Reply to the review by Anonymous Referee #2

We thank anonymous referee #2 for the positive review, which suggests useful additions to the manuscripts. These suggestions are highly appreciated. Answers to the comments are included in blue font right under the unmodified comments from the review.

REVIEW

Review of 'weather@home 2: validation of an improved global-regional climate modelling system' by Guillod et al. The paper is a useful and fairly thorough documentation of the w@h system with a focus on Europe, although some gaps remain that can easily be filled. Also, there are some qualitative statements that can be converted to quantitative ones with, I think, relatively minor effort. I am looking forward to a more complete version, which would be very informative.

We appreciate the overall positive tone of the reviewer's comments as well as the relevant points raised, for which we mention our intentions for the revised manuscript.

Major comments

1. Can you comment how the biases in the global model compare to other, state-of-the-art, GCMs, eg in Chapter 9 of the IPCC WG1 AR5?
   Following the referee's comment, we have had a closer look at w@h biases and reproduced a few figures of the mentioned IPCC chapter for our model. Figs. R1, R2 show the absolute error in annual-mean temperature and precipitation, and can be directly compared (visually) to the following figures in Chapter 9 of the IPCC WG1 AR5: 9.2(c) and 9.4(c), respectively. Overall and for both variables, HadAM3P performs similarly well to state-of-the-art CMIP5 models (despite being an older model, but possibly because of the prescribed SSTs). Larger errors in temperature are however found in HadAM3P over Greenland and Eastern North America. For precipitation, no such hotspot or error is found. For a regional and seasonal quantification, we have reproduced Fig. 9.39 for SREX regions (Fig. R4, but note that the regions in a different order than the IPCC AR5 Fig. 9.39). Here, too, HadAM3P performance looks similar to the CMIP5 models, with a few cases where HadAM3P biases are larger. We have therefore added the following two sentences, at the end of the two paragraphs dealing with temperature and precipitation biases in Section 3.1: "For most regions, the performance of HadAM3P is similar to state-of-the-art coupled climate models from CMIP5 (Flato et al., 2013)" and "Like with temperature, the model performs similarly to typical CMIP5 models (Flato et al., 2013)".

2. For attribution, a correct representation of variance is as important as trends (Uhe et al, 2016). Please add the equivalent of Figs 1-3 for the variance, preferably of daily data but monthly should be OK if these were not saved. In that case CRU-TS can also be used as ground truth, for daily data Berkeley Earth has temperature fields and CPC precipitation fields over the required period.

   Daily data were not saved in the global model. Therefore, we have followed the referee's suggestion and have computed bias maps of the standard deviation of monthly averaged temperature, precipitation and 500hPa geopotential height. Due to the large number of figures in the paper, these are placed in the Supplementary Information (Figs. S1, S3 and S4), but the main text (Section 3.1) mentions the main results from these.

3. Section 4.3. It would be useful to explicitly comment to what extend the biases in extremes can be corrected by a simple additive (temperature) or multiplicative (precipitation) bias correction.

   We thank the referee for this useful suggestion. We have added some comments on bias correction and the suitability of various techniques in the last paragraph of Section 4.3, with a direct reference to the quantile-quantile plots.

4. Section 4.4 Given the strong connection between the reliability and trends, please add trend maps of the observations and model results in addition to the reliability diagrams, preferably also with SLP trends.

   We appreciate the referee's suggestion to add information about trends. However, we have found that trend maps are highly variable within the ensemble, and using the ensemble mean leads to a spuriously smoothed spatial pattern, as the effect of internal variability is removed. Therefore, rather than trend maps, we have chosen to show trends for regional averages of temperature and precipitation (Fig. 17), which allows us to display the spread in trends from individual w@h2 time series (constructed by randomly sampling 1 ensemble member per year). We have also added some text in Section 4.4 related to that figure, and have renamed that section "Reliability and trends".

Minor comments

- p.5 l.30 Why is Z500 taken from the ancient ERA40 reanalysis rather than a more modern one? JRA-55 covers the period 1961–1990.

  We have replaced ERA40 with JRA-55 in the manuscript.

