The authors are grateful to Reviewer #2 for sharing his concerns with the authors of this submission. Your concerns were carefully taken into consideration for the manuscript revision, leading to a more consistent and scientifically sound manuscript. In the following pages we present a point by point reply to your general and specific/technical comments.

***General comments:***

*1. The title of the manuscript does not reflect the general topic of the manuscript. Also within the main body of text, the authors mix up the definition of water stress (which is a function of demand versus supply) with the definition of hydrological drought/low flows. I suggest the authors to stick to the definition of average and low flows/runoff and hydrological drought throughout the manuscript.*

REPLY: Following the reviewer's indication, the title was changed to better correspond with the topic of the manuscript. The new title is "High-end climate change impact on European runoff and low flows. Exploring the effects of forcing biases". Moreover, references on water stress were eliminated from the manuscript.

*2. The introduction section is too long and consists of redundant information. I takes too long for the reader reaches the main goals of the manuscript. Please remove redundant text and restructure this section please, specifically related to: a. Flooding (not studied in this paper: can be removed, also in the discussion section) b. Comparison of all the GCMs/LSMs: keep it short and focus on results for the EU continent. Do not present everything here, might be more appropriate in the discussion section. c. Do not mix up the concepts of drought and water stress: focus on drought also when referring to literature in this section. d. The model JULES is now explained both in the introduction and in the methods section. I would suggest to replace the majority of this piece of text to the methods section. Only briefly mention JULES in one/two sentences in the introduction. e. Use consistent namings: e.g. when referring to the climate change scenarios +2/+4 degrees global warming. f. Please rephrase the research goals. The goals in itself are fine but they could be defined more precisely.*

REPLY: The reviewer's suggestions were carefully taken into account into the revised version of the Introduction. The Introduction now is shorter and more focused on the topics of this study as redundant information have been removed. Specific changes made per paragraph are:

- Paragraph 4 was deleted as its content (added stresses of population growth and human activities on climate change impact) is not relevant with this study. A reference on the significance of this topic is made in Paragraph 1 of the Introduction.
- Paragraph 5 on multi-model assessments was significantly shortened and moved to the end of Paragraph 3
- Paragraph 9 was removed from the manuscript as it was repeating the information of Paragraph 8.
- As the content of Paragraph 10 (discussion on other LSMs) was not very relevant to the context of this study, only a short definition of the LSMs has been kept of this paragraph. This is now merged into a paragraph where the JULES model is briefly mentioned.
- The discussion about JULES in Paragraph 11 was removed from the Introduction section and was moved to the Data and Methods section (merged with Section 2.2.)

- The research objectives in the last paragraph of the Introduction have been reformulated as follows:
-

"The research objectives set by this study are the following:

i) To identify changes posed on the hydrological cycle (mean state and lower extremes) at +4 oC global warming compared to a baseline situation, and relative to the target of 2 ºC warming.
ii) To analyse the effect of bias correction on projected hydrological simulations. To achieve this, both raw and bias corrected Euro-CORDEX data were used as input forcing in the impact model.
iii) To assess the effect of the observational dataset used for bias correction.
iv) To identify climate change induced changes in drought climatology at the basin scale."

*3. One of the goals is to evaluate the average runoff and low flows under a +2 and +4 degrees global warming scenario. I would suggest therefore to add to table 1 the years in which a +2 degree global warming is reached for each of the GCMs.*

REPLY: The time-slices centering at +2 degrees of warming are added to Table 1. The new version of Table 1 is:

Table 1. Euro-CORDEX climate scenarios used to force JULES.

| | **GCM** | **+2 SWL time-slice** | **Exceeded warming level (ºC) in the +2 SWL time-slice** | **+4 SWL time-slice** | **Exceeded warming level (ºC) in the +4 SWL time-slice** | **Equilibrium Climate Sensitivity (K)** |
|---|---|---|---|---|---|---|
| 1 | GFDL-ESM2M | 2040-2069 | 2 | 2071-2100 | 3.2 | 2.44 |
| 2 | NorESM1 | 2036-2065 | 2 | 2071-2100 | 3.75 | 2.80 |
| 3 | MIROC5 | 2037-2066 | 2 | 2071-2100 | 3.76 | 2.72 |
| 4 | IPSL-CM5A | 2018-2047 | 2 | 2055-2084 | 4 | 4.13 |
| 5 | HadGEM2-ES | 2024-2053 | 2 | 2060-2089 | 4 | 4.59 |

*4. Only 2 of 5 models reach a 4 degree global warming before 2100, can we really speak of a 4 degree global warming scenario then? And is it fair to compare the output of these GCM modelling results with each other or to estimate and ensemble-mean value? Please elaborate.*

REPLY: The main reason for not having models capturing +4 degrees, was data availability. For the models that did not reach +4 SWL, data were not available after 2100 so as to extend the analysis. For the 5 available models, the average exceeding warming level is 3.74 $^{0}$C. We considered that this is close enough to +4 to examine the models as an ensemble at +4 SWL. In the manuscript it is stated that the term "+ 4 SWL time-slice" will be adopted for all models for reasons of consistency. The authors believe that it is fair to compare the output and derive ensemble mean of the 5 models, as it is very typical for climate change assessment studies to extract ensemble means using time-slices (eg 2071-2100), without considering how different the temperature projections are for each ensemble member in that period.

*5. The +2 and + 4 degrees refer to a 'global warming condition' whereas this study looks specifically to the European conditions. Could you elaborate a bit more on the temperature differences (and differences in precipitation accordingly) between the GCMs for the European continent when using the +2/+4 global warming scenario time slices? How could these differences influence your analysis/results?*

REPLY: In the Introduction section there is a comment on the way +2 degrees global warming reflects on Europe: "The effect of a 2 $^{o}$C global warming for the European climate was examined by Vautard et al. (2014). The study revealed that warming in Europe is projected to be higher than the global average of 2 $^{o}$C".
To elaborate more on this, the following piece of text was added:
"Temperature increases of up to 3 $^{o}$C were found for the winter season over north-western Europe and for the summer months over sourthern Europe. Heavy precipitation was found to increase over the whole continent for all seasons, with the exception of southern Europe during summer."
And on the effect of +4 $^{o}$C in Europe: "The +4 $^{o}$C global warming scenario is also translated in more intense temperature increases in Europe, especially for the summer season (World Bank, 2014)."

*6. In this study, only the JULES model is being used for hydrological simulations. I would suggest the authors to elaborate a bit more on the performance of JULES compared to other models, both in the baseline situation and given the future simulations. Moreover, it would be good to show/discuss how well the JULES model matches observational data, preferably with a focus on the pan-European continent.*

REPLY: A piece on the hydrological performance of JULES over Europe has been added to the description of the model section in Section 2 (Data & Methods).
*"Other studies give insight into the hydrological performance of JULES specifically. Blyth et al. (2011) extensively evaluated the JULES model for its ability to capture observed fluxes of water and carbon. Concerning discharge, their findings suggest that for the European region seasonality is captured well by the model. For temperate regions (like most of central Europe) to model exhibited a tendency towards underestimating river flows due to overestimation of evapotranspiration. Prudhomme et al. (2011) assessed JULES' ability in simulating past hydrological events over Europe. In general terms the model was found to capture the timing of major drought events and periods with no large-scale droughts present were also well reproduced. The model showed a positive drought duration bias, more profoundly present in northwest Spain and East Germany-Czech Repuplic. Prudhomme et al. (2011) argue that this feature is related to overestimation of evaporation by the model. For regions where droughts tend to last longer, JULES exhibited a better ability of reproducing the drought events' characteristics. Gudmundsson et al. (2012) compared nine large scale hydrological models, and their ensemble mean, based on their skill in simulating the interannual variability of observed runoff percentiles in Europe. According to the overall performance (accounting for all examined percentiles*

