Reply to Referee #2

We thank the Referee for the valuable and insightful comments. Below we reply point to point and describe the modifications that we plan to introduce in the revised version of the manuscript. Our replies are evidenced in blue and italic.

This paper calibrates a 12-parameter conceptual hydrologic model (HYPERStreamHS) for the 9850-km2 upper Adage River Basin (Italy), using observed data and bias-corrected data of three regional climate models (EURO-CORDEX), for the 1982-2010 reference period. The model is parameterized for the climate model data using (i) the Kolmogorov Smirnov (KS) statistic for the empirical distribution functions of annual extremes and (ii) flow duration curves. The KS test is subsequently used to test if the observed and simulated extremes are drawn from the same probability distribution. The paper also plots the parameter ranges of the 200 best solutions and models future streamflow extremes.

Reply

We thank the Referee for this careful summary of our work.

General comments

(1) The first key weakness of the research is that the authors calibrate 12 parameters of a conceptual hydrologic model using just 29 annual daily stream flow extremes. This is, of course, a terrible over-parameterization. The effect of over-parameterization on the streamflow simulations needs to be quantified.

Reply

We understand the concern by the Referee of a possible overparameterization. In the following we show that this is not the case in our work. In the submitted manuscript we considered the issue of overparameterization in two different ways:

1) *we investigated identifiability of model parameters for the case in which KS metric has been used (Figure 5) on the basis of the best 200 retained simulations. Results indicate that, though not all the parameters are equally important, some of them present a very good identifiability, thus suggesting that overparameterization could be less than what could be suggested by the number of parameters;*
2) *Evaluation of parameterizations obtained by using KS as efficiency metric in terms of R_{FDC} led to satisfactorily results (in the range between 0.45 and 0.80) for all the investigated cases (see Table 2). Since R_{FDC} is a metric that uses the entire time series of observational data, we consider this as an additional evidence that the use of KS metric on annual maxima is not terribly biased towards over-parameterization.*

However, we do agree with the Referee that this aspect deserves a supplement of analysis and discussion. In the revised manuscript we plan: i) to expand the discussion of the aforementioned points; and ii) to introduce new analyses supporting our conclusion that the use of KS metric does not lead to terribly overparameterized models, as it will be outlined afterwards.

To demonstrate the low level of overparameterization, similar to most applications, of our proposed approach we performed the following additional simulations. As a first *step we evaluated the performance of the model in the time window 1950-1980, not used for calibration, at the gauging station of Ponte S. Lorenzo at Trento. The verification is done by using precipitation and air temperature provided by the ADIGE dataset. The parameters used were those obtained by calibrating with KS with the annual maxima of the time interval 1982-2010. Similarly, to the cases presented in the original manuscript the first two years of simulations have been used as spin-up period. Results are presented in Figure R1a which shows the simulated and observed ECDFs of annual streamflow maxima and the associated p-value of the Kolmogorov-Smirnov test. Reproduction of observed ECDF is satisfactorily, especially for high flow quantiles, with a KS value of 0.233 (it was 0.067 in calibration) and p-value = 0.372. In a strict statistical sense KS metric provides simulated samples of annual streamflow maxima belonging to the same population of observations also in the time window 1952-1980; the reduction of p-value from calibration to validation is not negligible but altogether rather common in hydrological models.*

We also performed a spatial validation of the KS-ADIGE parameterization by simulating streamflow at the Bronzolo gauging station (about 6000 km^2) in the time window 1982-2010. These data were not used in the parameters calibration phase. Results presented in Figure R1b highlight an excellent reproduction of the observed ECDF of annual streamflow maxima with a KS = 0.133 and p-value = 0.951. This is in our opinion a noteworthy result which indicates that the KS-ADIGE parameterization is reliable and does not introduce distortion in the spatial distribution of hydrological processes, particularly for those associated to high streamflow events, i.e., runoff generation and streamflow concentration processes.

