We first would like to thank the reviewer comments which will help to improve our manuscript.

These results are certainly interesting and potentially important, but I have two main concerns. First, I found the paper very hard to read, mainly because of the writing style and the fact that the authors decided to show nearly all results rather than making an informed decision on the most relevant parameters and results.

We agree that a lot of results and statistics are presented. In function of the other reviewers comments, we could put in Supplementary Material the validation of the melt extent as well as several tables.

Second, the results are described in a merely qualitative fashion, without a more quantitative analysis of the reasons of the substantial differences between the various model runs. As a result a clear conclusion is lacking as to the quality of the re-analysis products for reproducing Greenland surface mass balance.

We think that the differences between the MAR simulations are well explained in respect to differences in the forcing reanalysis and the conclusions list the best reanalysis (pg 12, lines 14-15). However, we agree that additional general conclusions will be needed in the revised version of our manuscript. Eg: Which are the best reanalysis ? Over which periods the results are reliable ? … This is clearly mentioned in the revised conclusion of our manuscript.

These concerns are described in more detail below, together with some textual comments where clarification is needed (listing not complete). Major revisions will be necessary to bring the paper to a publishable level.

Major comments

The paper is very difficult to read, owing to the multitude of model acronyms and the overwhelming number of figure frames and tables contents. Already early in the paper I got confused by all MAR model versions presented. On page 3 alone, there is mention of MAR, MAR3.5.2, MARv2, MARv3.x, MARv3.2.

Between Fettweis et al. (2013) using MARv2 and this paper using MARv3.5.2, some MAR biases have been identified in several papers using intermediate MAR versions. Therefore, we judged important to list these biases (which have been corrected in MARv3.5.2) and to clearly mention the different MAR versions used in the text. Several papers used MAR outputs as comparison without giving the model version while each one has its own advantages and drawbacks. By explicitly listing here the MAR version used, we aim to inverse this trend.

. ... Some more sobering statistics: the manuscript contains six tables with about 500 numbers, sixty lines in line graphs, more than 40 maps, and the acronym MAR is used almost 300 times!

As proposed above, some comparisons can be added in the Supplementary Material. In addition, in function of the other reviewer remarks and editor suggestions, some less interesting MAR simulations (MAR forced by NCEP2, by JRA-55, by ERA-20C not corrected, by 20CRv2 not corrected) could be put in Supplementary Material although it is a pity to mask a part of these results showing sensitivity of the reanalysis used.

Readability can be further improved by not combining multiple results in a single sentence. Moreover, results should be easily traceable in figures. This now is not always the case. For example, p. 9, l. 17 reads: “MAR also underestimates accumulation in the south-east versus ice cores but overestimates versus BOX13 because this data set is based on RACMO2 outputs which are known to underestimate accumulation in this area (Noël et al., 2016).” This sentence starts by stating that MAR underestimates accumulation in the southeast compared to ice cores. But in Fig. 6c I see red colours,
indicating an overestimation? The sentence continues that MAR overestimates (accumulation) vs. BOX 13, but Fig. 6c shows red and blue colours? Then it concludes that BOX13 is wrong in this area because RACMO2 outputs are underestimating accumulation in the southeast. So what is the conclusion?

Sorry, there is a mistake in the text. It is at the north-east and not at the south-east. In respect to Fig 6b (using observations only), most of the comparisons with ice cores in the accumulation zone are blue at the north-east while, over this same area, MAR overestimates BOX13 estimations (which is based on model results and not observations). Only ice core observations (shown with blue cross in Fig. 6a) are discussed here. The conclusion is that both MAR and BOX13 underestimates accumulation in this area. These sentences will be rephrased to be clearer in the revised version and the fact that only the accumulation observations are discussed here will be explicitly mentioned in the revised manuscript.