*and evaluation metrics), JULES was ranked third best out of the 10 models, after the multi-model ensemble mean and the GWAVA model. For low and moderately low flows, expressed as 5th and 25th percentile respectively, JULES is also in the top three models regarding the representation of interannual variability in runoff. In the study of Gudmundsson et al. (2012b), where an ensemble of hydrological models is evaluated for their ability to capture seasonal runoff climatology in three different hydroclimatic regime classes in Europe, JULES exhibits a good performance, comparable to that of the best performing multi-model ensemble mean. In other studies employing multi-model ensembles, focusing on the whole European region (Gudmundsson and Seneviratne, 2015) or a single basin in Europe (Harding et al., 2014; Weedon et al., 2015) JULES' simulations also correspond with these of the other models."*

*7. The authors used two hydrological indicators to identify changing climate trends, the average and 10th percentile of runoff production. Reading the manuscript, it did not became clear to me however how the authors applied these indicators. Did they use monthly or yearly values? And if they used monthly values, did they used a variable threshold approach to estimate the 10th percentile values? Or did they use a fixed 10th value over all months? Please clarify.*

REPLY: Average and 10th percentile runoff production were deduced from monthly runoff data. For the analysis of the gridded results with the SWL time-slice approach, each indicator was computed from the monthly values of all years in the time-slice.
For the analysis of basin averaged runoff regime, the two hydrologic indicators were calculated per year, for all the years of the simulation. This resulted in time-series of basin aggregated average and 10th percentile runoff production, spanning from 1971 to 2100. Clarifications for the derivation of the hydrological indicators have been added to the manuscript:
"The two hydrological indicators were deduced from monthly runoff data. For the analysis of the gridded results at pan-European scale with the SWL time-slice approach, each indicator was computed from the monthly values of all years in the time-slice. For the analysis of basin aggregated runoff regime, the two hydrologic indicators were calculated per year, for all the years of the simulation. This resulted in time-series of basin aggregated average and 10th percentile runoff production, spanning from 1971 to 2100."
Note that a different approach was used in defining the drought threshold (described in current Section 2.5, which was 2.4 in the first manuscript). For this, daily values of discharge were used, and a daily varying threshold computed based on the values of the historical period was applied to the 1971-2100 discharge time-series.

*8a. With respect to the examination of drought climatology the authors explain that they 'counted the number of days per year that extreme lows in flow occur'. First, using a 10th percentile value is not yet really an extreme low (I would say extreme lows would be using a 5th percentile or 1st percentile).*

REPLY: The limit set for defining low flows was defined considering its relevance for drought formation. The 10th percentile value has been used in numerous studies as a limit for identifying drought conditions (Hannaford et al., 2011; Prudhomme et al., 2011, 2014; Roudier et al., 2015; Stahl, 2001). As the use of the word "extreme" mostly refers to lower percentiles, we substituted this word in the manuscript with "particular": "changes in drought climatology, i.e. the number of days per year that particular lows in flow occur."

*8b. Secondly, I'm wondering whether the authors used any buffering methods (defining minimum/maximum inter-event times or minimum length of dry conditions in order to be considered a drought) to estimate the drought climatology, see for example Tallaksen et al. (1997) and Sung and Chung (2014). Thirdly, could the authors argue why they used a 'total number of days per year with extreme lows', rather than a 'total/max number of consecutive days with extreme lows', which might be a more appropriate indicator for drought climatology? Finally, Wander et al. (2015) argue that – when evaluating drought conditions under climate change- it is better to use a transient variable threshold approach as (aquatic and terrestrial) ecosystems are able to adapt to changing drought conditions. I would suggest the authors to at least discuss the use of this transient variable threshold here or in the discussion section.*

REPLY: We did not use a method to buffer the discharge time-series before comparing them with the daily varying drought threshold. The authors acknowledge the different methods that the reviewer refers to, although we have made different choices for our analysis. Comments on the different options (not followed in this study) for drought identification are added to the discussion section:
*"Our analysis of drought climatology at the basin scale was based on the total number of days under a predefined daily varying drought threshold. We did not employ any buffering criterion for the days under threshold to be accounted for in the total sum (as discussed for example by (Sung and Chung, (2014) and (Tallaksen et al., (1997)). The use of such a criterion would have decreased the calculated dry days. However, as the interpretation of the results of this study is mostly oriented in identifying trends of change rather than absolute numbers describing the future regime, the lack of a buffering criterion is not supposed to notably affect the extracted conclusions. (Wanders et al., (2015) employed a transient variable threshold for the assessment of the drought conditions under climate change, considering a gradual adaptation of the ecosystem on the altered hydrological regime. This is an interesting alternative, especially for climate change mitigation and adaptation studies. In our study we aimed to identify global warming induced changes in the future hydrological state without considering adaptation, thus the same historically derived threshold was applied to the whole length of the simulated runoff time-series."*

*9. Please mention that the results presented in section 3.1 refer to the use of the Euro-CORDEX data without bias-correction.*

REPLY: "raw" was added before the "Euro-Cordex forcing data" to indicate that these data have not undergone any bias adjustment. This was also added to the caption of the relative figures.

*10. Section 3 is quite wordy about (significant) positive and negative trends to describe the changes in average and 10th percentile runoff values towards future conditions. However, only few statements are actually backed up with numbers/statistics. I would suggest to execute some extra statistical analyses to give your results some extra body. E.g. the trend observations could easily be backed up with a simple regression analysis giving a 'number' to the trend (coefficient) and a feeling of (in)significance of the* *result (R-squared and significance level of the estimated coefficient in the regression). Outcomes could be mentioned briefly in text – e.g. between brackets- and in the figures.*

REPLY: Following the reviewer's suggestion, a regression analysis was performed with the annual time-series of basin aggregated runoff production and total number of drought days per year. The analysis employed a linear regression model for trend estimation and the trend significance was tested at the 0.05 significance level. Relevant information have been added to Section 2.4: "The trend of the annual time-series was investigated employing a linear regression analysis to estimate the sign and the average rate of the trend. The significance of the trend was tested at the 95% confidence interval via a Student-t test."
Trend and p values have been added to the relevant figures and tables of summary statistics of the regression analysis have been added to the ESM (Table S1,S2,S3).

