Response to the comment of L. Samaniego (Referee)

We would like to thank the reviewer for his positive and insightful comments on the manuscript. Below is our response to the issues raised in the review. The original comment is printed in plain font, our response is printed in italics.

This manuscript is based on the presumption that the combination of statistical analysis, process-based modeling using climate and stochastic projections as well as expert judgement is the best way to assess climate impacts on low flows. Without any further analysis, one could dare say that this premise should be true considering that this approach has much more information than any single analysis and thus should have less chance of not finding an answer that is closer to the true one. The authors strive to demonstrate the advantages of the proposed approach and the validity of this premise with a regional study conducted in four Austrian river basins. The manuscript is well written although it is a bit too long in my opinion. The topic of the study is relevant for HESS but the manuscript requires a substantial revision before publication. Below, I provide a number of issues to be clarified before publication.

We would rephrase the above statements in saying that the three pillar approach is a plausible way to assess climate impacts (not necessarily the best as we do not compare it with other approaches) and that we strive to demonstrate the usefulness of the premise rather than its validity, as validity can never be demonstrated for the future. We have now removed Figure 1 which may have been suggestive of the claim of a "best method".

• My first remark refers to the terminology chosen for this manuscript. My impression after reading the abstract and the introduction is that the names given to the various methods and the proposed "three-pillar" approach can be considerably simplified without diminishing the message that the authors try to convey. On the contrary, it will help the reader. I wonder, for example, what a data-based method has to do with a downward approach (downward refers to "toward a lower place, point, or level")... and conversely a mechanistic one with an upward approach ... I know that these terms have been used in current literature, but in my opinion, these buzzwords can be replaced by method A and B without changing the meaning of the sentences. I suggest either to justify the meaning of "downward" and "upward" in the present context or even better, to simplify the text. In my opinion, the so-called "downward approach" is a classical statistic method, so I wonder why not calling it simply like that.

The terminology of upward and downward approaches (Sivapalan et al., 2003) reflects the alternative avenues towards obtaining understanding of how a system operates which is unrelated to whether the methods are statistical or deterministic. The upward or mechanistic approach is based on a preconceived model structure that puts conceptual components such as runoff generation together (hence upward), while the downward approach infers the catchment functioning from an interpretation of the observed response at the catchment scale (fingering down to smaller scales, hence downward). We realise there are subtleties involved and the terminology is not essential for the paper, so we have removed it.

• In this study, old IPCC nomenclature for emission scenarios (A1B, B1, A2 etc) are still used instead of the newer RCPs proposed by the IPCC. Newer climate projections (e.g., CMIP5) are readily available for quite some time. Please explain why.

Jacob et al. (2015) showed that the most recent regional climate simulations over Europe, accomplished by the EURO-CORDEX initiative (RCPs, Moss et al., 2010), are rather similar to the older ENSEMBLES simulations with respect to the climate change signal and the spatial patterns of change. For consistency with related studies in Austria (e.g. Parajka et al., 2016) we have therefore chosen the older emission scenarios. We are now noting this in the manuscript.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P., and Wilbanks, T. J.: The next generation of scenarios for climate change research and assessment, Nature, 463, 747–756, 2010.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., Kröner, N., Kotlarski, S., Kriegsmann, A., Martin, E., Meijgaard, E., Moseley, C., Pfeifer, S., Preuschmann, S., Radermacher, C., Radtke, K., Rechid, D., Rounsevell, M., Samuelsson, P., Somot, S., Soussana, J.- F., Teichmann, C., Valentini, R., Vautard, R., Weber, B., and Yiou, P.: EURO-CORDEX: new high-resolution climate change projections for European impact research, Reg. Environ. Change, 14, 563–578, doi:10.1007/s10113-013-0499-2, 2014.

• Authors do not formulate in the introduction a research hypothesis to be tested. I guess, the authors intend to test that the “Three-pillar approach” is superior than any of the single ones, but failed both to explicitly mention this hypothesis and to present statistic evidence that corroborates this assertion.

