Dear reviewers and dear Editor,

We uploaded a new version of the paper. We inserted modifications answering to most of the reviewers requests. Various modifications were done along all the manuscript but we report the main structure changes:

- We made modifications on figures and added most of requested information
- Section devoted to hydrological model was re-organized in subsections and more details about the implementation were given
- The performance of the hydrological model used with mean parameters on calibrated basins were evaluated in order to evaluate the possible impact on ungauged basins.
- Analysis on annual maxima of accumulated rainfall on 24 hours was added.
- We removed references to later sections
- The language was revised by an expert of English language along all sections; if editor and reviewers agree that the paper is now improved regarding the scientific point of view but it still needs improvement regarding the language, we are open to submit the manuscript to a language revision service.

We hope the paper is now suitable for publication.
Answers to the reviewer comments are reproted in the following and we briefly described how we modified the manuscript.
The authors.

Reviewer 1

The reviewed manuscript corresponds to the authors response to the mayor revisions of the article: "Analysis of the streamflow extremes and long term water balance in Liguria Region of Italy using a cloud permitting grid spacing reanalysis dataset" by Silvestro et al.

General comments

- The authors present an approach (modelling chain) to use a long-term and high spatial resolution (4km) regional climate model with further bias correction to force a hydrological model over a large region in Italy. Calibration of the hydrological model was performed when possible and then parameters were transferred to ungauged basins. On gauged basins the model seems to represent relatively well streamflow; however, I see a significant limitation in the approach used to transfer these parameters to ungauged basins, as these are purely empirical, and the average of the calibrated parameters were used in ungauged. A GEV distribution was fitted to the simulated Annual Discharge Maxima (ADM) at all gauged sites and compared with observed streamflow, for which I see a fair agreement as fitted GEV 95% confidence interval sometimes does not cover observations (3 out of 6 in my opinion). I think the authors underestimate the role of the hydrological model in the overall analysis and discussion through the manuscript; more emphasis should be given to it, as this is largely driving the simulated hydrological response of these basins. The Ratio(t) is defined to assess the performance of a regionalization approach in characterizing streamflow for several return periods (T). Honestly I have somewhat a hard time understanding the usefulness of this regionalization approach, as it does not seems to work very well. I encourage the authors to demonstrate that the Ratio(T) is being well represented, as I see problems in small and large scale basins (Figure 11). The analysis of the effect of downscaling precipitation on the streamflow should be revised (see specific comments). The final analysis of the mass balance components (runoff ratio) seems very useful and maybe it should be more deeply explored and the main focus of the manuscript.
- I acknowledge that the authors responded to most of the comments from the previous revision; however, there are still significant issues with the manuscript in my opinion.
- For latter revisions I would encourage adding more details to the author's response letter, specifically where the changes are located in the manuscript (page and line), that way it would be much easier for us to go through the revisions.
- I believe that the English of the manuscript needs to be further improved.

Specific Comments:

1 Introduction:

- Page 4, Line 22-23: I agree in the use of high-resolution RCM to reproduce small-scale rainfall events, but I think you need a reference to this or even better if you can show that this is the case with the original dataset in your study region.

Done, added four references (Buzzi et al., 2013; Marta-Almeida, 2016; Pontoppidan et al. 2017; Schwitalla et al. 2017).

Pgg 4, lines 19-20

2 Material and Methods:

Study area and case study:

- Figure 1: To help the description of the figure, use (a), (b) and (c) to refer to each of the plot. And explain each of them as an inset of the others. I'm not sure about the utility of including the Curve Number, if you want to keep this you need to reference the source of this information. Also for the legend use "Curve Number" instead of C.N., try to avoid acronyms, same for DEM, etc.

We changed the Figure as requested, CN map was removed

- Page 5, line 12: refer to table 1.

Ok done

- Page 6, line 5: What about the raingage distribution vs elevation, is orographic precipitation well represented by these stations? Discuss.

We added the fact that they are quite well distributed with elevation. In any case we used all available data.

Pgg 5, lines 18-22

- Page 6, line 5-7: be more specific, how low?