*11. I miss in the discussion section a piece of text elaborating on the use of JULES, the performance of JULES and the potential use of other GHMs/LSMs.*

REPLY: The piece of text added in Section 2, gives information on JULES' hydrological performance, with focus on the European region. It also provides information on the performance of JULES compared to other GHMs/LSMs. At the discussion section, a piece of text about the use of other GCMs and the uncertainty stemming from the use of a single impact model has been added:
*"In our study only one impact model (JULES) was used. Hagemann et al. (2013) argue that impact model induced uncertainty in future hydrological simulations is larger than that of the GCMS for some regions of the land surface and suggest using multi-impact model ensembles to deal with this issue. However useful conclusions can be drawn also from studies employing a single GHM/LSM. Examples of such single model climate change impact assessments performed recently are the studies of Schneider et al. (2013) and Laizé et al. (2013) with the WaterGAP GHM, the studies of Arnell and Gosling (2013), Gosling and Arnell (2013) and Arnell et al. (2013) with the GHM MacPDM and of Hanasaki et al. (2010) using the H08 LSM."*

***Specific/Technical comments:***

*12. 'GFDL and NorESM1 exhibiting generally wetter patterns' (p 7281, line 8-9). Looking at figure 2 I would say that NorESM1 is also generating relatively dry patterns for southern Europe.*

REPLY: The reviewer's observation is correct. The sentence was intended to refer to the wetter patterns GFDL and NorESM1 show in the northern regions but it was confusingly structured. This sentence has now changed accordingly: *"GFDL and NorESM1 exhibiting generally wetter patterns for northern Europe and Scandinavian Peninsula, and with IPSL describing drier patterns, especially for southern Europe."*

*13. 'all models agree' (p 7281, line 11 & 20): I would suggest to add a sub-figure to figure 4 & 5 that shows the modelling agreement with the ensemble mean in terms of estimated change (+/-) in average/10th percentile runoff.*

REPLY: As the reviewer suggested, a sub-figure showing model agreement has been added to the revised version of the figures. The model agreement here is defined as agreement towards a wetter change in the projected time-slice. Thus 100% agreement means that all the five models project more runoff than in the baseline period while 0% agreement means that all five models show a drier response in the projected time-slice. The revised Figures are shown below:

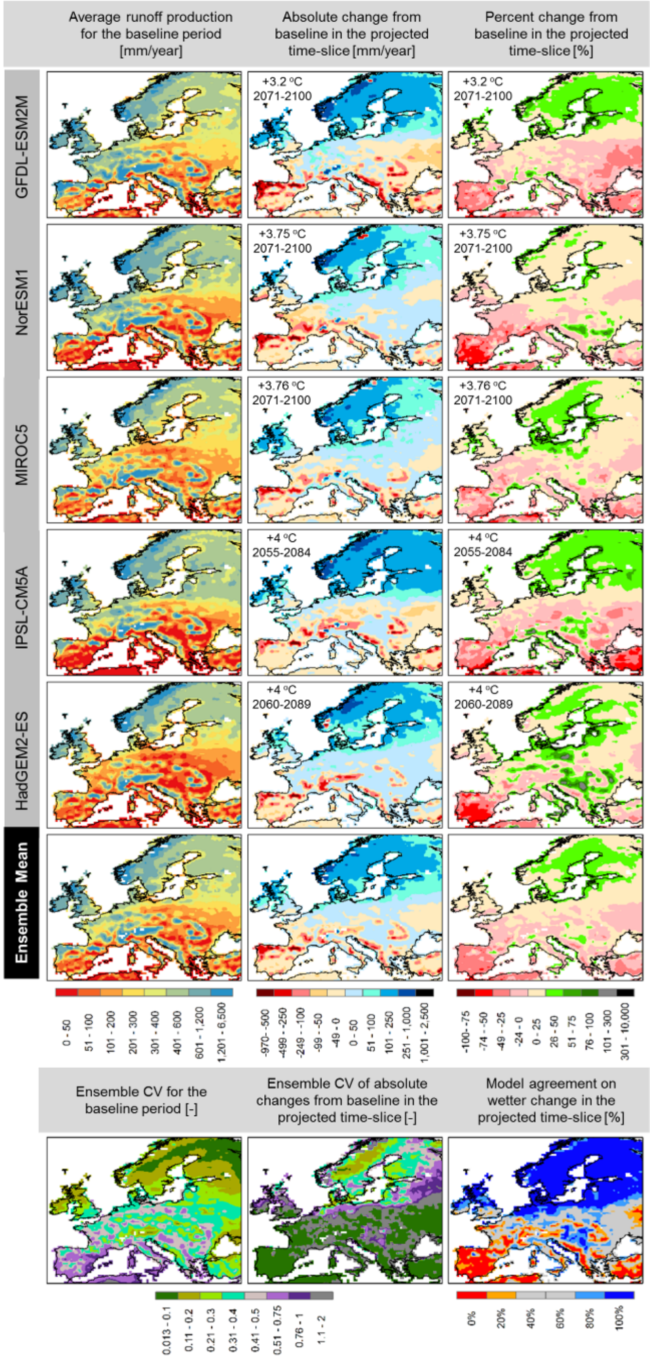


Figure 2. Average runoff production from raw Euro-CORDEX data for all dynamical downscaled GCMs and their ensemble mean. Runoff production averaged over the baseline period (1976-2005) (left column), absolute change in runoff in the +4 SWL projected time-slice (middle column) and percent change in the +4 SWL projected time-slice (right column). Bottom row: coefficient of variation of the ensemble members for the baseline period (left column), coefficient of variation of the projected absolute changes in the +4SWL projected time-slice (middle column) and model agreement towards a wetter change in the +4 SWL projected time-slice.

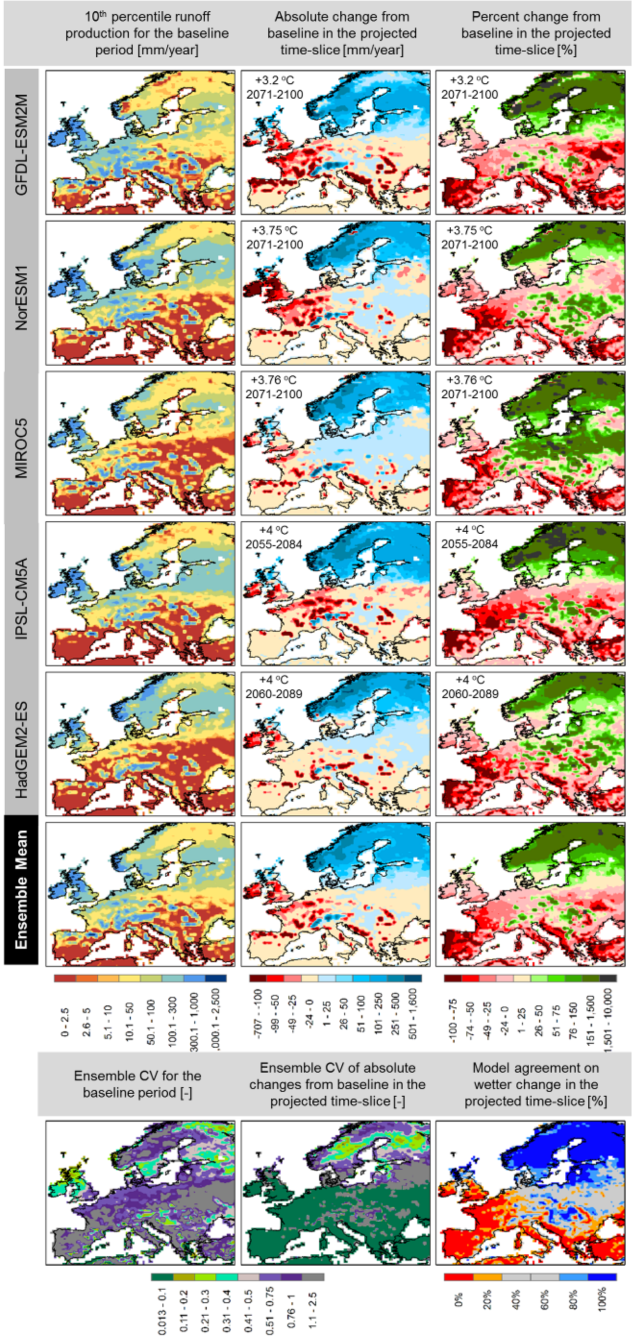


Figure 3. 10th percentile of runoff production from raw Euro-CORDEX data for all dynamical downscaled GCMs and their ensemble mean. 10th percentile runoff production derived on an annual basis and averaged over the baseline period (1976–2005), absolute change in 10th percentile runoff in the +4 SWL projected time-slice (middle column) and percent change in the +4 SWL projected time-slice (right column). Bottom row: coefficient of variation of the ensemble members for the baseline period (left column), coefficient of variation of the projected absolute changes in the +4SWL projected time-slice (middle column) and model agreement towards a wetter change in the +4 SWL projected time-slice.