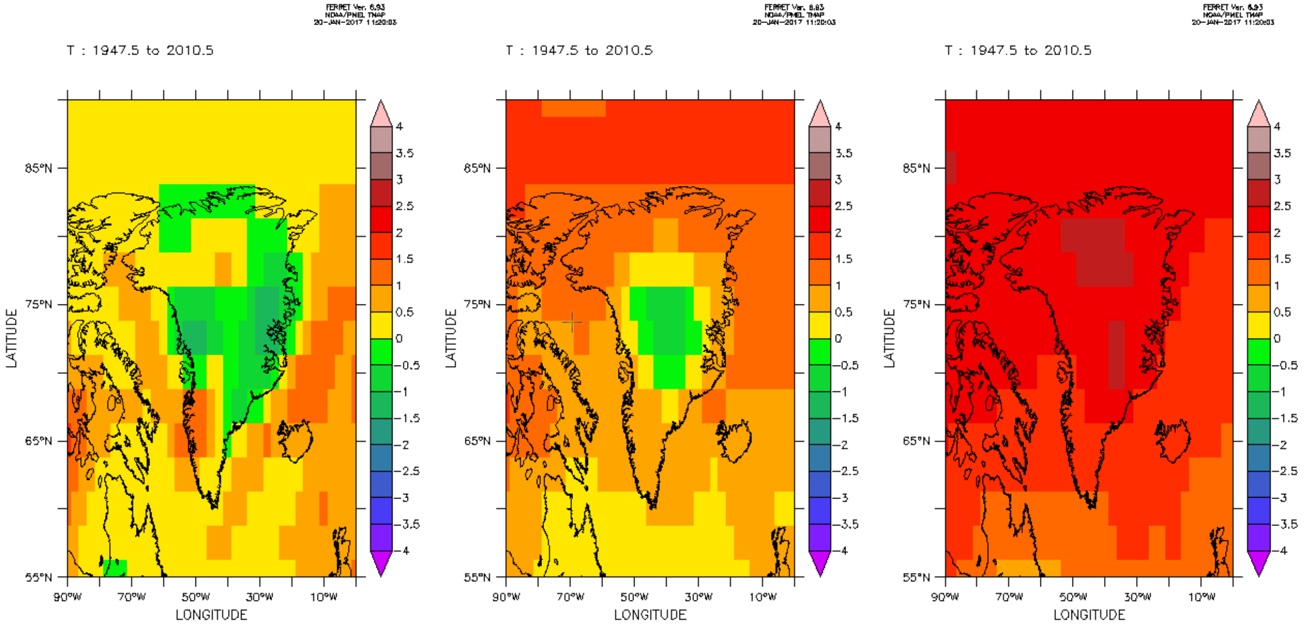
p. 3, l. 13: how can SMB be robustly calculated outside the ice sheet mask? If ice is assumed to be present in a certain grid cell that is currently tundra, then the local climate would be altered (cooled) and hence SMB would be influenced?

Indeed, there is a small influence but this impact can be neglected. For each MAR atmospheric pixel, there are 2 surface sub-pixels covered by tundra and permanent ice. At each time step, the surface energy balance and surface temperature are computed for each sub-pixel by the SISVAT surface model. They are averaged afterward over the whole atmospheric pixel for being used into MAR as input. Knowing that the average is weighted by respective cover of each sub-pixel and that the permanent ice cover over the "true" tundra is 0.001% (corresponding to the FORTRAN precision), we can resonantly assume that this impact is not significant in respect to a full tundra covered pixel.

This strategy allows to estimate SMB outside the MAR ice sheet mask (which is dependent of the spatial resolution) in the aim of forcing ice sheet models afterward. However, it is important to note that SMB for these "dominant" tundra pixels is underestimated because the near-surface temperature over these pixels is more representative of the tundra conditions than the ice sheet conditions and therefore, the melt is overestimated but it is better than nothing.

p. 3: Selecting the 700 hPa temperature as predictor for melt appears unfortunate, because this pressure level intersects with the ice sheet surface. This implies that T at 700 hPa at the lower parts of the ice sheet represents a free atmosphere temperature, at intermediate parts of the ice sheet it represents a boundary layer value (temperature inversion) and at the highest levels a below-surface (extrapolated?) value. This is confirmed by the blue lines in Fig. 1a, which show a conspicuous local minimum over the higher parts of the ice sheet, clearly caused by the ice sheet surface, a feature that would not be expected if a higher level (500 hPa) had been selected. The authors partly recognize this problem by masking out the level below 2000 m in Fig. 1, but that is again unfortunate because melt also does take place above this altitude so part of the interesting information is lost.