We added more details in the text: "…., respectively about 1/50, 1/200, 1/200, 1/60 km^2"

Pgg 5, lines 22-25

- Page 6, line 7-8: what data was used for calibration/validation?, I assume streamflow discharge, but it is unclear.

-This is described some lines after those indicated by the reviewer (Page 6 lines 15-17 old text) and in the Hydrological model section. In any case we now used subsections in section 2.4, this should help in better clarify model description and its validation.

Pgg 12, lines 12-21

Downscaling the rainfall

- Pag 10, line 10: Define Lr and tr**.**

Done, we added a sentence to clarify "The model takes into account the variability of precipitation at small spatial and time scales (e.g. L ≤ 1 km, t ≤ 1 hour), preserving the precipitation volume at the scales considered reliable (Lr, and tr) for quantitative precipitation forecasts. In other words Lr, and tr are those scales where we expect, on average, a reliable forecast of precipitation volume"

Pgg 9, lines 1-9

- Pag 10, line 10-12: add reference.

Done (Rebora et al., 2006a)

Pgg 9, lines 7-8

The hydrological model

- This section needs better organization, use subsections if needed to explain: model inputs, calibration/validation and model structure (parameters, spatial resolution, etc.)

The section is now re-organized in two subsections. The first describes generically the model the second focuses on the implementation for the current study-

Pgg 10 from line 10

- How is evapotranspiration being calculated?
We added a sentence and reference…"… The energy balance uses the "force restore equation" (Dickinson, 1988) that allows to explicitly model the soil surface temperature and estimating the evapotranspiration from the latent heat flux (Silvestro et. al 2013)…"
Pgg 11, lines 11-13

- pag 13, line 15-17: Calibration procedure is not very clear. What do you mean by 11 sections, what's a "section" (basin?).
These information are reported in Table 1, moreover we changed the text accordingly:".. The model was calibrated on 11 outlet sections where streamflow observations were available at hourly time resolution (see Table 1);"
Pgg 12, lines 12-13

What ground stations measurements were interpolated? I assume you calibrate and validate the model using the downscalled reanalysis data, please clarify.
No we used observations, we tried to clarify the text "…the hourly measurements of rainfall, air temperature, solar radiation, air relative humidity provided by the regional weather stations network were interpolated on a 1-km regular grid through a Kriging method fed to the model"
Pgg 12, lines 13-15

- At what temporal scale did you calculated NS and REHF, hourly or daily?
We added a sentence to clarify in section 2.4.2: "The observed streamflow data at 60- minute time resolution were compared with the model output in order to evaluate its performance. "
Pgg 12, lines 17-19

- What range did you use for calibration and why?, explain. How did you come up with the 500 mm and 1 (Vxmax and Rf) for all the basins, wild guess?
We added a table (Table 2 of new text) with the range of parameters and added a sentence in section 2.4.2 : "The parameters range values considered during the calibration process were defined considering their physical meaning, the mathematical constraints and the experience, they are reported in Table 2 (*Silvestro et al., 2015; Cenci et al., 2016)"*
Pgg 13, lines 15-17

Since Vxmax and Rf are less sensitive then the other 4 parameters we decided to fix them at regional scale similarly to what done in Davolio et al. (2017), we now highlight this fact in the text (section 2.4.2).
Pgg 12, lines 15-17

- Pag. 14, line 16-17: This point is crucial in the manuscript. I disagree that taking average parameters values from calibration into ungauged basins is a satisfactory approach, especially as this are highly empirical values, and this is critical to analyze the results. I would encourage the authors to demonstrate that the impact of taking average parameters in ungauged basins is minimal. For this purpose you can use the average parameter values in the model of a gauged basin and demonstrate that these parameters have little impact on the streamflow performance. I think a significant assumption is being made here, which is not easy to support given the empirical basis of the model.
We did the test requested by the reviewer. We added in table 3 (Table 2 old text) the statistics calculated on the model run by adopting average parameters also on calibrated sections. Clearly the performance are worse in respect using calibrated parameters, but as it can be clearly seen from table 3, the skill scores maintain good values. The new text highlights this finding.
Pgg 13,14 lines 21-2

- For the regional analysis it is necessary to know how much of your study region is covered by gauged and ungauged basins, this can help discussing the impact of the assumptions in model's parameters.