- p.6 l.8 "30 years period from 1961–1990". I understand that this is dictated by the short runs of w@h1. Can you add a comment on how different the biases of w@h2 are over the whole century?

  We have also plotted the w@h2 bias maps for the time period century (1900–2006), and they look very similar to the ones with years 1961–1990. We have therefore added the following sentence at the beginning of Sect. 3: “w@h2 biases look very similar when the whole time period, from 1900–2006, is considered”.

- Almost all figures would be more intuitive for readers with a left-to-right script if w@h1 was plotted to the left of w@h2.
We follow the referee’s suggestion in all figures.

- Please show Fig. S1 in the main text instead of Fig.3 as it is much more informative.
We agree that precipitation bias maps tend to over-represent wet regions when shown in mm/day. However, relative biases (in %) tend to over-represent dry regions in a similar way. After having tested both ways, we would like to keep this as is since we find that this is best for the discussion of the biases in Section 3.1.

- p.7 l.8 “suggesting that certain modes are not well represented”. To be nit-picking: misrepresentation of modes will affect the variability much more than the mean state. Just delete, as it carries no useful information.
The sentence was replaced with “The bias patterns are similar in both models w@h1 and w@h2”.

- p.7 l.22-31. You should mention that by prescribing SST you pretty much fix the trends over land as well (eg Shin et al, Clim.Dyn. 2011 and other papers from Sardeshmukh’s group). The agreement is therefore not all that surprising.
We have added the following sentence, albeit with a cautious formulation due to our new findings of Section 4.4, which show that local to regional trends exhibit large variability depending on the ensemble members and hence do not appear that strongly constrained by SSTs: “Although this may not be surprising since others have found that prescribing SSTs may strongly force trends over land (e.g., Shin and Sardeshmukh, 2011), we note that regional trends computed from various ensemble members suggest a large range of trends despite the prescription of SSTs (see Sect. 4.4).”

- p.7 l.22-31 Some formal analysis how many times the temperature falls outside the ensemble range seems called for, ie whether the ensemble is reliable: is the spread a good representation of variability? Note that this is not covered in section 4.4, as there the distributions are normalised to their own variability.
Thank you for this good suggestion. We have added to the anomaly time series (Figs. 5, S9 and S11) the fraction of years when the observation lies within the 5–95% confidence interval of the w@h2 ensemble. For the global time series these are 71% for temperature and 58% for precipitation. We have added the following sentences in Section 3.2: “For temperature ... CRU-TS mostly lies within the 90% confidence interval of the w@h2 ensemble (71% of the years, suggesting that variability at the global scale might be slightly underestimated)...

For precipitation... CRU-TS appears to lie more often outside the w@h2 ensemble for precipitation than for temperature (observed values are within the 5–95% range from w@h2 on only 58% of the years)...".

- p.7 l.32- The same holds for the regional time series.
  As mentioned above, this has also been added to the regional time series figures (Supplementary Figures S9 and S11).

- p.11 l.2 "and may be the subject of further work" is not useful information.
  Removed.

- p.11 l.14 Why did you not take a standard percentile for the shading, like the 95% CI, rather than the full range of 1000 bootstrap sample?
  We thank the referee for his suggestion. We have changed our plots to show the 95% confidence interval from the bootstrap samples. Besides the description of this in Sec. 4.3 and in the respective figure captions, no text was changed as the results are qualitatively the same.

- p.11 l.29 I am also not impressed by the cold extremes in France and the British Isles, especially with the non-linear behaviour there.
  Yes indeed. We have added the following sentence: "Extreme cold night in BI and FR, however, are also underestimated by the model (i.e., extreme cold night are not cold enough)".

- p.12 l.19 Can you make the connection between the "attribution of extreme weather events" and "seasonal temperature in the upper tercile" more explicit? What are the reasons to assume that if the model is reliable in the latter it is suitable for the former?
  Our reliability analysis focuses on seasonal averages, not on extreme weather events as such. However, both are related to some extent, as extreme weather events can have a significant impact on the seasonal average. In addition, if a specific set of forcings (greenhouse gases, SST pattern, ...) is conducive to higher temperature, it will lead to higher seasonal averages and likely also hotter heat waves. We have added the following sentence in the first paragraph of Section 4.4 to emphasize these points: "While seasonal averages are not directly related to extreme weather events, the drivers of both are likely similar (e.g., higher CO_2 leads to increased mean and extreme temperature), and the occurrence of a few extreme events may strongly impact the seasonal average".