The reason of this good agreement obtained in both validations is in the fact that, although the model is calibrated to the annual maxima daily streamflows, it is applied to simulate the entire streamflow time series and therefore the maxima are reproduced correctly only if the interaction between the precipitation and streamflow relevant during high flow extremes is correctly reproduced. We propose to comment this in the revised manuscript in addition to the loss of identifiability of parameters not relevant in the reproduction of high flow extremes such as parameters controlling snowmelting and groundwater contribution, the latter being relevant only for low flows (Figure 5 of the submitted manuscript). This shows that we cannot exclude that additional analyses could be envisioned for limiting overparameterization (e.g., reduced number of parameters, presence of constraints in the parameters range, etc.) in applications dealing with different hydrological models and different data availabilities (e.g. lower number of streamflow extremes). However, these evidences are in our view enough to consider the model reliable.

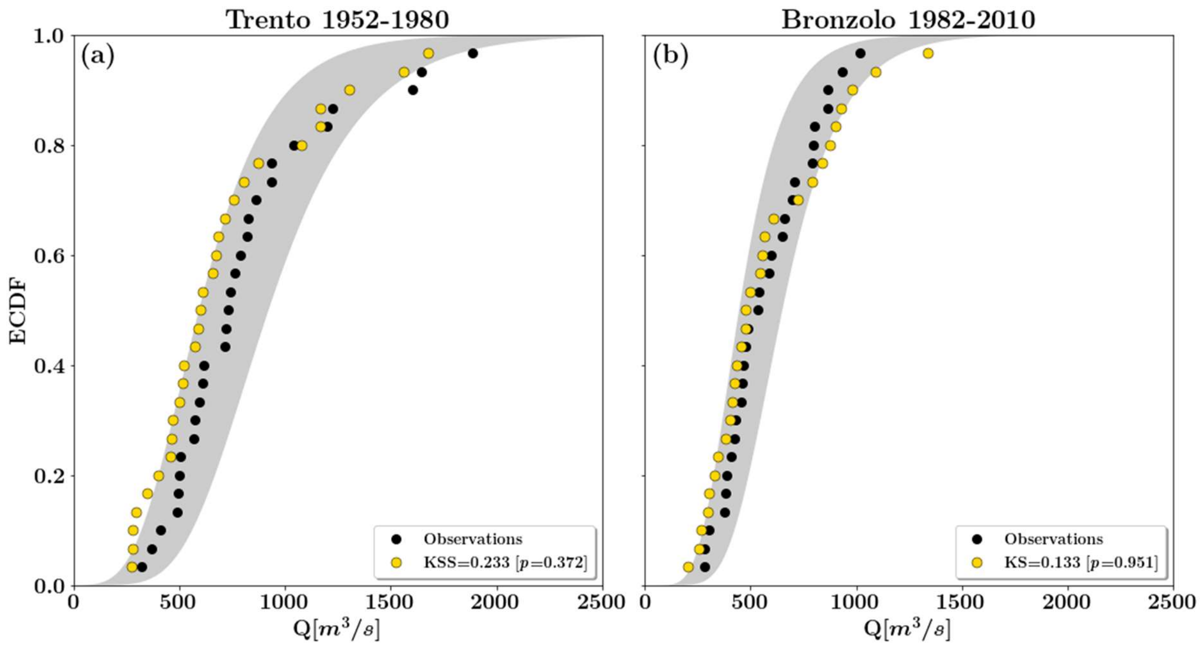

Figure R1. a) ECDFs of annual maximum daily streamflow at Trento gauging station in the period 1952-1980 and b) ECDFs of annual maximum daily streamflow at Bronzolo gauging station in the period 1982-2010, obtained by using as input the observational dataset ADIGE.

As a concluding comment, we remark that our main objective is to present a methodological framework for improving high flow extreme determination in climate change scenarios, not to propose a new multipurpose calibration procedure. We investigated the overparameterization issue as requested by the Referee and we believe we provided arguments that our methodological approach does not lead to a "terribly overparameterized" model: a physiological deterioration of modeling performances in validation is indeed common to all hydrological models and applications.