Actually, we are not intending to test a hypothesis in this paper. The aim of the paper is to present an approach to assess climate impacts on low flows from different sources of information. The objective is twofold, to present the concept and to illustrate the viability of the approach. A hypothesis that the three pillar approach is superior to any of the single methods would be testable in a synthetic world (where the future is generated and assumed to be perfectly known) but this would probably be a rather trivial exercise. The real world is more complex, so we confine ourselves to illustrating the feasibility of the approach very much in the spirit of ensemble predictions. We are now making the underpinning philosophy of ensemble predictions more explicit in the paper.

• L19, P9. If a hydrologic model is used in this study, I do not understand why a runoff index is not used instead of a meteorological drought index like SPEI. Streamflow, and thus low flow characteristics, are the outcome of the whole hydrologic system that is represented by a hydrological model. Moreover, it is well documented in the literature that atmospheric drought indices are quite transient whereas those related to soil moisture, groundwater, and runoff are not (Samaniego et al JHM 2013 and sources therein). Thus, the stochastic dependence of SPI or SPEI with any low-flow index is, in general, not significative (Kumar et al. 2016 HESSD). It should also explained why a Gaussian transformation (perhaps due to a long tradition...) should be applied a variable than is definitely non-Gaussian (i.e., P □Ep). L14 P9. A more reliable approach to “check the realism” of the ensemble climate simulations would be to estimate a runoff index over a historial period in which reanalysis (or hindcasts) and historial meteorological forcings are available. This is probably the best way to know whether a RCM or a Numeric Weather Prediction Model output can explain observed low-flow spells or other kinds of drought events as proposed by Thober et al. 2015.

We agree that a number of methods can be used for testing the realism of ensemble climate simulations (and we find the methods suggested by the reviewer useful), but the jury is probably still out on what is the most suitable method in a particular hydro-climatological setting. Kumar et al. analyse groundwater anomalies rather than low flows, so their results are not fully applicable to the present case, while Haslinger et al.. (2014) did find significant links between SPEI and low flows in the study area. The SPEI has been adopted here for its simplicity and because it can be calculated from the HISTALP data (Auer et al., 2007) back to the year 1800. Given this is a side issue in the paper, in our opinion, comparing different methods would go beyond the scope of this paper. The hydrological modelling later in the paper allows a more detailed comparison in the spirit of the references suggested by the reviewer. We now give an explicit justification of the use of SPEI.

• L18 P.5 It is not clear to me why the “first and second pillars” do not use local information used in the third pillar. After all, trends are based on local meteorological observations and any rainfall-runoff model, to my knowledge, uses local observations of rainfall, temperature, and discharge. Please elaborate why they have to be different (L22)?

We appreciate the comment as the wording has indeed been lacking clarity. The first two pillars do not use observed changes in the stochastic rainfall characteristics while the third pillar (stochastic extrapolation) does. We have reworded the sentence for clarity.

• L17 ff, P5. I guess authors demand too much from downscaled GCM-RCM forcings. GCM and RCM are climate models describing the evolution of physical processes in the atmosphere, ocean, cryosphere and land surface at large temporal and spatial scales (about 2.5_). They are not intended to describe transient states, consequently one can not say that they are reliable or not. They do not have all the process necessary to describe rainfall generation at smaller scales like high resolution numerical weather models have if they are run at 1 km to 2 km spatial resolution. RCMs at 1/4_ resolution and larger would be hardly able to estimate convective precipitation over mountainous areas like Austria. For GCMs, this is almost an imposible job. If this is known, I wonder why the hydrology comuntiny insist on getting “reliable” daily precipitation (say from RCMs inreanalysis mode) from these models so that low-flow statistics can be estimated ... Dynamic and stochastic downscaling may help a bit but many studies have shown, for example, that very few RCMs from the ENSEMBLES project are even able to get extreme statistics of the observed rainfall fields at monthly time scales (see e.g., Soares et al. 2012 JGR in Portugal, and Thober & Samaniego JGR, 2014 in Germany). As a consequence, low-flow statistics and its variability (e.g., Q_{95}) obtained from reanalysis (e.g., WATCH) should be evaluated as expectations over reasonable periods (e.g., over decades). Likely yearly statistics are too short a period. See for example Schewe, J. et al. as an alternative.