We added a sentence in section 2.6 "…The method was conceived and tested especially for the Tyrrhenian catchments of Liguria Region, so the present analysis was carried out only for this area; the 45-50% of which is located upstream the calibrated basins…"
Pgg 1, lines 19-22

- Table 5: need to include the periods available for discharge, and ADM.
As explicitly mention in the text (sections 2.6 and 3.5) data for ADM are not continuous and they are made of 20 to 50 years length time series often not continuous, the time windows of data availability are often not overlapped with ExpressHydro simulation. Similarly runoff ratio have been deducted by the official documents (we inserted also the link), which again are not continuous in time.
These are the reasons why we did not inserted the exact periods, but we explain the fact in the text.
Pgg 14,15 , lines 24-4

3 Results
- Figure 2 and 3: I would recommend showing the difference in precipitation as a percentage with respect to the observed precipitation, or as (mm), so it is easier to compare with the other maps which are in mm.
Done

- Pag 17, line 17: not sure what you mean by "are largely confirmed"?. Please change the units of the difference map so the comparison is easier. And do the same when you analyze this difference.
Ok, we changed the sentence: "..results on annual rainfall depth confirm the findings of Pieri et al. (2015) both on eastern and western Liguria sides"

- Figure 4: Include coefficient of determination of correlation and mean bias.
Done

- Figure 5 and 6: Values presented in this figures are good for the analysis; however, I think you already have enough figures regarding the performance of EXPRESS-Hydro (Fig 2 to 6) I would try to merge or remove some, maybe by adding more information to Table 3 you can remove one or two plots without losing much information for the analysis.
Figure 6 were added to satisfy the request of a reviewer during the first review, on the other side the regional perspective (Figure 5) is in our opinion interesting. We believe that they are both to be maintained, the total number of figures is high but similar to other published papers. In any case if editor and reviewer retain necessary remove one of them we will do it.

- As ADM are a key aspect of your analysis, I would also show the performance of the bias-corrected precipitation time-series in representing peak precipitation, you could look at this by comparing the probability distribution functions of simulated and observed precipitation. I think this is very important as your peak flows are driven by the peak precipitation events, not by the precipitation volume. This will help you in the discussion of the ADM model's performance.
Done, see pgg 18 lines 6 to 15 and Figure 7.

- Pag 18, line 21-24: why did you chose these 4 basins and nor others?
We added a comment in section 3.1: "..the basins locations was spread from east to west side of the region to investigate if different behaviours arise along the study area.."
Pgg 17, lines 23-24

- You should try to merge figure 7, 8 and 9.
We believe that it is better to maintain different figures in order to have more flexibility during editing (we have other big multi-panel pictures, as figure 3..); moreover we did some attempts to make a unique figure but we believe it results poorly readable and makes difficult to analyze the results
If editor and reviewer retain this point necessary we can do it, but we do not believe this improves the presentation of our work .

- Why did you choose these 4 basins to contrast the analysis?
We added a comment in section 3.1: "..Six basins were chosen in order to evidence the variability of results, showing either good and poor performances"
Pgg 18, lines 18-20

- Explain what is that you are testing with Kolmorogov-smirnov test, and this should be in your methods too. I know you are talking about ADM distribution but it should be clear from in your manuscript.
In section 2.6 we added a sentence "…Moreover the comparison between observed and modeled ADM was also done using the Kolmogorov-Smirnov test with a 5% significance level, in order to verify if they belonged to the same distribution.."
Pgg 15, lines 12-14
In section 3.2 we changed a sentence in order to explicitly refer to ADM…" The Kolmogorov-Smirnov test with a 5% significance level on ADM was applied to all the selected stations and the corresponding results are summarized in Table 5"
Pgg 19, lines 11-12

Section 3.3
- Figure 10: are you using data from every grid point in your model to construct this curve or selected basin outlet?, this need to be clarified in the text. Avoid shortening the words in the figure; you have enough room in the plot, change 'Calib' to calibration, 'simul' to simulations, etc. This applies to all plots/tables.
We modified the figures 7 to 10 (now 8 to 11).
The caption already mentioned explicitly that we used all grid points. We slightly modified it to better clarify: ".. Bottom panels: results without and with rainfall bias correction using all the grid points with drainage area larger than Ath."