(2) The second key weakness is that there is no evaluation (validation) of the parameterized models with an independent data series. How can we call this reliable and accurate? (Highlights, l.18-19)

Reply

Successful validation of the KS-ADIGE parameterization in the time window 1952-1980 was presented in the previous Reply. Additionally, we performed the validation of parametrizations obtained by the NSE-ADIGE and R_{FDC}-ADIGE metrics in the time window 1952-1980. These simulations led to NSE and R_{FDC} values (NSE = 0.803 and R_{FDC} = 0.804) which are only slightly lower than those obtained in calibration (NSE = 0.822 and R_{FDC} = 0.874), thus indicating once more the reliability of the adopted modeling framework. In the revised manuscript we plan to present and comment these additional results.

(3) Why not present the characteristics (figure, table) of the rainfall extremes of the observations and the climate models?

Reply

Evaluating extremes of precipitations is not the objective of this work and we are afraid that including them may cause a loss of focus, considering also that the comment n. (6) invites us to revise deeply the manuscript in order to avoid unnecessary details or discussions of relative importance with respect the main objective. The Euro-Cordex simulations have been widely studied in the literature and we feel that referencing these studies will suffice to describe the context and allow the interested reader to deepen this aspect not directly considered in our work. We believe that in this way we obtain a good compromise between the request of curtailing the paper and the need to introduce the additional results presented in the Replies 1 and 2 of this rebuttal.

(4) It is not surprising that the KS test for comparing the empirical distributions of observed and modeled annual flow extremes will give a better result for the model optimized for these extremes with the KS statistic than for the models optimized with the flow duration curves or with the NSE. However, we can also understand that the KS test has its limitations (see Figure 3), so please present in this light.

Reply

Our objective is to show that calibration should be tailored to the application and in particular if the objective are the maxima, minimizing KS (i.e. imposing that experimental and numerical ECDFs are as close as possible) is a better strategy that imposing correspondence between the FDCs or between chronological time series. In Figure 3 we showed that low statistical coherence (i.e., small p-values of the Kolmogorov-Smirnov test) can be achieved when employing parameterizations obtained with a calibration approach directly targeting the statistics of extremes, but still using observational data as input, i.e., KS-ADIGE. We agree with the Referee that each metric has its own limitations and trade-offs and in the revised manuscript we will try to further stress this point.

(5) The paper is written in a wild wild way. We find Methods in the Introduction, Methods in the Results and Discussion, Introduction in the Results and Discussion, no specific research objectives in the Introduction, inexact language, superfluous text and many repetitions.

Reply

We consider with extreme attention this comment and we plan to restructure significantly the manuscript by introducing all the suggested modifications.

(6) In summary, the paper needs to be completely restructured and rewritten in a concise and quantitative manner. Uncertainties stemming from the two key weaknesses (1 and 2 above) need to be quantitatively addressed, metrics and p-values of section 4.1 and 4.2 should be summarized together in one clear table. Expressions such as statistical coherence, forward simulations, extrapolations, 100% confidence bands (!?) need to be defined in the Methods and possibly reworded.

Reply

We already clarified the modifications that we plan to introduce to address major comments (1), (2) and (5). Concerning the remaining issues, we plan to summarize in a single Table all the p-values, as well as to revise significantly the description of the methodological section in the view of the Referee's suggestions.

Specific comments (non exhaustive)
l.8: error prone RC: Please quantify. The majority of your models are accepted, according to your KS p-value.

Reply

Here we were referring exclusively to the parameterizations derived from calibrations conducted using observed meteorological data. Though it is true that many of these models pass the Kolmogorov-Smirnov test (assuming as customary a level of significance of 0.05), their p-values are low and this is reflected in the estimation of high flow quantiles as depicted in Figure 4 of the original manuscript. In the revised manuscript we plan to revise this sentence to be more explicit.

l.39 Much less?

Reply

Thank you for noticing this. We will modify the sentence accordingly in the revised manuscript.

l.57: iii) due to the impossibility of obtaining totally unbiased climate simulations there is no a-priori guarantee that simulations fed by climate models produce samples (e.g. time series of simulated annual streamflow maximum) that are statistically coherent with observations. RC: Your approach cannot address this problem either.

