there any explanation for that? In particular basins, Ma, An, Pi, Ka, where you have really bad results during the calibration but still validate the models and got very good results.

This relates to some of the comments Reviewer #1 made, too. For us, it is partly related to the geological characteristics of those watersheds and partly to the fact that they are located at high elevation and part of the precipitation during Jan-89 event occurred as snow. We managed to slightly improve the hydrograph simulation of these watershed (still they are not optimal) modifying deep drainage and soil properties based on geology. We have modified the 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 a maximization of the drainage from the soil column to the groundwater bucket. Also, we modified the soil type of the area from clay loam (MODIS database) to sandy loam, based on field evidence. The overestimation of the peak got reduced up to 40% but still an overestimation remained. The same model parameterization results in positive NSE and negative BIAS for the Nov-1994 event. This combination of results led us to believe that the main issue is an underestimation of the snow. Please refer to the answers to the general comment 3 and specific comment 12 of Reviewer #1 for details on the added analyses and manuscript modifications.

6. Also, it would be better to use percentage bias instead of bias alone to have a more general indicator of biases.
   We modified the manuscript accordingly.

7. Figures 5 and 6, it is very hard to see the differences in precipitation. And then, why to use incorrect model precipitation at all.
   We changed the color of modelled rainfall, we hope figures are clearer than in the previous version. Figures numbering changed because we added an extra figure to discuss the sensitivity analysis results. The mentioned figures are now Fig. 6 and Fig. 7.
   Regarding rainfall, in our intention we do not use incorrect rainfall, we use the best available modelled rainfall. Please refer to the answer to your general comment 1.

8. If we look at figures 7 and 8, we can see that WRF-hydro does a really bad job (Figure 8) in basins where WRF precipitation has good performance (Figure 7), can you explain why?
   Fig. 7 and Fig. 8 are now Fig. 8 and Fig. 9. WRF-Hydro forced with modelled rainfall seems to poorly simulate especially watersheds Af (Nov 1994), Pg (both events), Le (both events), Mk (Jan 1989), Li (Nov 1994), An (Jan 1989), Pl (Jan 1989). For An and Pl, we believe that the problem is the same as for observed rainfall, therefore related partly to the difficult parameterization of a highly fractured bedrock and partly to the high elevation causing snow. Watershed Af, for Nov-1994 event show a very high rainfall PBIAS. Watershed Le shows rather low rainfall NSE for both events and a very high rainfall PBIAS for Jan-89 event. Watershed Li show a medium-high MAE for Nov-1994. Watersheds Pg and Mk are those characterized by the lowest average discharge during both events. Therefore, small discharge variations cause higher performance loss for them than for all other watersheds. They are the perfect exemplification of what we wrote in the manuscript about the small shifts in the space-time rainfall fields causing important performance losses. Figures S3 and S4 in the supplementary material show it. We have modified the manuscript as follows to stress this point.

   Results, section 5.3 WRF-Hydro simulations with modeled precipitation, Line 470-475: "These results indicate that a small shift in time or space of modelled rainfall, in comparison to observed precipitation, can strongly modify the hydrologic response of small watersheds to extreme events. This is particularly evident in watersheds Pg and Mk, which are among the smallest and those characterized by the lowest average discharge in both events (Fig. 6, Fig. 7, Fig. S3, Fig. S4). Although their rainfall performance indices (Fig. 8) do not show particularly large errors (except a negative NSE for Mk in Nov 1994), streamflow fit indices present very negative values and streamflow PBIAS is very high as well (Fig. 9)".

9. In the conclusion you mention this "Streamflow obtained with WRF-modelled rainfall forcing showed high discrepancies with observations, despite the good agreement between modeled and observed precipitation (average NSE of 0.83 and 0.49 for Jan1989 and Nov 1994, respectively)." Did you try to calibrate the model with the modeled precipitation data and evaluate the observation?
   This is a good point. We thought about it during the design of the methodological approach but then we preferred to focus on the performance loss passing from a calibration on observed rainfall to a simulation with modelled rainfall. Also, we are aware that the observed gridded dataset could be characterized by errors but the high density of stations and the study done on the creation of the daily dataset, including the evaluation of the final product (Camera et al., 2014), gave us some confidence about the quality of the input data.

References not present in the manuscript

Simmons, A, Soci, C, Nicolas, J, Bell, B, Berrisford, P, Dragani, R, Flemming, J, Haimberger, L, Healy, S, Hersbach, H, Horányi, A, Inness, A, Munoz-Sabater, J, Radu, R, Schepers, D, 2020. Global stratospheric temperature bias and other stratospheric aspects of ERA5 and ERA5.1. ECMWF Technical Memoranda 859, Reading (UK), 40 pp.