- Can you comment or interpret the step-changes in the observed growth curve of figure 10 in terms of the dominating hydrological processes in the region, and why you don't see that in the simulated values?
The change of slope is quite evident in observations and simulated peaks on calibrated sections, for values of Qpeak/Mean around 2; it was found also by Boni et al. (2008) and it is mainly due to presence of extreme events in the time series. When considering the simulated series on all the grid points the curve is smoothed because of the availability of a larger amount of samples (30 years of ADM for every grid point of the region). The fact of having a lot of time series causes also the increase of slope for Qpeak/Mean very high (>4-5): the most extreme flow events pass on various grid points with similar effects in terms of Qpeak/Mean. Still the "double-component" behavior of the observed growth curve of figure 10 is still present in the simulated curve.

- It seems to me that the relatively better performance in the regional growth curve, both calibrated and total area, could be the results of errors compensation as figure 7 to 9 show that observed distribution of ADM lies outside of the confidence boundary in some cases (Bisagno, Magra, Argentina; 3 of 6). This needs more discussion in the manuscript. If this approach is meant to be used in ungauged basins, the model needs to prove that it works in the calibrated basins. This can be a major problem in the proposed modelling chain. If the authors can prove that the model in calibrated basins can represent ADM, this will help a lot in supporting the proposed modelling chain.
We added some comments to discuss more this issue, but we partially disagree with this point. The regional approach to build the growth curve is indeed used to reduce errors on single basins (see section 3.2, some ADM fitting are good some are bad). This is also highlighted (as already shown on the paper) by the comparison of the regional curve built with all grid points with the one built with the calibrated basins only. Her is the new text…"…. It is important to highlight that regional approach allows to reduce the errors that can be found for single basins (Boni et al., 2007), and which are shown in section 3.3; on one side the normalization of each ADM series with its average reduce the effects of bias (due for example to a bad hydrological model calibration), on the other side the ADM time series of each outlet section (or grid point) is only a small sub-sample of the entire sample used to build the regional curve…:"

Pgg 21, lines 1-5

Text already present:"…. The curves that used only calibrated sections are really similar to the others, proving that the latter configuration enhances the robustness of the regional curve estimation without introducing evident errors…"

Moreover in section 3.2 we already presented some comments about the fact that single bad fitting does not means that the general hydro-climatology of the region is badly described:"…. We would like also to highlight the fact that simulated ADM distributions have often similar shapes to the observed ones and suffer of a sort of bias (for example Bisagno closed at La Presa, Figure 8), while in other cases the simulated ADM distribution is only partially out of the confidence intervals (example Argentina closed at Merelli, Figure 9). The average hydrologic regime on the study region could be only partially affected by the local bad fittings..."

Finally, as requested by the reviewer, in section 2.4.2 the scores for calibrated basins are calculated using average parameters in order to evaluate the expected errors of hydrological model on un-calibrated outlet sections

- Figure 11 shows that small basins are underestimated (as discussed in the manuscript), and it also shows an overestimation for the Ratio(T) for large basins, how can the authors explain this?. Also I'm not convinced that the B.C. Ratio(T) shows an improvement in Figure(11), as small basins that used to have a good performance (near 1) now they are overestimating (Ratio>1). Overall I'm doubtful about the usefulness of this Ratio(T) as it does not seem to produce good results. I think the authors need to show that this index, which tries to show that the regional curve is suitable for all basins, works.