We fully agree with the remark that RCM outputs should be assessed at time scales longer than a year and we did not intend to convey the impression that individual years should be taken at face value. In the discussion we are now making it clearer that the focus is on decadal rather than yearly scales and this is how the figures should be interpreted.

• L13 p8. The area of the river basins and the sampling size used in this study are probably too small to derive conclusive results. Authors should consider that the area of a GCM grid cell like ECHAM5 is at least $9 \_ 10^4$ km^2 and that of a RCMs used in Reclip:century is approximately $1 \_ 10^2$ km^2 (based on the project report). As a rule of thumb, due to the Courant–Friedrichs–Lewy condition, it is not recomendable to use prognostic values of state variables or fluxes obtained by numeric integration for areas less than four times the area of a typical grid cell. This implies that the minimun area to be consider in this case is a basin with at least $4 \_ 10^2$ km^2. Three of the study areas do not fulfill this condition. As a result, the uncertainty of the numerical model plus that of the downscaling techniques would increase dramatically which, in turn, would negatively affect the impact analysis. I recommend to test this approach in large basins that fulfill this condition and to enlarge the sample size considerably.

Yes, the spatial scales of applicability of RCM simulations is on the order of hundreds of km^2. This is exactly the reason why we put the smaller catchments into a regional context (Figure 3, now Figure 2). This was acknowledged by reviewer #1: "the paper also works with a large dataset condensed to a few representative examples ... that ensure that patterns are not emergent from a few preselected sites or times." As suggested by the reviewer we are now making the scale considerations of the climate simulations more explicit in the manuscript with respect to Figure 3, now Figure 2.

• L15 P11, I suggest to use a non-parametric test to estimate confidence bounds considering that the underlaying variable is certainly non-Gaussian. In this case, parametric t-Student estimations for confidence bounds do not apply.

This is a good point. We therefore reanalysed the data by a nonparametric approach based on bootstrapping to estimate distribution-free confidence intervals. The results are given in supplement A of this response. The bootstrap distributions of predicted values turn out to be very close to Gaussian so the results change very little. The expected changes never differ by more than 4% from those of the method used in this paper, and their 95% confidence bounds never differ by more than 21% (period 2021-2050) and 33% (period 2051-2080) from those of this paper. However, we do see the value of the nonparametric approach and have adopted it therefore in this paper, replacing the Gaussian approach in the original manuscript.

• The structure of the manuscript is cumbersome in some sections. I suggest that methods and results from every approach is presented separately to easereading. The number of sections is quite large for a research paper in my opinion. This manuscript is a bit long too.

In response to this comment we have reorganised the paper, merging the methods sections into one chapter and condensing the entire manuscript by about 30%.

• L31, p19. Authors do not attempt to estimate “how strongly the pillars agree”. It will be very enlightening to see a statistical analysis in this respect.

We appreciate the idea and have added a figure (now Fig. 11) showing the probability density functions (pdfs) of the low flow projections from the three methods for the period 2021-2051. We have tested the consistency of the pdfs by a two-sample Kolmogorov-Smirnov test which, however, gives lack of significant agreement for most cases which does not provide a lot of insight. We have therefore chosen to limit the quantitative comparison to the new figure.

• L2 ff p 26 As I said earlier, I have no doubt of this statement. In general, more information should lead to more reliable results. I do not see novelty on this statement. This can be inferred, for example, from simple parametric statistical tests by gradually changing the sampling size and estimating the effect on the confidence bounds for a given statistic. L29 ff is a consequence of this. Authors should present results and make statistical tests that demonstrate with large degree of certainty that adding information gradually leads to better results in this case. I have, however, reservations, on how soft data (e.g. historical reports), or subjective impressions can be used in a formal statistical analysis to “correct” confidence bound.

We agree that, to some degree, more information leading to more reliable results is an obvious statement. On the other hand, this is exactly the basis of multi-model ensemble projections. We have now changed the tone of the presentation in order not to imply that the use of more information is novel, rather the particular implementation in the context of low flow projections. Of course this can be formalised, for example by Bayesian methods that can handle subjective information (eg. Viglione et al., 2013) but this would go beyond the scope of this paper.