MAR is not directly sensitive to free atmosphere temperature biases over the Greenland ice sheet but rather at its lateral boundaries where the 850 hPa level is even more relevant as most of the melt occur below 1500m. But this level will be too masked by the ice sheet topography to be shown on a figure. The 500 hPa level is above the ice sheet summit but biases at this level less impact the melt amount simulated by MAR. Therefore, the 700 hPa is a good compromise to be shown although this level crosses the ice sheet topography. Anyway, the choice of the 600 or 700 hPa level to evaluate the melt variability is justified in Fettweis et al. (2013) and evaluations at other levels (850 and 500 hPa) than 700hPa will be provided in the Supplementary Material of the revised version.



20CRv2 – NCEP1 at 850hPa

20CRv2 – NCEP1 at 700hPa

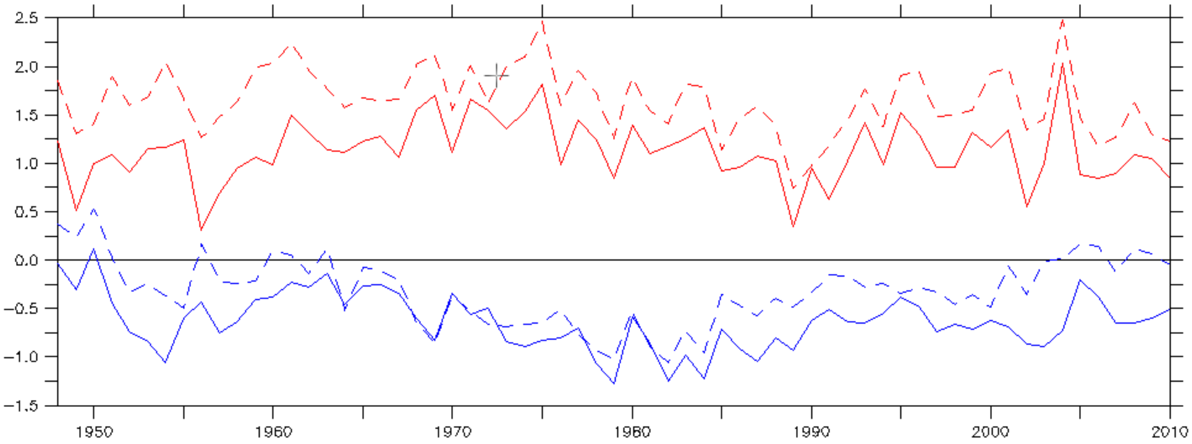
20CRv2 – NCEP1 at 500hPa

On the figure above, you can see the mean difference over **1948-2010** of the JJA temperature at 850hPa (left), 700hPa (middle) and 500hPa (right) of 20CRv2 in respect to NCEP-NCARv1. We can see that 20CRv2 is too warm at each level.