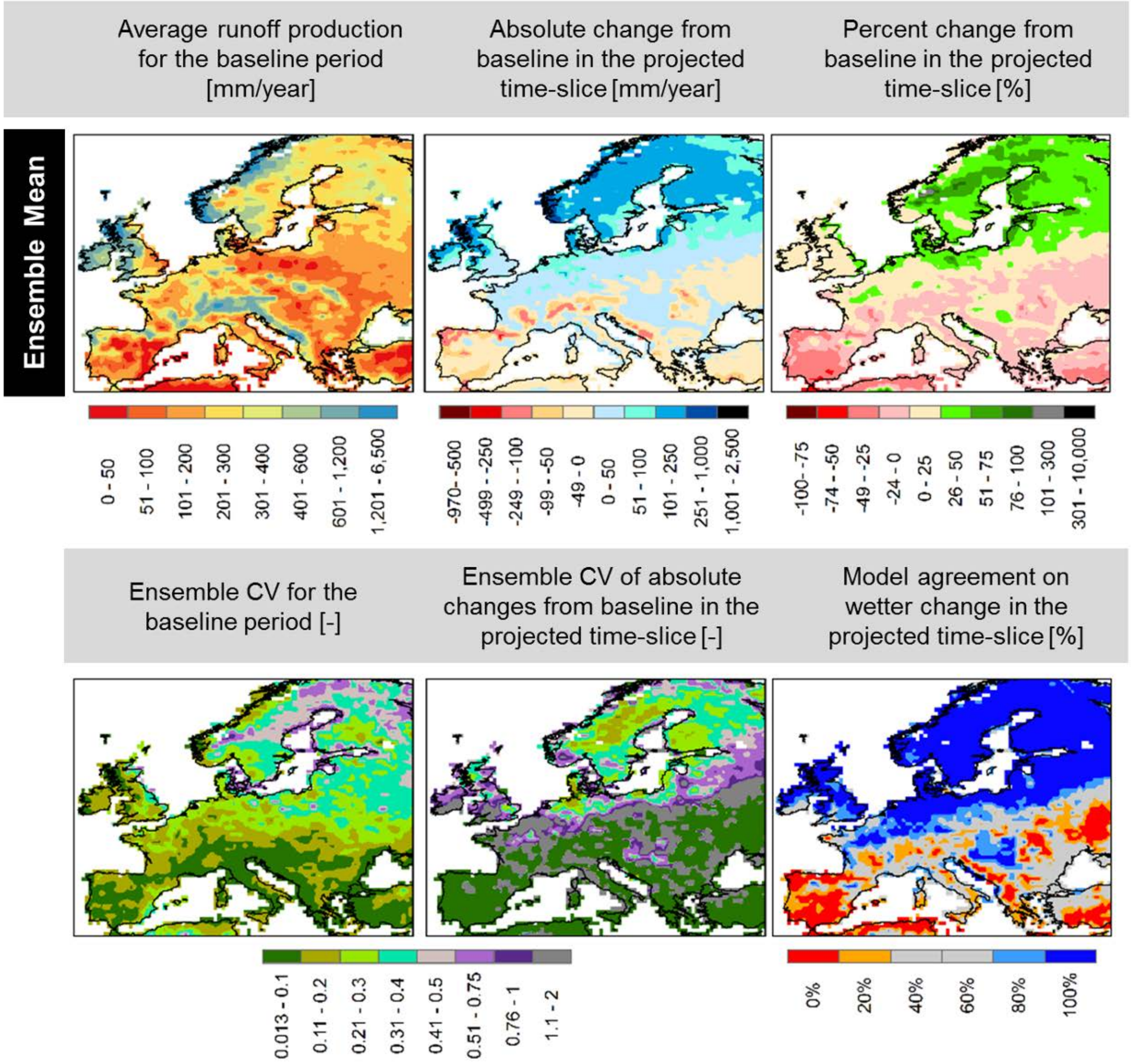


Figure 4. Ensemble mean of average runoff production from Euro-CORDEX data bias adjusted against the WFDEI dataset. Top row: Runoff production averaged over the baseline period (1976-2005) (top row), absolute (middle row) and percent change (bottom row) in ensemble mean runoff in the +4 SWL projected time-slice. Bottom row: coefficient of variation of the ensemble members for the baseline period (left column), coefficient of variation of the projected absolute changes in the +4 SWL projected time-slice (middle column) and model agreement towards a wetter change in the +4 SWL projected time-slice.