• Fig 11 is quite dense. It is supposed to be a synthesis, but I hardy can understand it. Sorry. In my opinion, this manuscript could become a nice contribution to the field if these issues are addressed before publication.

While reviewer Luce did note that the graphics of the paper are well constructed we can see the point here. To assist in the interpretation we have added a new figure (now Fig. 11) which is simpler and more clearly demonstrates the similarities and differences of the pillar projections.

Luis Samaniego

References

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., et al. (2014). Multimodel assessment of water scarcity under climate change. Proceedings of the National Academy of Sciences of the United States of America, 111(9), 3245– 3250. http://doi.org/10.1073/pnas.1222460110

Thober, S., & Samaniego, L. (2014). Robust ensemble selection by multivariate evaluation of extreme precipitation and temperature characteristics. Journal of Geophysical Research-Atmospheres. http://doi.org/10.1002/(ISSN)2169-8996

Soares, P. M. M., R. M. Cardoso, P. M. A. Miranda, P. Viterbo, and M. Belo-Pereira (2012), Assessment of the ENSEMBLES Regional Cli- mate Models in the representation of precipitation variability and extremes over Portugal, J. Geophys. Res., 117(D7), D07114, doi:10.1029/2011JD016768.

Samaniego, L., Kumar, R., & Zink, M. (2013). Implications of Parameter Uncertainty on Soil Moisture Drought Analysis in Germany. Journal of Hydrometeorology, 14(1), 47–68. http://doi.org/10.1175/JHM-D-12-075.1

Thober, S., Kumar, R., Sheffield, J., Mai, J., Schäfer, D., Samaniego, L. (2015). Seasonal Soil Moisture Drought Prediction over Europe Using the North American Multi-Model Ensemble (NMME). Journal of Hydrometeorology, 16(6), 2329–2344. http://doi.org/10.1175/JHM-D-15-0053.1

R. Kumar, J. L. Musuuza, A. F. Van Loon, A. J. Teuling, R. Barthel, J. Ten Broek, J. Mai1, L. Samaniego, and S. Attinger, (2016). Multiscale evaluation of the standardized precipitation index as a groundwater drought indicator, HESSD. http://www.hydrolearth- syst-sci-discuss.net/12/7405/2015/hessd-12-7405-2015.pdf

SUPPLEMENT A

Original CI

Table #2 Trend projections FOR MID OF PROJECTION PERIOD 2035 for (2021-2050) and 2065 for (2051-2080)

	Hoalp	Muhlv	Gurk	Buwe
Predicted discharge 2050 (m^3/s)	0.28 m^3/s (0.19, 0.38) m^3/s	0.67 m^3/s (0.36, 0.97) m^3/s	1.17 m^3/s (0.48, 1.87) m^3/s	0.02 m^3/s (-0.10, 0.14) m^3/s
Change 2050 (%)	+42% (-5, +88)	-10% (-51, +32)	-36% (-74, +1)	-89% (-156, -21)
Predicted discharge 2080 (m^3/s)	0.35 m^3/s (0.20, 0.51) m^3/s	0.58 m^3/s (0.07, 1.09) m^3/s	0.74 m^3/s (-0.42, 1.90) m^3/s	-0.08 m^3/s (-0.29, 0.12) m^3/s
Change 2080 (%)	+78% (1, 156)	-21% (-91, +48)	-60% (-123, +3)	-145% (-258, -33)

BOOTSTRAPED CI (5000 replications)

Table A.2 Trend projections FOR MID OF PROJECTION PERIOD 2035 for (2021-2050) and 2065 for (2051-2080)