Reply

What we meant here is that climate model simulations are also affected by bias and they may differ from observed meteorological data. In this respect, we showed that calibrating the model by using the chronological time series of observations leads to biased extremes. Instead, calibrating by using climate models and observed streamflow lead to a better reproduction of observed extremes. In some way this approach can be considered as an additional bias correction removing the biases that the standard procedure was not successful to remove. We do agree with the Referee that interpretation of point iii) may be misleading and we plan to remove it in the revised manuscript.

l.61: by directly targeting. RC: non-scientific language

Reply

Agreed. We will modify the writing.

l.64-73: These are Methods

Reply

We will move most of this information in the Methods section maintaining here only the parts needed to highlight the novelty aspects of our work.

l.66: are constrained to maximize the chances?

Reply

Agreed. We will modify the writing.

l.67: Statistical coherence RC: Please define statistical coherence or use another expression.

Reply

In our view statistical coherence implies that two samples are characterized by a large probability of belonging to the same population. In particular, we used the p-value associated to the well-known Kolmogorov-Smirnov test as a metric of coherence between observed and simulated streamflow maxima. On the light of Referee's comment, we will modify the Methods section to make this definition more explicit.

l.75: Do we really need six references for "goal-oriented"?

Reply

Agreed. We will reduce the number of cited papers.

l.102: Section 2.2 RC: It would make more sense to present this after Section 2.4

Reply

We will revise significantly the Methods section to incorporate this and other comments raised by both Referees.

l.111: A similar definition has been introduced for observed streamflow. RC What writing style is this?!

Reply

Agreed. We will modify the writing.

l.113-119: State your null hypothesis and condense this text.

Reply

The null hypothesis is presented at the beginning of the subsection; however, we do agree that all this part of Methods section should be condensed and rewritten.

l.120-125: Does this need a numbered Section?

Reply

Agreed. We will remove this subsection.

l.121: daily average RC: average daily

Reply

Thanks for noticing this, we will modify accordingly.

l.135-158: The efficiency criteria are without the max and min.

Reply

Though we agree with the Referee that efficiency metrics are typically shown without this notation, we believe that in our work this additional remark is needed to avoid confusion between the definition of the efficiency metric KS (eq. 5) and the two-sample Kolmogorov-Smirnov statistic (eq. 1). The additional advantage is that we highlight how, contrary to NSE and RFDC, KS metric should be minimized.

l.138: sensitive?

Reply

Here we meant that the metric considers the chronological time series of simulated and observed daily streamflow. We will modify the writing.

l.146: repetition

Reply

Thanks for noticing this, we will modify accordingly.

l.162-166: RC: Please condense.

Reply

Agreed. We will condense here.

l.171: adaptation?

Reply

Here we meant statistical inference using a Gumbel distribution. We will modify the writing.

l.172: for comparison purposes in order to extrapolate RC: Now what is it?

Reply

The term extrapolation refers to the common practice of estimating high flow quantiles for a return period beyond the available number of simulation years, procedure that cannot be done on the basis of the ECDFs and that necessarily is to be performed by means of a statistical inference of a theoretical distribution. This is needed in impact studies concerning the effect of climate change on high flow extremes. We will try to make this aspect clearer in the revised manuscript.

l.181: portion?

Reply

Here we meant upper part of the Adige river basin. We will modify accordingly.

l.288: parametric errors, RC: Without comma and what do you mean? All models are simplifications of reality.

Reply

What we meant here is that since model predictive errors are always present we cannot expect that a model reproducing streamflow time series properly is also expected to provide a good reproduction of high flow quantiles. We will modify the text to make this aspect clearer.

l.233: provide an assessment?

Reply

Agreed. We will modify the writing.

l.244: the correction used in the reference period 1989-2010 is extended to the period 1980-2010 RC: This is not clear. Is this done by you and if so how?