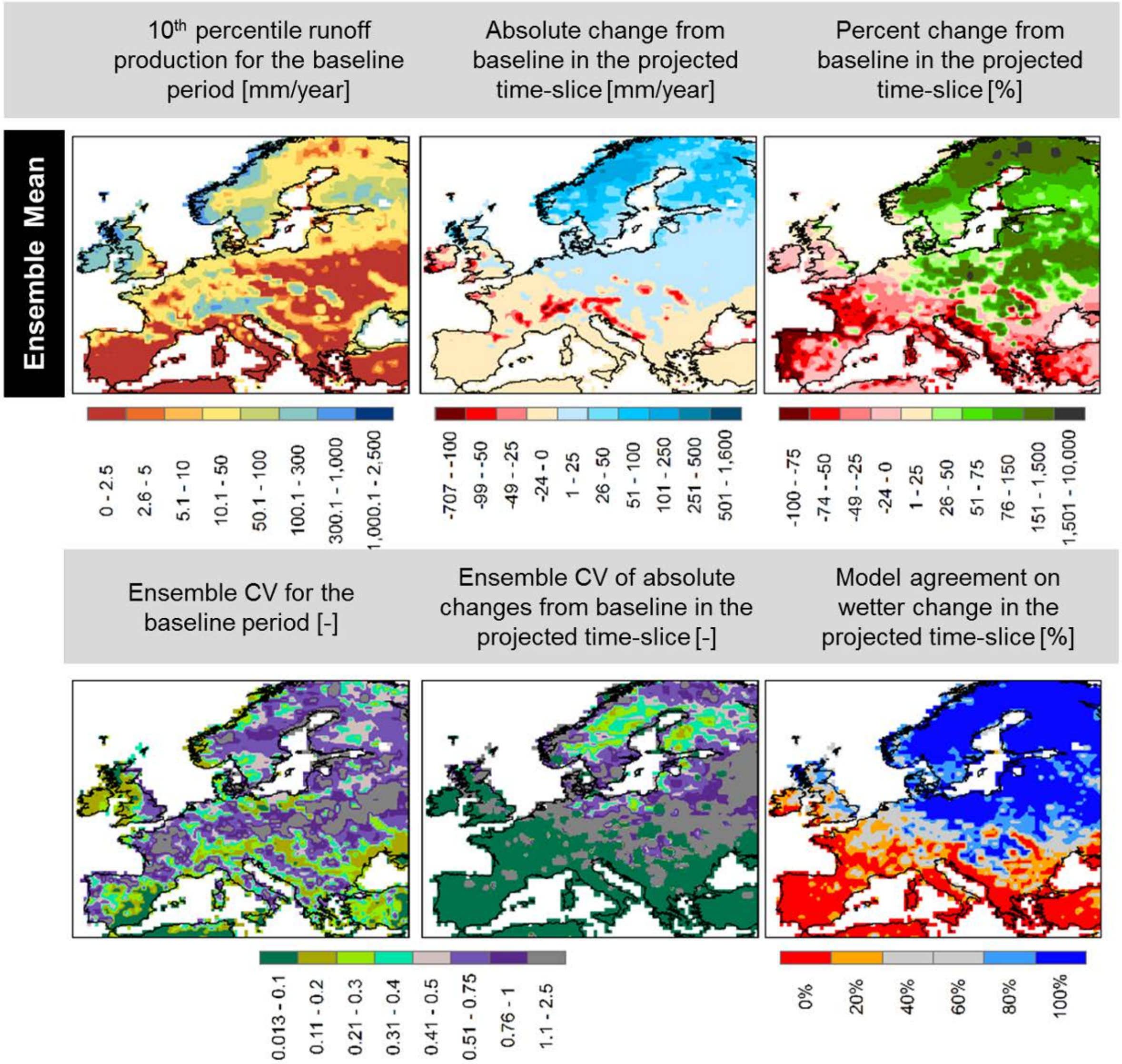


Figure 5. Ensemble mean of 10th percentile runoff production from Euro-CORDEX data bias adjusted against the WFDEI dataset. Top row: 10th percentile runoff production derived on an annual basis averaged over the baseline period (1976-2005) (top row), absolute (middle row) and percent change (bottom row) in ensemble mean runoff in the +4 SWL projected time-slice. Bottom row: coefficient of variation of the ensemble members for the baseline period (left column), coefficient of variation of the projected absolute changes in the +4 SWL projected time-slice (middle column) and model agreement towards a wetter change in the +4 SWL projected time-slice.

*14. 'with MIROC5 being the only ensemble member that expands this wetter climate down to central Europe': Looking at figure 3 I would argue that HadGEM2-ES is also showing some significant expansion of this wetter climate down to central Europe.*

REPLY: The reviewer's observation is correct. MICOC5 is distinguished in the fact that has a stronger spatial coherence in the expansion of the wetter changes towards the south and also higher values of change. In the revised version this sentence has been modified to include the HadGEM2 model: "…with MIROC5 and HadGEM2 being the two ensemble member that expand this wetter climate down to central Europe".

*15. 'Thus averaging … projected changes' (p 7282, line 2-3): Isn't this always the case with taking an average ensemble-mean?*

REPLY: It is. In the revised version this sentence is again used but as an observation rather than as a finding: "Concerning the ensemble mean, smoothing of the projected changes due to averaging has revealed clear patterns of change…".

*16. 'making it easier to identify clear patterns of change' (p 7282, line 3): Is this really the case? I would argue that an ensemble-mean might be useful but that it could also create pseudo-results (the average value is not per se the true value namely), therefore it is important to consider the full spread of GCM-forced outcomes as plausible results (unless you have information on the reliability of the different GCM-forcings).*

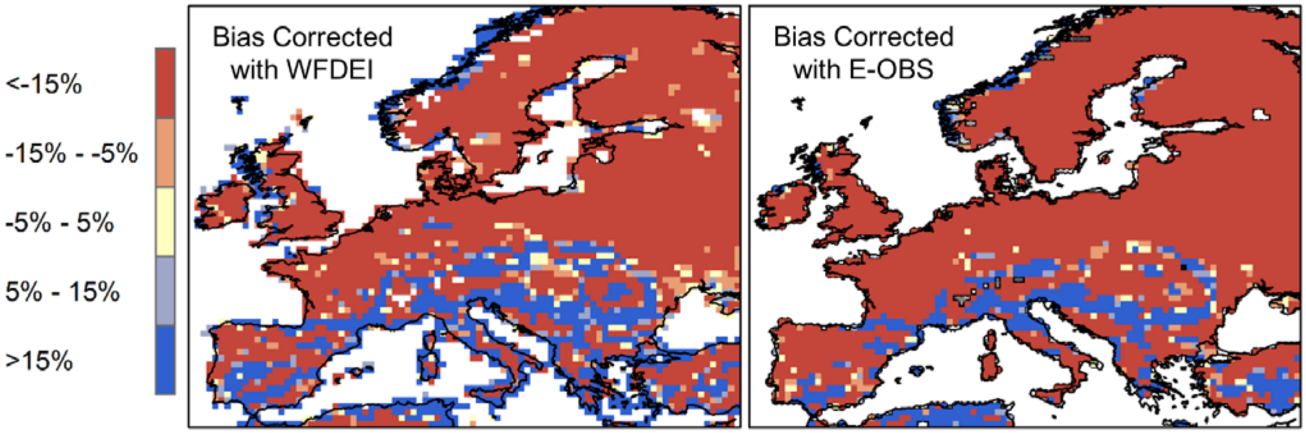
REPLY: The reviewer's comment is very legitimate. The sentence has been changed to reflect that ensemble-mean results can be misleading if they are not interpreted in conjunction with the ensemble members' agreement. Here we tackle the model agreement issue by examining both the cv of the changes (gives us information on "how much the models agree") and the agreement towards wetter of drier response ("what they models agree on"). The changed sentence is the following: "Concerning the ensemble mean, smoothing of the projected changes due to averaging has revealed clear patterns of change, which however have to be interpreted considering the full spread of the GCM-forced outcomes and the agreement between them in order to avoid misguided conclusions."

*17. 'For 10th percentile … part of Europe' (p 7282, line 16-19): Could you argue why this is the case?*

REPLY: This could be explained by the cv calculation formula (standard deviation/mean). For 10th percentile runoff both the mean and standard deviation are lower than for average runoff. If the mean of 10th percentile runoff decreases more than the standard deviation does, the computed cv for 10th percentile runoff will be lower compared to the cv of average runoff.

*18. 'bias adjustment of the forcing data resulted in a drier hydrological response from the JULES model' (p 7282, line 24-25): Could you support this statement with some numbers? E.g. xx % of the total pan-European land surface area shows a drier output using the bias-corrected forcing data compared to using the non-bias corrected forcing data, xx % shows insignificant change, xx% shows a wetter output.*

REPLY: Following the reviewer's suggestion, a Figure has been added to the ESM (Figure S3), showing the difference between bias corrected and non-bias corrected output in the baseline period. Changes between -5% and 5% have been classified as insignificant, >5% as wetter output and <-5% as drier output. The percent of pan-European area each change category occupies, along with the average value of the change (in absolute and percent terms) have been calculated. Additions based on these findings have been made in the manuscript: "Bias adjustment of the forcing data resulted in a drier ensemble mean runoff for the baseline period for 70.40% of the pan-European land surface, in comparison to 26.01% of the land area that had a wetter response after bias adjustment. The remaining 3.59% of the European area had changes that were classified as insignificant". Figure S3 is shown below:



| | Bias Corrected with WFDEI | | | Bias Corrected with E-OBS | | |
|---|---|---|---|---|---|---|
| | Drier output | Wetter output | Insignificant change | Drier output | Wetter output | Insignificant change |
| Percent of pan-European land area | 70.40% | 26.01% | 3.59% | 83.62% | 14.67% | 1.70% |
| Average percent change | -44.15% | 148.77% | -0.53% | -56.10% | 215.33% | -0.87% |
| Average absolute change [mm/year] | -231.44 | 159.37 | -2.96 | -285.70 | 131.97 | 1.49 |

Figure S3. The effect of bias correction on the ensemble mean of average runoff production for the baseline period (1976-2005). Figures: Relative difference between the ensemble means of bias corrected (left:with WFDEI, right:with E-OBS) and raw forcing data. Differences between -5% and 5% are classified as insignificant, differences <-5% as drier output and differences >5% as wetter output after bias correction. Table: percent of land area that falls into each category of change and average of the changes.