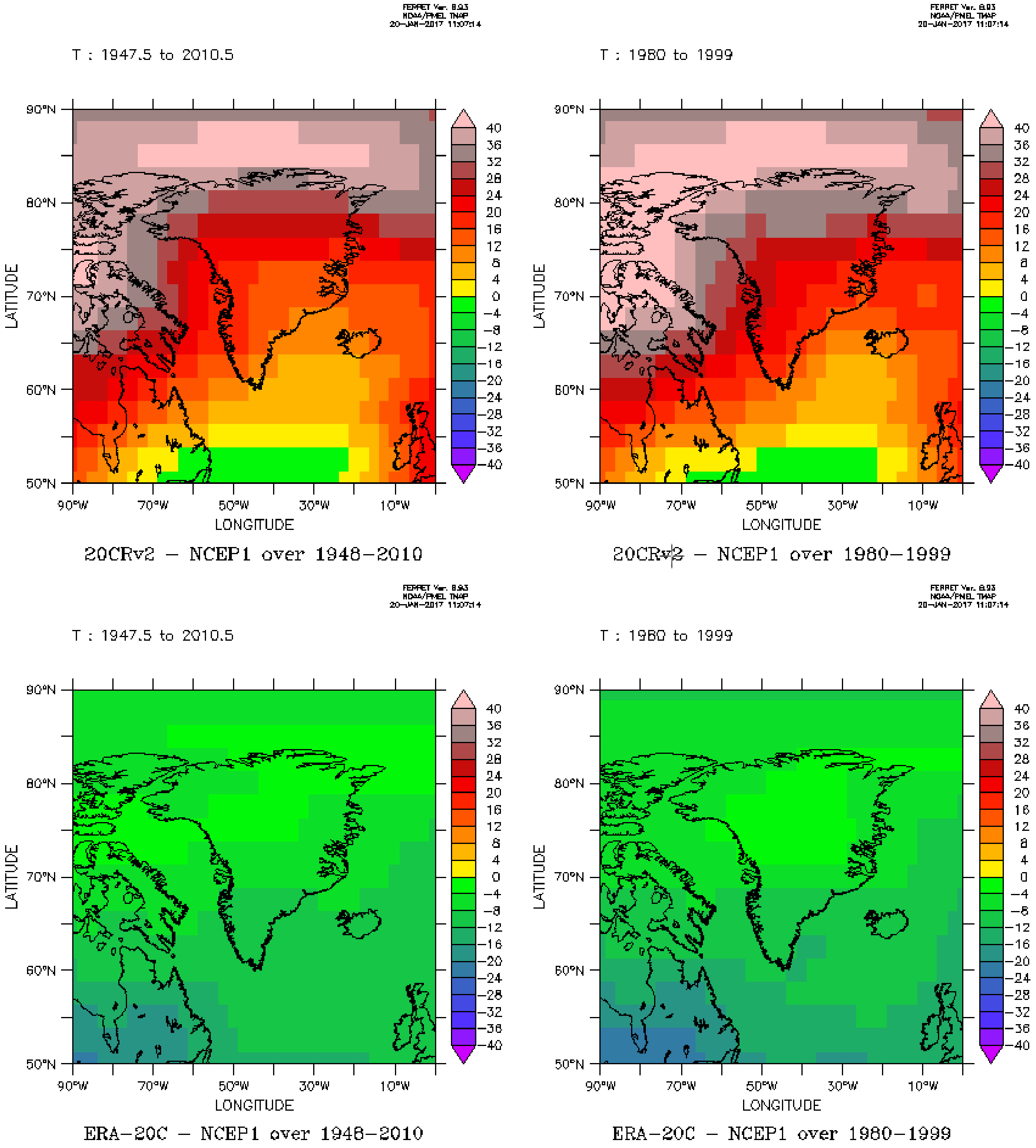
p. 4: Why using the obsolete 20CRv2 at all when an improved/corrected product is available (20CRv2c)?

Using both 20CRv2 (where a temperature correction is needed) and 20CRV2c (where no temperature correction is applied) as forcing allows to show the sensibility (in particular at the beginning of the last century) of MAR results to very similar forcings as well as, in the same time, the uncertainties in the reanalysis. Knowing that there are only 2 available reanalysis before 1950, that the MAR results forced by 20CRv2 and 20CRv2c are significantly different before 1950, showing both time series is relevant for us. To improve the clarity of several figures, we could remove the JRA and NCEP2 forced time series as these simulations do not bring any new stuff in respect to the ERA and NCEP1 forced time series.

p. 5: A temperature correction of +/- 1 C is applied to 20CRv2 and ERA-20C, based on a 1980-1999 comparison with ERA-Interim. Can one be sure that these biases have been constant before and after this period? Can you show how stable this bias has been in time for the full period 1958-2015 for which ERA-Interim and ERA-40 are available, which appear to perform adequately? Why is this only done for NCEPv1 (line 26)? Why are these results, which are instrumental for the interpretation of the results in this paper, not shown? The same question applies to the height of the 500 hPa level displayed in Fig. 2. How constant are these biases in time over the period with reliable forcing (∼1950-present)?