Reply

Thanks for noticing this typo. Climate models' biases have been corrected (by comparison with observations) as part of the Euro-Cordex experiment. The correction obtained during the overlapping period between climate models' outputs and observational data (i.e. 1989-2010) is extended to the previous 9 years in order to produce bias corrected scenarios for the entire control period 1980-2010. Since the adopted Euro-Cordex scenarios are the standard used in climate change impact studies in the revised manuscript we plan to significantly reduce this part of the text.

l.253-257: RC: Methods

Reply

Agreed. We will move this part to the Methods section.

l.260: On the other hand. RC: Which other?

Reply

Agreed. We will modify the writing.

l.278: cast?

Reply

Agreed. We will remove this term.

l.279-284: RC: Introduction

Reply

Agreed. We will move this part to the Introduction section.

l.311: coined here as Hydrological Calibration on Extremes (HyCoX) RC: Repetition

Reply

Thanks for noticing this. We will remove the repletion.

l.326-335: RC: Methods and Introduction

Reply

Agreed. We will move this part into the suggested sections.

l.370: Fig 3

RC: It will be easier to follow if all metrics and p-values are presented together in Table 2.

Reply

Thanks for the suggestion. In the revised manuscript we plan to summarize in a single Table all the p-values,

Reply

l.373: Forward?

Here we meant evaluation runs with parameterizations derived from a given calibration experiment. In the revised manuscript we plan to modifying this terminology.

l.377: the 90% confidence interval of the observed ECDF. RC: of the fitted extreme value distribution function?

Reply

Exactly, we evaluated confidence intervals by means of a parametric bootstrap. We will clarify this aspect throughout all the revised manuscript.

l.407: Extrapolation?

Reply

We already replied in a previous comment. We will modify the writing in the revised manuscript.

l.418: 100% confidence bands?!

Reply

Here we consider as a metric of uncertainty for the calibrated parameter the range between the maximum and minimum value of each parameter in the 200 retained simulations. We will modify the writing in the revised manuscript to clarify this aspect.

l.429: Furthermore, we verified a-posteriori that the optimal parameters are inside the range of variation. RC: The methods are unclear. Is this “range of variation” (please use a better expression) for the 40,000 simulations? How can the optimal parameter fall outside the range?

Reply

Agreed. We will use the term “parameter range” in place of “range of variation” and we will clarify that this is valid for all the 40,000 simulations. In general, we plan to revise significatively the Methods section to implement all the useful suggestions provided by the Referee. Notice here that 40,000 simulations are simply the maximum number of particle positions, given by the number of particles times the maximum number steps used in the particle swarming algorithm.

We are not sure to understand what the Referee meant with the last part of her/his comment since in Figure 5 the optimal value as well as all the parameter values of the retained simulations fall within the normalized parameter range.

l.440: The differences observed in the optimal value of model parameters are due to structural errors in the GCMs and RCMs. RC: Really? And now we use these errors to make an erroneous hydrologic model, without any independent model validation.

One can understand that there are two modelling approaches each with assumptions and uncertainties. So please stick to quantitative evidence.

Reply

We agree with the Referee that this is an unsupported statement. We will remove it in the revised manuscript. Considering the aspect of model validation, we believe we provided in the first two replies evidences that our proposed methodological framework is accurate and reliable.

l.446: Furthermore, our approach provides an answer to the need of reducing uncertainty in climate change impact assessments. RC: Please quantify your uncertainty reduction.

Reply

What we have shown is that using in calibration the same climate model later used in the projection ensures statistical coherence, which is lost when calibration is performed by using observed meteorological input (given that projections should be done by using the climate models). In this respect, our goal-oriented calibration framework aims at improving the estimation of extremes by directly calibrating the selected hydrological model to the quantities of interest. In the revised manuscript we will modify the writing to better clarify this message and avoid any misunderstanding.

l.452: Marked dashes?

Reply

We will modify the term with "bold horizontal dashes".

l.490-594: RC: Please be concise. Answer your research objectives, which you should have stated in the Introduction.

Reply

Agreed. We will reduce significantly the Conclusion section.