*19. 'with increases in runoff in northern Europe getting more pronounced in the runs after bias correction (p 7282-7283, line 28-1): does this hold both for the absolute and percentage change? Or only for the percentage change? And should in that case the difference be considered as a results of the baseline values becoming reduced in magnitude?*

REPLY: This applies only to percent change. The reviewer's comment is right that this is probably due to baseline values getting smaller after bias correction. In order to avoid confusion, this statement was replaced with the following in the revised manuscript: "Projected changes from bias adjusted data exhibit very similar patterns and magnitudes with the raw data derived changes".

*20. 'sign change' (p 7283, line 2): how about significance of the values? Doesn't it just all fall under 'insignificant change'?*

REPLY: Changes are in the lower region (from the class:-25% to 0 to the class 0 to 25%). However the effect that bias correction had on the baseline mean on the same areas (Figure S3 in ESM), was not classified as insignificant (apart from very few gridboxes).

*21. 'bias correction has … model agreement' (page 7283, line 3-4): That makes sense as we bias-correct all forcing data-sets using the same WFDEI data-set. Point for discussion should be whether this is actually a desired outcome (all outputs merging towards 'one single line'). How big is the confidence in the WFDEI data-set, for example?*

REPLY: This is indeed a very interesting topic for discussion as in most analyses, one has to assume that the "observations" are correct (and here the WFDEI dataset serves as observations of historical precipitation and temperature). We introduce this issue in our study by comparing the effect of two different observational datasets (WFDEI and E-OBS) on projected runoff output from bias corrected data. The subject of how much the observational datasets can be trusted and how they affect uncertainty in simulations (Biemans et al., 2009), relates to model sensitivity to input forcing. In the now running phase of ISIMIP2, historical validation simulations were performed with three different observational datasets, in attempt to study this model sensitivity and quantify the resulting uncertainty in simulations.

*22. 'For the baseline …southern Europe' (p 7283, line 6-8): Could you explain/clarify this?*

REPLY: The cv calculated for the baseline period has in general lower values than the cv calculated for the +4 SWL projected period. This makes sense as the models were bias corrected against years that include the baseline period. This sentence however has been removed as it does not apply to the new version of the manuscript (now we do not present the cv for the projected period but the cv of the differences between projected and baseline periods).

*23. 'basin average runoff production' (p 7283, line 20): Please clarify how you got these averaged values. Did you first averaged all runoff values and afterwards took the average and 10th percentile? Or did you basin-averaged all average (temporal) and 10th percentile values?*

REPLY: From the gridded runoff output, basin aggregated time-series were produced (first spatial averaging). Based on these, the average and the 10th percentile were calculated per year, resulting in annual time-series of average and 10th percentile runoff production.

*24. 'A common observation … input forcing' (p 7283,26-27): These decreases are really large. Did you find any corresponding decreases in the literature, or did you check the values with observed time-series (from gauges). Could you somehow explain these large decreases?*

REPLY: These decreases are reasonable based on the new Figure S3 of the ESM (average of negative changes after bias corrections is around 44% of 231.44 mm/year). In current Figure 10 (previously Figure 12) there is comparison between observed and modelled historical runoff production. The decreases after bias correction are probably that large because the raw data produce too high values of runoff. Bias corrected data result in modelled runoff that is closer to observations compared to non-bias adjusted data.

*25. Figure 9 is not correct. The figure shows twice the results for the Danube and Rhine whilst the results for the Elbe and Guadiana are missing.*

REPLY: This mistake has been eliminated in this revised version of the manuscript. The correct figure is shown below (the trend and p value from the regression analysis are also added):

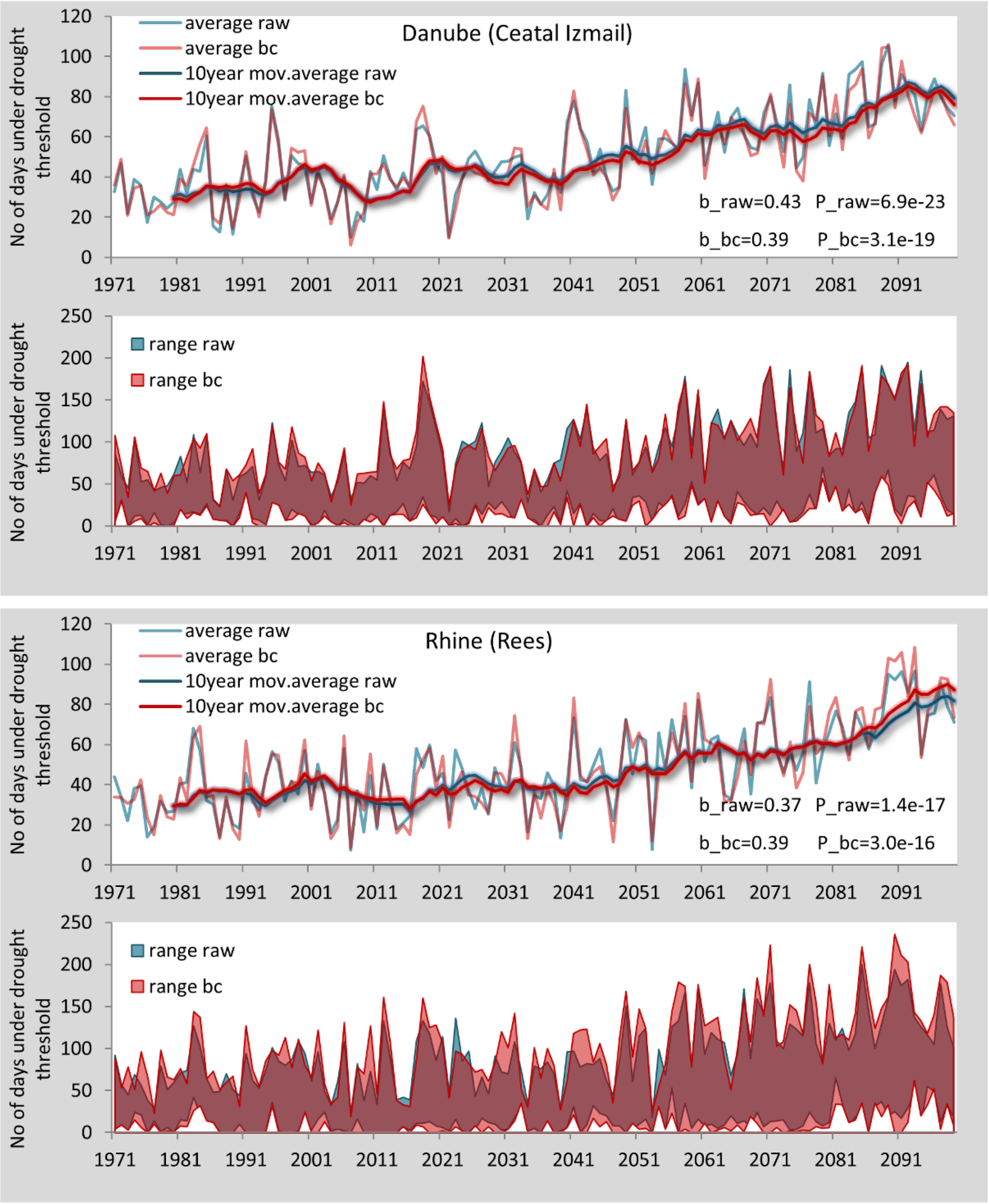


Figure 7. Number of days under drought threshold per year for raw and bias adjusted Euro-CORDEX data. Ensemble mean and 10-year moving average of the ensemble mean (top), ensemble range (bottom).