Table 2

	Hoalp	Muhlv	Gurk	Buwe
Predicted discharge 2050 (m^3/s)	0.28 m^3/s (0.19, 0.37) m^3/s	0.68 m^3/s (0.45, 1.02) m^3/s	1.19 m^3/s (0.58, 2.00) m^3/s	0.02 m^3/s (-0.14, 0.14) m^3/s
Change 2050 (%)	+39% (-7, +71)	-8% (-41, +34)	-36% (-72, -1)	-90% (-177, -22)
Predicted discharge 2080 (m^3/s)	0.35 m^3/s (0.22, 0.45) m^3/s	0.60 m^3/s (0.15, 1.14) m^3/s	0.74 m^3/s (-0.23, 2.01) m^3/s	-0.08 m^3/s (-0.33, 0.12) m^3/s
Change 2080 (%)	+74% (0, 123)	-21% (-79, +51)	-59% (-113, +9)	-148% (-282, -36)

Figure A.1. Bootstrap distribution of trend projection for Hoalp, period 2065 for (2051-2080)

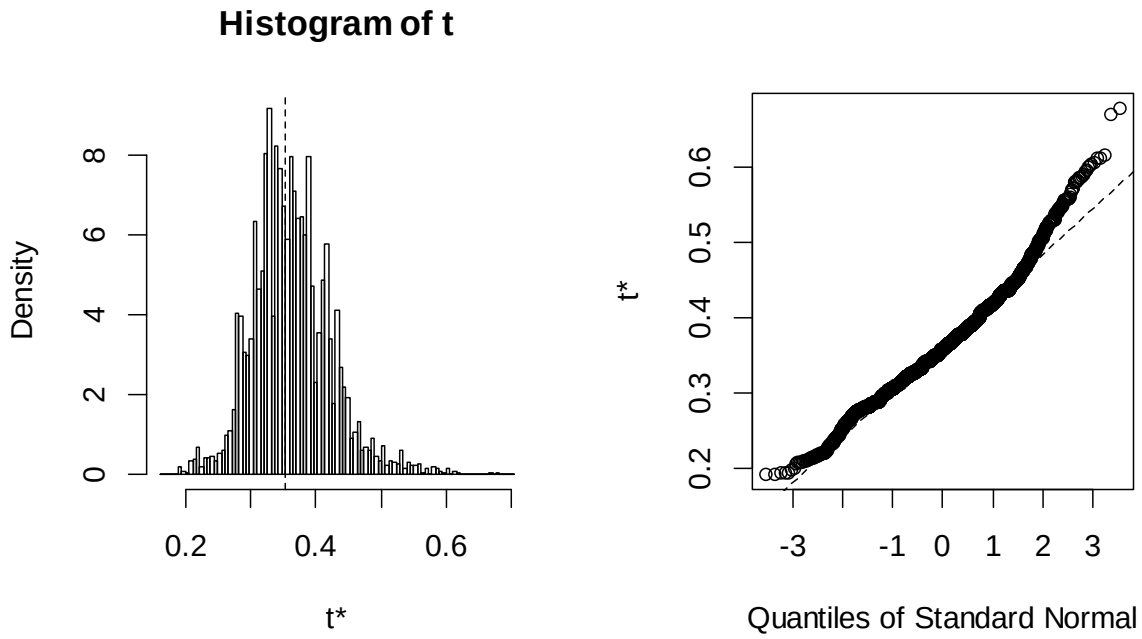

Figure A.2. Bootstrap distribution of trend projection for Muhlv, period 2065 for (2051-2080)