Ratio(T) is a simple way to compare the results with the benchmark (1 perfect; <1 underestimation; >1 overestimation). Many comments can be found to discuss results of Figures (11, 12 and also 13, now 12,13 and 14): page 21 and part of page 22 were devoted to discuss results and explain the reasons of these results.
In the new text we also evidenced that B.C. seems to introduce overestimation on large basins "….B.C. introduce an overestimation on larger basins (A > 200-300 km2)", while in text is already discussed the overestimation of the central part of the region also for small basins. The further analysis described in pages 21-22 and figure 14 evidences that mean Ratio improve using BIAS correction for all T, with values that remain in most of the cases (Fig 12 and 13) in range [0.5 1.5] and maximum range [0.3 2]; this is not a foregone results.
It is finally important to highlight (this comment is present also in the text) that we are comparing results with a "benchmark" and not with observations. As a consequence, also the benchmark can be affected by uncertainties, thus it appear to us an interesting achievement the fact that Quantiles have the same order of magnitude and in many cases differences lower than 50%.
We would like to highlight again that we are not presenting a definitive operational methodology, but we are exploring the potentialities of the presented modeling chain to reproduce ADM in an environment made by small basins, in complex topography areas, especially testing the usage of a high resolution (cloud permitting) reanalysis. To our knowledge this is one of the first works of this kind done in the considered study area and we have not the claim to solve all the possible issues and problems in a unique work, or to define a definitive methodology that can be applied in any case.

- Figure 12 has little discussion in the manuscript and the comparison that the authors do (line 9-11, page 22) is hard to see from the figure.
We added a sentence to help in reading the legend .." Areas where modelling chain is really close to the benchmark are in green/light blue, whereas dark blue and purple point out where under or overestimation is high (absolute difference larger than 70-100%)...".

We would like to notice that discussion of figure 12 (now 13) continues after lines 9-11 of page 22 of old text, results of figure were discussed even in the new text (page 22). In page 23 comments continue using both figures 12 and 13, and then introducing figure 14.

- Page 21, line 16-18: Then why not excluding these from the analysis?
We think the reviewer refers to page 23 not 21 i.e..".... leads us to consider that the underestimation of quantiles for very small catchments (i.e. A < 30-50 km^2) is a structural problem of themodelling chain…". The answer is that this is a finding (or result...) of the analysis, there were no a priori reasons to exclude them from the analysis.

Section 3.4
This analysis should be changed. In order to fairly compare the effect of downscaling in the mean annual streamflow, which is what I assumed you are comparing (clarify), you should show results from the hydrological model calibrated using the non-downscaled precipitation (unless you did so, but it is not clear from the text) versus the calibrated hydrological model using downscaled precipitation. If you want to show this analysis you should also include the effect of downscaling on Nash-Sutcliffe, REHF and bias against observations.

As described in section 2.5 (now 2.4.2 since we have inserted two sub-sections) the model was calibrated-validated using recent hourly data (2013-2014 period used for validation). The calibrated model is then used for simulating the 1979-2008 period using as input ExpressHydro Reanalysis with and without rainfall downscaling. So we have not different calibrations, but a unique calibration done for a period where time series of meteo gauges and discharge data are available.

It is not possible calculating the statistics like Nash-Sutcliffe, REHF…etc.., for 1979-2008 period, for 2 reasons: 1) no observed discharge time series are available 2)The ExpressHydro Reanalysis can not reproduce the exact sequence and timing of the real events in period 1979-2008, but only a reasonable climatology of the area. Theoretically it could be also reproduce particular events but with very high errors in geolocation and timing and large uncertainty from a quantitative point of view. As a consequence even in the case we had 1979-2008 observed discharge time series they could not be directly compared with simulated ones (both with and without downscaling)

The presented analysis is done to highlight the effects of downscaling especially on small basins. It appear to us that the presented graph which presents Ratio versus Area helps to synthetically show the downscaling effect, and the reason why we used it. In fact when ratio is far from 1 (lower than 1) it means that downscaling has an important role especially for small catchment (as commented in the text). Furthermore the fact that ratio is always < 1 means that the rainfall downscaling correctly plays the role of enhancing the runoff formation (we explicitly inserted this point in the new text)

We also added in the text the description of terms of the equation 9 to clarify the meaning of RatioDS