- p.12 l.20 It is not clear to me whether these reliability diagrams are computed using all grid points in the region, as the Met Office group does, or using the area-averaged value for the region. Please clarify.
  The reliability diagrams use area-averaged values for the region. This was clarified by adding "regional area-averaged" in the following sentence of section 4.4:

"For each type of event (e.g., high summer temperature, defined as JJA averaged temperature in the upper tercile), the probability of the event is computed for each year from **regionally averaged** w@h2 model output ("forecast probability")".

- Fig.13 Please explain the difference between the red and green dots.
  There is no green dot on Figs. 13–16, so we assume this refers to the black dots. As explained in the figure caption, "bins containing less than five years shown in black". Red dots, on the other hand, are for bins with at least five years. This was clarified in the caption by inserting "(red dots indicate bins containing at least 5 years)". It should be mentioned that black dots were, therefore, mostly not considered in the description of the results as they do not correspond to robust values. We have added a sentence to make the reader aware of this in the main text: "Results for bins containing at least 5 data points (i.e., years) are shown in red, while for other bins, shown in black, values are not very robust and should be interpreted with caution".

- p.12 l.31 How does this assessment that the model performs well after calibration compare to publications that w@h1 and other RCMs are very poor at simulating trends in heat waves (Min et al, 2013; Sippel et al, 2016)?
  The reliability analysis is based on seasonal averages, while heat waves usually last a few days. We have not investigated trends in heat waves specifically in our analysis. Nonetheless, the two studies mentioned by the referee (Min et al., 2013; Sippel et al., 2016) point to an underestimation in heat wave trends by RCMs compared to observations. This is nicely consistent with the "underconfidence" that we find for hot summer: the model's sensitivity to greenhouse gas forcings may be too low. We therefore added the following sentence: "Interestingly, this underestimation of the sensitivity of hot temperatures to forcings is consistent with the tendency of RCM to underestimate trends in heat waves over Europe (Min et al., 2013; Sippel et al., 2016)"

- p.13 l.6 "For low summer precipitation (Fig. 15), the reliability is found to be rather good in IP, AL, EA, ME" I do not see that by eye. Please use a more objective criterion, such as the fit by Weisheimer and Palmer (2014).
  We thank the referee for this useful suggestion. We have implemented the fit and bootstrap sampling proposed by Weisheimer and Palmer (2014) and we now display their proposed categorisation on the upper left of each plot in Figures 13–16 (note that as more simulations have been completed since our initial submission, some of the figure have changed slightly). We have also added a table (Table 3) which summarises the five categories. For temperature, very good performance is found, with categories 4 and 5 in almost all cases. For precipitation, performance is much lower, with categories 1–3 being most prominent. We have substantially edited Section 4.4 to include the information provided by this met-

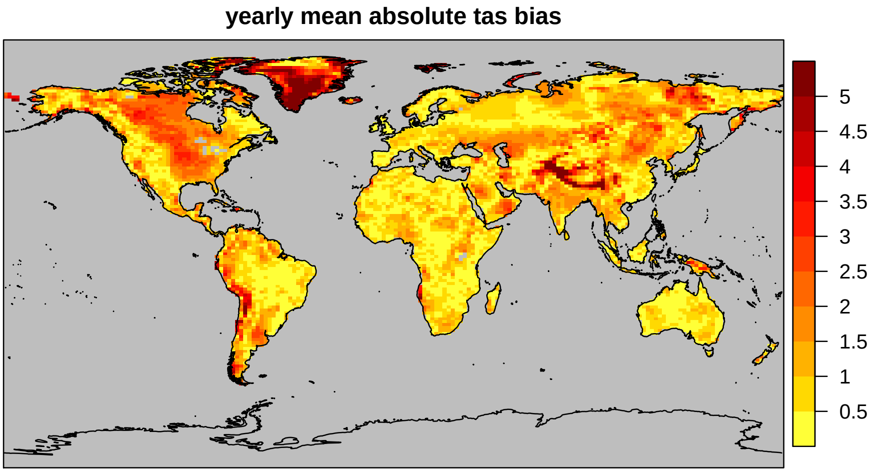

Figure R1: HadAM3P bias in yearly mean 2m-temperature with respect to CRU-TS (degrees C).

ric.