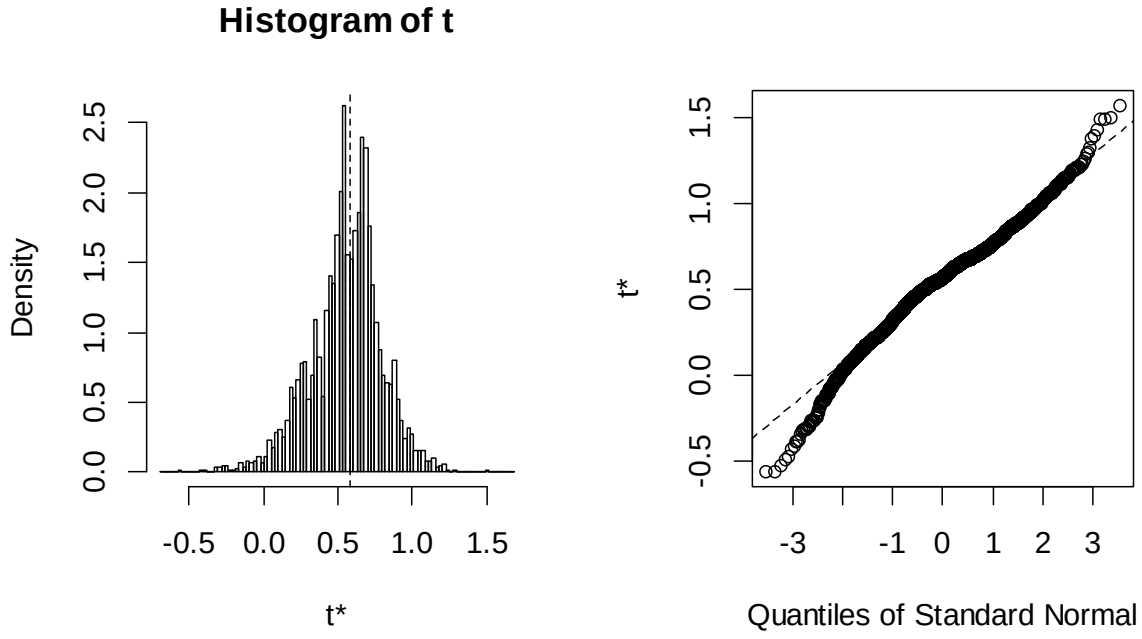

Figure A.1. Bootstrap distribution of trend projection for Gurk, period 2065 for (2051-2080)

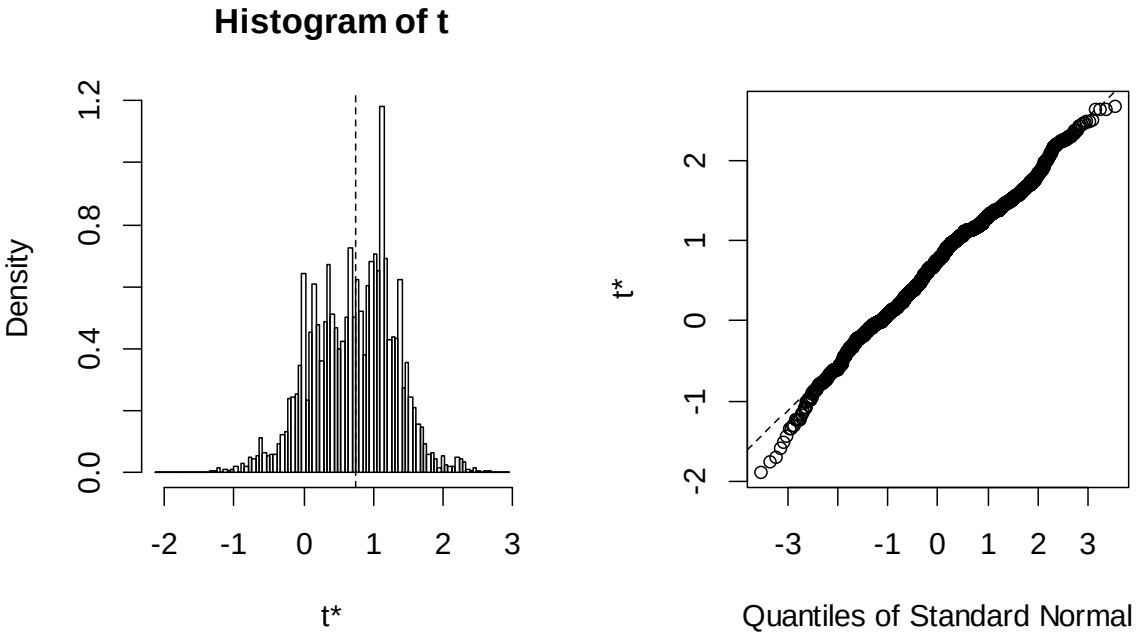

Figure A.2. Bootstrap distribution of trend projection for Buwe, period 2065 for (2051-2080)