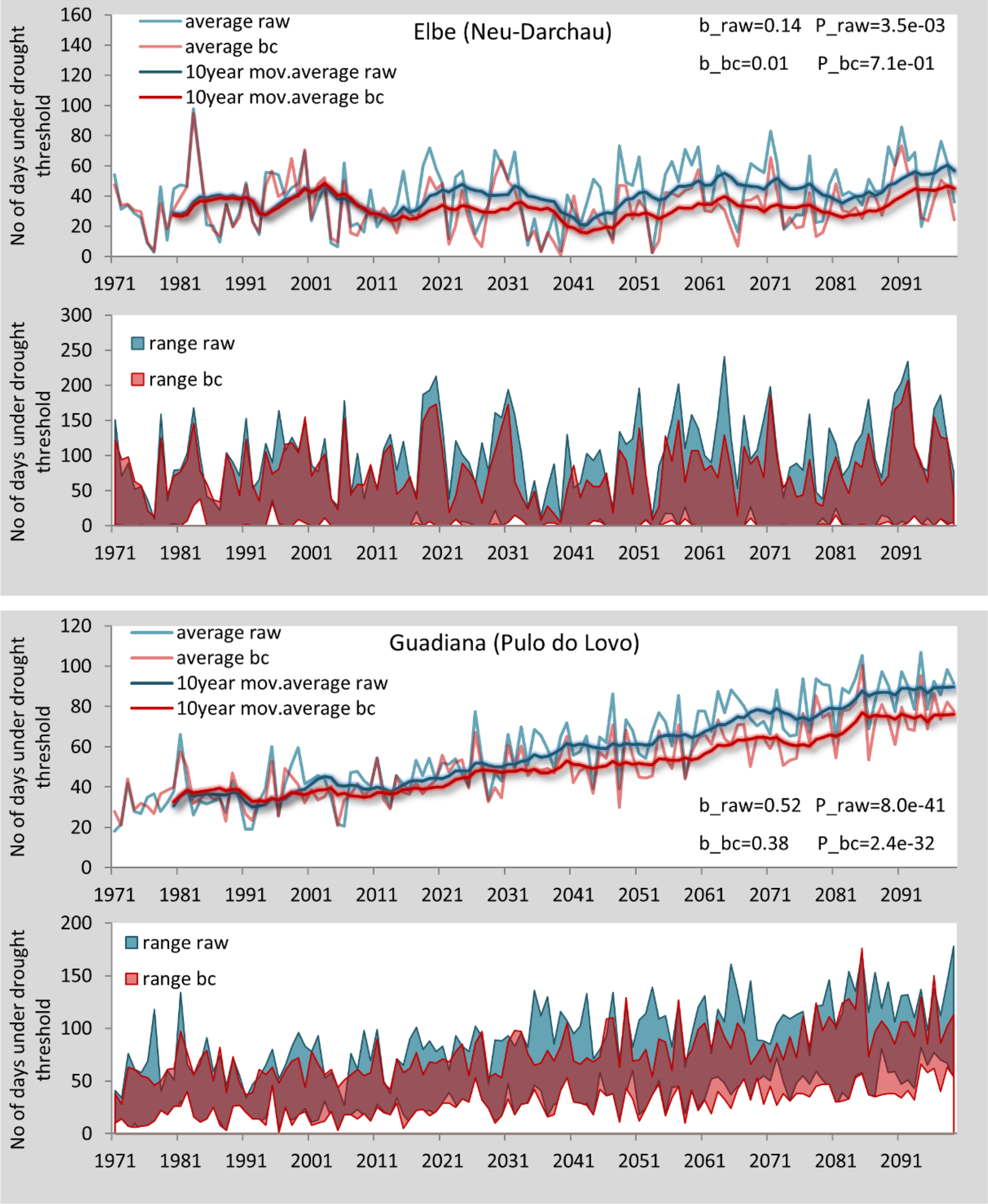


Figure 7 (continued)

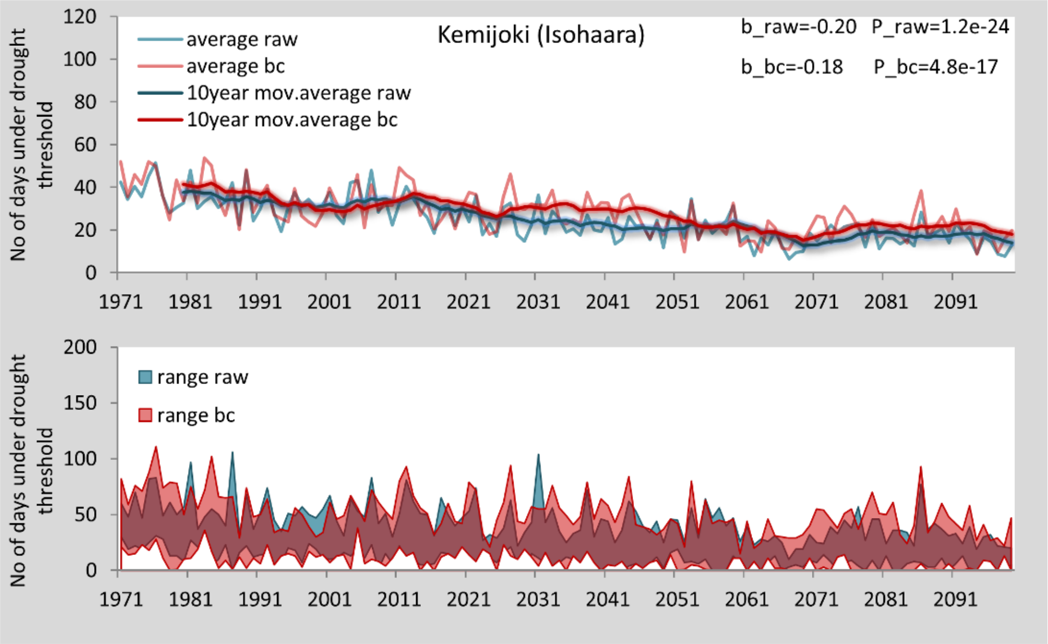


Figure 7 (continued)

*26. 'the effect of climate warming is far more pronounced for the low flows', 'significantly' (p 7285, line 16-17): Is this really significantly? Did you tested this? Please use some statistical methods to support these statements.*

REPLY: As this is not directly checked in this part, the statement was removed from the manuscript. However, having statistically checked the trend of the time-series from 1971 to 2100, we have a good approximation of the trend between the baseline and the +4 SWL time-slices. The scope of presenting Figure 8 (previously Figure 10) is to see how the basin aggregated runoff values change in between that previously estimated trend, i.e. check the values at +2 compared to +4 SWL. At the same time it allows us to compare the runoff values calculated by the different ensemble members and see how these deviate from the ensemble mean.

*27. 'there is a significant decrease from 0 to +2 C' ( 7285, line 23): How did you estimated the values for 0 degrees Celsius warming as they are, following figure 10, not equal to the baseline values. Please, clarify this in the methods section.*

REPLY: We did not estimated values for 0 degrees Celcius. 0 degrees correspond to the reference period of the pre-industrial state (1861-1880), set for the calculation of the SWLs. This means that by the baseline period (1976-2005) there has already been a change in temperature. In Figure 8 (previously Figure 10) we allocated the basin aggregated runoff values of the baseline period to their corresponding temperature change from pre-industrial state (this temperature change ranges from 0.3-0.5 $^{o}$C between the models).