On the figure above, you can see the mean difference over [90°W-0°W, 50°N-90°N] of the JJA temperature at 700hPa (in solid) and 500hPa (in dash) from 20CRv2 (in red) and from ERA-20C (in blue) in respect to NCEP-NCARv1 Reanalysis (1948-2010) (Unites are °C). As shown in Figs 1-2, NCEPv1 compares very well with ERA-Interim and ERA-40 (not shown here). We can see that the biases vary a few with time but 20CRv2 is every time too warm while ERA-20C is too cold. It is clear that applying a same correction over the whole period could induce additional biases because the quality of the reanalysis is not constant in time (we have no information to evaluate this one before 1950). Nonetheless, a constant correction over time allows to compare different periods, knowing that we can assume that the corrected time series is homogeneous. This issue will be more discussed in depth in the manuscript and remember in the conclusion. Finally, it should be noted that no correction is applied when MAR is forced by 20CRv2c.



On the figure above, you can see the mean difference over **1948-2010** (left) and over **1980-1999** (right) of the annual mean geopotential height at 500 hPa from 20CRv2 (top) and ERA-20c (below) in respect to the NCEP-NCARv1 reanalysis (units are meters). As 20CRv2 (resp. ERA-20c) is too warm (resp. too cold), 20CRv2 (resp. ERA-20c) overestimates (resp. underestimates) the annual geopotential height at 500 hPa. However, these biases are constant in time and not impacted by the reference period chosen.

Similar figures will be added in the Supplementary Material of the revised version of our paper to valid our assumption.

p. 7: To improve the logical sequence of the paper, the evaluation of the ERA-Interim forced run (section 4) should be presented before the other results in Figs. 1-3.

An other solution should be to start by showing MAR results (Figs 3-4) and afterward explaining the differences between the MAR simulations in respect to discrepancies in reanalysis (Figs 1-2). But, we prefer to keep the logic used in Fettweis et al. (2013) where the forcings were discussed first. In addition, as MAR results using corrected reanalysis (20CRv2 and ERA20C) are shown in Fig. 3-4, keeping this sequence allows to justify afterward to show MAR results using corrected and not corrected reanalysis as forcing.

Figure 5: Knowing that absorbed shortwave radiation is the main energy source for melting, it is quite remarkable that ablation is so well reproduced in Fig. 5c, while surface albedo is clearly too high in MAR, by up to a factor of 2. This would imply that, all other things being equal, ablation would be significantly overestimated had albedo been correct. How can this be reconciled, are there compensating errors?

There are obviously compensating errors as it is the case in each climate model. MAR is firstly tuned to successfully simulate the SMB and ablation but not albedo. In this version (3.5.2) of MAR, MAR overestimates the bare ice albedo but underestimates longwave (LWD) and overestimates shortwave (SWD) because it underestimates the cloudiness (like RACMO for example). However MAR is able to successfully simulate ablation and (near) surface temperature, meaning that its Surface Energy Balance (SEB) is OK. But it is obvious that there are biases in the individual fluxes of SEB which, when they are summed, are compensated. In the next version of MAR, there are more clouds reducing the biases of SWD and LWD and allowing to have a lower bare ice albedo value better in agreement with that the observed one.

This problem of compensating errors will be explicitly mentioned in the revised version of our paper in Section 4.1

Textual comments
p. 1, l. 9: validated -> support
ok, thanks. This will be corrected in the revised version of your paper.
p. 1, l. 13: The period 1961-1990 is not commonly chosen as reference period because there was approximate balance, rather it is the only official climatological period that can be chosen before significant changes occurred in Greenland.
ok
p. 1, l. 19: “stationarity assumption” Unclear, please explain.
We mean here that the SMB has been quite stable after 1930 until the end of the 1990's. This sentence will be rephrased.