- p.13 l.15 “Therefore, these results may be dominated by the long-term trend arising from increased greenhouse gas concentrations”, This is fairly certain, as seasonal predictability in Europe is dominated by the trend.
  We have changed “may be” to “are”.

- p.14 l.30 “Overall, weather@home is an excellent tool for the investigation of extreme weather events.” should read “may be a useful tool if proper bias corrections and other caveats are taken into account”. As with every climate model.
  Done.

References

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W. J., Cox, P., Driouech, F., Emori, S., Eyring, V., et al.: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Climate Change 2013, 5, 741–866, 2013.

Min, E., Hazeleger, W., van Oldenborgh, G. J., and Sterl, A.: Evaluation of trends in high temperature extremes in north-western Europe in regional climate models, Environ Res Let, 8, 014011, URL `http://stacks.iop.org/1748-9326/8/i=1/a=014011`, 2013.

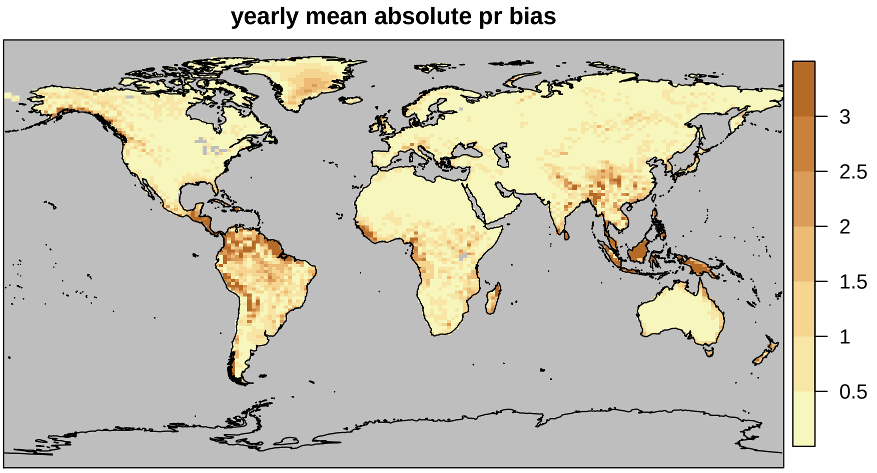

Figure R2: HadAM3P bias in yearly mean precipitation with respect to CRU-TS (mm/day).

Shin, S.-I. and Sardeshmukh, P. D.: Critical influence of the pattern of Tropical Ocean warming on remote climate trends, Clim Dyn, 36, 1577–1591, doi:10.1007/s00382-009-0732-3, URL `http://dx.doi.org/10.1007/s00382-009-0732-3`, 2011.

Sippel, S., Otto, F. E. L., Flach, M., and van Oldenborgh, G. J.: The role of anthropogenic warming in 2015 Central European heat waves [in Explaining extreme events of 2015 from a climate perspective], Bull Am Meteorol Soc, 97, 551–556, 2016.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11, doi:10.1098/rsif.2013.1162, URL `http://rsif.royalsocietypublishing.org/content/11/96/20131162`, 2014.