*28. 'probably due to its very low values of 10th percentile runoff' (p 7286, line 7-8): In how many of the GCM runs you reach zero flow? And how reliable is zero flow for this river? Please mention this in the text.*

REPLY: For all the GCMs, for the +4 SWL projected time-slice, bias corrected data give values of around 0.001 mm/year for Guadiana while for the baseline time-slice, basin aggregated 10th percentile runoff varies between 0.0015 and 0.0033 mm/year. Raw data give higher values for the baseline period (0.004-0.73 mm/year) but project values of the same order of magnitude as the bias corrected data. Only NOAA and MIROC give values higher than the other models (0.009 and 0.004 mm/year respectively). The other three give values of around 0.001 mm/year. Thus the absolute projected values for raw and bias corrected data are very close. However, the calculated percent change from baseline gives higher values for raw data, as the raw baseline values are higher than the bias corrected ones. The close to zero values of 10th percentile runoff indicate that the river exhibits intermittent flow regime. This is relevant for this particular river, as it is located in a semi-arid region and intermittent flows typically characterize its hydrological regime (Filipe et al., 2002; Collares-Pereira et al., 2000; Pires et al., 1999). Given the changes that are projected for the Iberian Peninsula at +4 SWL, it is expected that the intermittent flow regime in Guadiana might intensify.
A piece concerning this topic was added to the discussion section of the revised manuscript.

*29. 'E-OBS corrected data … the observed values' (p 7287, line 11-14): Could you think of an explanation for this observation? Are there any differences between E-OBS and WFDEI that might clarify this result?*

REPLY: The two newly added Figures in the ESM (Figure S1 and S2), describe the effect of bias correction on the forcing variables and help us understand the differences between the two observational datasets. (Figure S1 shows the effect of bias correcting against the WFDEI dataset and Figure S2 against the E-OBS dataset. Results are shown for precipitation and temperature as the rest of the forcing variables were not bias adjusted. The absolute differences between bias corrected and raw input (bc-raw) are shown for all the participating GCMs and for their ensemble mean. The cv between the ensemble members before and after bias correction has also been calculated. In each sub-figure, the spatial average of each illustrated map is noted in each sub-figure.)

The precipitation differences after bias correction with E-OBS are quite larger than the differences after bias correction with WFDEI (for ensemble mean the values are approximately 2.3 times bigger), meaning that E-OBS corrected precipitation has lower values than the WFDEI adjusted precipitation. It is thus reasonable that this difference in precipitation will also reflect on the output of the hydrological model, producing lower runoff values when forced with the E-OBS bias adjusted dataset.

*30. Section 3.6 focuses on the basin averaged average runoff whilst I think it is (more) interesting to show also the results for the 10th percentile runoff and using not the basin-averaged numbers. Optionally figures could be placed in a supplementary.*

REPLY: A Figure of the same analysis described in section 3.6 but for 10th percentile runoff has been added to the ESM (Figure S6). The following piece of text was added to Section 3.6:

“From the application of the same analysis on 10th percentile runoff production (Figure S6 of the Supplementary Material), it is deduced that for the low flows the E-OBS corrected data again produce lower values of runoff compared to WFDEI. In this case, however, even the raw forced output (which is wetter than the bias corrected) underestimates the observed 10th percentile runoff values. Regarding the percent projected changes, results from bias corrected data produce smaller values compared to the raw data while E-OBS adjusted data result in decreased changes compared to output from WFDEI adjusted forcing.”

*31. Could you elaborate a bit more on how the differences between the two bias corrected data-sets and their hydrological output develops, comparing the difference in input data (precipitation, temperature) with the order of magnitude differences in hydrological outputs (local and routed runoff)?*

REPLY: As mentioned in the response to comment 29, the precipitation differences after bias correction with E-OBS are approximately 2.3 times larger than the differences after bias correction with WFDEI (for the ensemble mean, comparing spatially aggregated differences). For the historical period, the ensemble mean of precipitation differences for WFDEI bias corrected results is -91.17 mm/year compared to -209.95 mm/year for EOBS bias corrected results.
Using the values of Table 2 we can make a rough comparison of the effect bias correction has on input precipitation and output runoff. For WFDEI forced runs, the average of the difference between bias corrected and raw forced output for the five basins, for the historical period is around -135 mm/year. The respective number for EOBS adjusted data is -197 mm/year. Runoff differences after bias correction with E-OBS are about 1.5 times smaller than the differences after bias correction with WFDEI. Comparing this with the respective number characterizing precipitation differences (2.3) we could argue that the effect of the different observational dataset is weakened in the hydrological model output. Another comment could be that the effect of bias correction on input precipitation and output runoff are of the same order of magnitude (-91.17 mm/year vs -135 mm/year for WFDEI and -209.95 mm/year vs 197 mm/year for E-OBS).

*32. ‘it is … model agreement’ (p 7287, line 14-15): Incomplete sentence*

*33. ‘changed’ (p 7288, line 9): change*

*34. ‘it is … climate change’ (p 7289, line 1-2): please leave out as this is only deduced and not studied.*

REPLY: The flood related unsupported statement was removed from the manuscript.

*35. ‘It should be … average-state’ (p 7289, line 4-5): Please back-up with some references.*

REPLY: Similar results with increased model spread expressed as cv for low flows compared to average state flows were found by Koirala et al., (2014).

*36. ‘of’ (p 7289, line 15): delete*

*37. ‘remarked’ (p 7291, line 25): remarkable/significant*

REPLY: “Remarked” was changed to “remarkable”

*38. ‘thus’ (p 7292, line 1): and*

*39. ‘are expected’ (p 7292, line 2): please remove floods. Moreover I wouldn’t say this so frankly, replace with: ‘could be expected’*

REPLY: The statement about floods was removed and the replacement the reviewer suggested was done.

*40. ‘two degrees warming’ (p 7292, line 12); +2 SWL (p 7292, line 14): please be consistent in naming*

REPLY: Here we wanted to make a statement on the effect that “two more degrees of warming” (reaching +4 SWL) has on the change already documented at +2 SWL. To clarify this the last sentence was modified to the following:

“For the rest of the European region where trends are not clear or ensemble members do not agree towards the change, the effect of the further warming from +2 SWL to +4SWL, does not seem to severely affect the hydrological state, which is however already significantly altered at +2 SWL compared to pre-industrial.”

*41. Table 1: (1) All RCMs are RCA4 - column could be deleted. (2) Is it of interest to present the equilibrium climate sensitivity?*

REPLY: The column containing the RCM was deleted as suggested by the reviewer. We believe that the ECS is of interest as it gives some information on the behavior of the GCM.

*42. Table 2 and 3: Absolute and percentage change to what? Please clarify in table or heading.*

REPLY: Absolute and percent change refer to changes from baseline in the projected time-slice. Relative clarifications have been added in the Tables.

*43. Please merge figure 2 & 4 (3 & 5): add one/two rows to figure 2 and 3 to show the ensemble-mean changes and the modelling agreement.*

REPLY: Following the reviewer’s indication Figures 2 -5 have been restructured. Figure 2 was merged with Figure 4 (and Figure 3 with Figure 5). Moreover, the CV for the projected period was substituted with the CV of the absolute differences. Finally, a sub-figure showing model agreement towards a wetter change in the projected time-slice has been added to the two new Figures.