p. 1, l. 20: “. . .only suggests. . .” Unclear, please explain. Does ‘only’ refer to ‘suggests’ of the 1920-1930 warm period?
We mean here that only the ERA-20C forced simulation suggest that … and not the 20CRv2(c) forced simulations. This sentence will be rephrased.
p. 1, l. 20: last sentence of abstract contradicts earlier statement of “..unprecedented melt. . .” after 1990.
SMB in the 1930's was "comparable" to SMB anomalies observed now but it was the results of BOTH positive anomaly of melt and negative anomalies of accumulation while currently, only melt anomalies drive the current SMB anomalies. Therefore, the current melt rates are well unprecedented.

p. 2, l. 5: Please also cite studies that imply a connection between subglacial meltwater injection and frontal ablation of marine terminating glaciers.
ok
p. 2, l. 10: considered -> mentioned (?)
ok
p. 2, l. 11: could not be -> is maybe not
ok
p. 2, l. 19: attractive -> useful, powerful, robust (?)
"powerful" is a better word indeed.
p. 2, l. 23: validated -> evaluated (please use this throughout manuscript: models are by definition an approximation of reality and can therefore not be validated)
ok
p. 3, l. 12: factional –> fractional
ok
p. 4, l. 7: all of the. . ..I think it cannot be claimed that ‘all’ observations are really used.
Perhaps the greatest fraction? Can this be supported by a reference?

greatest fraction is indeed more adequate here.

p. 4, l. 12: surface marine winds -> near surface winds over the ocean surface
ok
p. 4, l. 13: “As this reanalysis assimilates much less data than ERA-40/ERA-Interim,” Is this also true for the overlapping period?
yes
Reliable -> accurate, reliability -> accuracy. Has the increase in accuracy been published in literature?
Yest, in the ERA_20C paper (Poli et al.,2016) which will be referenced here.
p. 4, l. 16: “covering the half of the last century” Second half.
ok
p. 5, l. 3: Confusing: here summer T at 600 hPa is mentioned, in line 6 summer T at 700 hPa.
ok
p. 5, l. 7: ‘drives’. I suspect that what you mean to say is that T at 700 hPa is a good predictor for melt variability in MAR?
Yes. This sentence will be rephrased.
p. 5, l. 13: “Surprisingly, the comparison is worse with the 2nd generation of the NCEP reanalysis, which is warmer than ERA-Interim in summer except at the South-East of Greenland” I don’t see this in Fig. 1f: the southeast is also too warm?
It is at the southeast of the domain presented here (Iceland). This sentence will be rephrased.

p. 5, l. 17: “. . .too warm (see Fig. 1c) and too cold (Fig. 1g). . .” I assume ‘warm’ and ‘cold’ must be swapped in this sentence.

The reference of the figure must be swapped indeed.

p. 5, l. 21: Why is the performance of 20CRv2c not discussed here? It was not corrected? Using the acronyms ‘20CRv2-corr’ and ‘ERA-20C-corr’ is somewhat confusing because the ‘c’ in ‘20CRv2c’ presumably also stands for ‘corrected’.

The performance of 20CRv2c is discussed p5, line 28-20. We agree about he confusion 20CRv2c and 20CRv2-corr. The corrected reanalysis could be write CORR-ERA-20C and CORR-20CRv2.

p. 6, l. 13: “Both ERA-20C forced simulations also significantly underestimate precipitation along the south-western coast. . .” This does not become clear from Fig. 4d.

Latitude boundary will be added in the text to well specify that it the "south" part of the south-western coast.

p. 6, l. 19: “However, both simulations underestimate precipitation along the south-east coast with respect to MAR_ERA−Interim.” But these deviations are also mostly hatched, i.e. are not significant according to the definition used here?

These biases are not significant indeed. This sentence will be rephrased.

p. 6, l. 26: the same -> similar

ok

p. 6, l. 28: What does the ‘+40%’ mean?

40% means the runoff overestimate of MAR forced by 20CRv2 without correction. This sentence will be rephrased.

Caption Fig. 1 and elsewhere: Celsius degrees -> degrees Celcius

ok

Caption Fig. 2: Please include explanation of the wind vectors in these plots, do they represent anomalies in the wind field?

They represent well anomalies of wind field. The caption will be updated.