Reviewer 2

The article presents a potentially useful study, but there are many aspects that still need to be improved before considering publication. Though the revised version has been partially improved, there are still portions of the methodology and text that are not clear. The organization is messy, forcing the reader to

move back and forth across sections.
Some of this issues and concerns were included in the previous reviews and unfortunately the authors were not able to address them in the revised manuscript.
For example, reviewer 1 of the previous round suggested moving the EXPRESS-Hydro reanalysis section to the introduction, as the text in lines 1 to 19 in page 7 are not part of the methodology. This was not done, and this section is still unclearly linked to the methodology (and this text belongs to the introduction)

Ok. We already moved large part of this section on introduction, we now moved some other sentences and eliminated the subsection 2.2 (Express Hydro dataset)

. He also suggested, replacing the Pieri et al., data by the Express-Hydro reanalysis everywhere in the text. This was not done everywhere, see for example line 9 in page 7, and later line 15 of page 17.... This is just one example of a problem, repeated in other aspects, that makes the manuscript (and work) very difficult to follow.

Ok. Reference on page 7 has been removed. On section 3.1 we replaced Pieri et al.,with data by the Express-Hydro reanalysis when referring to the data set; we maintained the reference to Pieri et al. only when we refer to their results since we made similar analysis (comparison between ExpressHydro reanalysis and rainfall observations) but using a different set of observations. "…Pieri et al. (2015), using EURO4M-APGD reference observational dataset (Isotta et al. 2013, with about 50 daily raingauge stations over Liguria), already showed an overall underestimation of the WRF rainfall depths on annual basis in Liguria,..."...." The same analysis was repeated in this study, using 95 raingauge stations over Liguria.."

There are constant references to later sections which again, make the manuscript difficult to follow.

We removed references to later sections and added some references to scientific works along the manuscripts.

Page 13 (and 14): What do you mean by "it was possible to calibrate the model for 11 sections?" how are "sections" defined? Which are these sections? Please clarify. Please comment on the range of variability of the parameter values, particularly as you the use a mean value for the ungagged basins ("sections?"). It might be worth doing a sensitivity analysis (using the range of calibrated parameters) to investigate the possible impact of using mean values.

We now explicitly refer in the text to basin, we also changed the sentence to clarify which sections "…The model was calibrated on 11 basins where streamflow observations are available at hourly time resolution (see Table 1)…"

Under the suggestion of the other reviewer we did the run on calibrated outlet sections using average parameters (scores are reported in table 2), this should help to evaluate the impact on ungauged basins due to the use of average parameters.

We did not carry out a sensitive analysis because it is out of the scope of the paper, but a detailed sensitive analysis of the model parameters is reported in Silvestro et al. (2013).

Page 17, section 3.1 is unclear. Are you referring to the EXPRESS-HYDRO reanalysis data?. Please, be clear in the introduction about this dataset, and refer here (in the same terms) to your new results. Is the bias

correction considered here? (it seems so, from line 14 in page 18, but please mention it at the beginning).

Yes we changed the sentence at the beginning of the section: "The comparison between EXPRESS-Hydro reanalysis and precipitation climatology over Liguria from observational data was undertaken at the annual, seasonal and monthly scales"

The comparison of precipitation (Express Hydro versus observation) was done before bias correction was applied, we changed the sentence (page 18 line 14) which refers to skil scores calculation. "…while in Table 3 we reported the values of two statistics: BIAS and the Root Mean Square Error (RMSE)."

Only the analysis of 24 hours accumulation annual rainfall maxima is done with Bias Corrected reanalysis as requested by reviewer 1. This is explicitly mentioned in the text.

Text of the section has been revised

The authors should consider a thorough English revision, possibly by a colleague with good English skills to improve readability and English usage. Please note also that the manuscript is full of one-sentence paragraphs, which makes the paper hard to read.
In short, the lack of clarity is compromising the delivery of the message of the paper.

All the manuscript has been revised by an expert of English language

references

Silvestro, F., Gabellani, S., Delogu, F., Rudari, R., Boni, G.: Exploiting remote sensing land surface temperature in distributed hydrological modelling: the example of the Continuum model. Hydrol. Earth Syst. Sci., 17, 39-62, 2013. doi:10.5194/hess-17-39-2013.