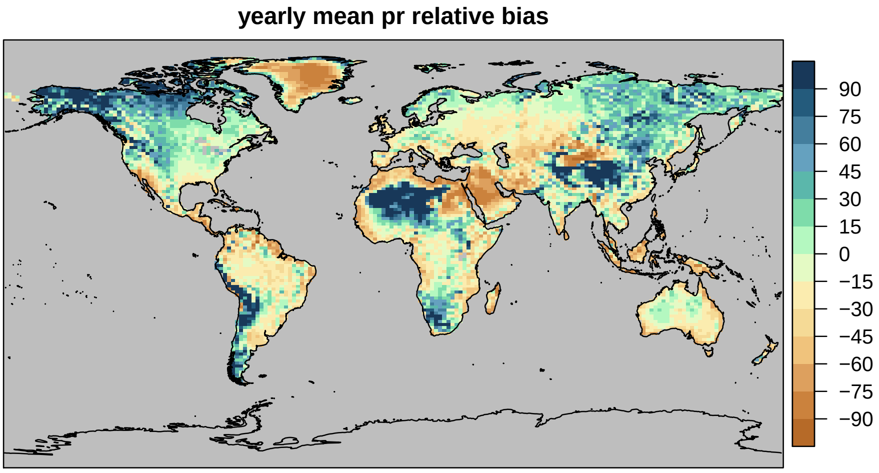

Figure R3: HadAM3P relative bias in yearly mean precipitation with respect to CRU-TS (%).

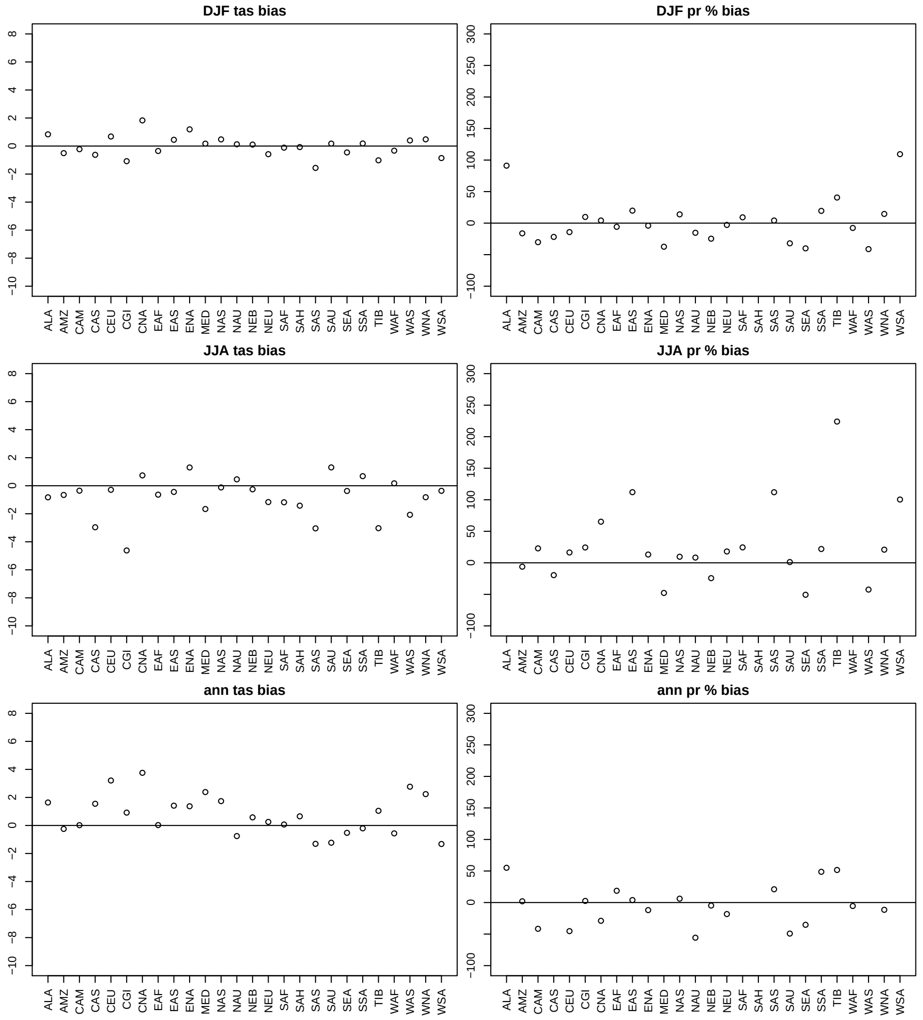

Figure R4: HadAM3P biases in (left) temperature and (right) precipitation in DJF (top), JJA (middle) and annually (bottom), for the SREX regions. As Fig. 9.39 of Chapter 9 of IPCC WG1 AR5.