Reply to the review by Anonymous Referee #1

We are very grateful to the anonymous referee #1 for the constructive and positive review. We include our answers to the comments in blue font right under the unmodified comments from the review. We note that the line numbers provided by reviewer #1 refer to the originally submitted document and do not correspond to the published discussion manuscripts. When referring to a particular line in our answer here, we provide line numbers for both documents to avoid confusion.

REVIEW

The manuscript describes the validation of the improved global-regional climate modeling system weather@home2. Reading this manuscript has been a real pleasure! The manuscript describes in a concise and very well written way the changes compared to the previos version of the model system and their impact on the global and regional climate over Europe. Figures illustrate the important results. I can recommend this manuscript for publication after the authors have addressed a few questions and comments.

We thank the referee for his positive and encouraging review.

Abstract and Conclusions: I wouldn't fully agree to the statement that European biases are reduced. It is certainly true for the temperature, but precipitation? Look at Fig.9, the w@h1 0.22 deg results are often better than w@h2! I suggest you differentiate between temperature and precipitation biases.

We agree that for precipitation the improvement in w@h2 is not clear, however we think the different resolution of the two models leads to some confusion: the w@h1 0.22 degree and the w@h2 0.44 degrees results are based on interpolated/aggregated data, as the models are not run at these resolutions. From Figure 9, root-mean square biases are reduced when precipitation is either interpolated or aggregated, which highlights an issue of precipitation location in both models. Hence, the apparent better performance of w@h1, 0.22 degree may be just an artefact. This was explained in section 4.1, from line 21 on page 9 until the end of section 4.1.

Table 2 in the paper lists the mean regional bias for both models in the sub-regions. w@h2 is better than w@h1 in many cases as well, but indeed there is overall not a clear improvement compared to w@h1. Therefore, we follow the referee's suggestion and change the following sentence in the abstract:

"The European RCM biases are overall reduced, in particular the warm and dry bias over eastern Europe, but large biases remain"

to

"The European RCM temperature biases are overall reduced, in particular the warm

bias over eastern Europe, but large biases remain. Precipitation is improved over the Alps in summer, with mixed changes in other regions and seasons."

And we have added the following sentence in the conclusion: "Precipitation biases in HadRM3P, on the other hand, do not exhibit substantial improvements overall".

Sec2.3: it is not clear to me how you create the initial conditions for each simulation. Are single-year spinup simulations part of the 13-mnth long experiment (i.e. making it 25 month long), or how exactly is it done? Please explain.

We agree with the referee that this needs clarification. Two separate sets of simulations were run. A first set of 12-months long simulations (December to November) have been run in a first experiment in order to create spun-up conditions (the initial state for the spin-up simulations comes from a long HadAM3P simulation with MOSES 1 and was reconfigured to MOSES 2). The end state from the spin-up simulations were then used to initialise the 13 months long experiment. We have clarified this in the first paragraph of Section 2.3, which now reads:

"A large ensemble of w@h2 consisting of more than 100 simulations per year from 1900–2006 is analysed. First, a restart file from a century-long HadAM3P simulation with MOSES 1 has been reconfigured for MOSES 2. This initial condition file is then used in a spin-up ensemble consisting of 12-month simulations (from December to November, with multiple simulations for each year), providing spun-up initial conditions on December 1st each year. The simulations analysed in this paper are then initialised on the 1st of December each year from the end state of the spin-up ensemble and are run for 13 months. (...)".

Sec2.3: How good is the initialization of soil and vegetation variables? Soil has a memory in excess of 1 year, so a 1-yr spin-up may not be sufficient for soil temperature and humidity. You have made a large effort to improve the land surface and vegetation components in your model, yet an inaccurate initialization could make these improvements worthless. Could you comment on that?

This is a very good point. The initial conditions used for the spin-up simulations are derived from a multi-decadal HadAM3P simulation. The land surface model in that simulation was MOSES 1, therefore the soil initial conditions are spun-up to that model. As the referee correctly points out, one year is however a rather short spin-up to the more recent land surface model MOSES 2, although one might expect these to be not too different. Unfortunately soil temperature was not saved as an output in these simulations, but we have looked at soil moisture and could find a small spin-up effect from our simulation output.

Fig. R1 and R2 display the difference in soil moisture between end of year 2 and end of year 1 (monthly average in December, which are months 13 and 25 from the restart with MOSES 1 conditions) for the 4 soil layers, scaled by the standard deviation of soil moisture at the end of year 2 (i.e., month 25). In some regions, large changes are

found (GCM: North Africa in all layers, and Asia/Western North America in the deepest layer, RCM: mostly only the deepest layer in Europe). This suggests that the soil has partly, but perhaps not fully equilibrated with the model. Fortunately, the upper 1m of the soil, corresponding to the root zone in most regions and therefore most critical for evapotranspiration, appears relatively well spun-up over Europe. Unfortunately, it is not possible to assess whether an additional year would lead to further changes, as these are not available.

To nonetheless test the spin-up effect on our analysis, we display the biases in temperature and precipitation in HadAM3P and HadRM3P for both years in Figs. R3–R6. The largest impact is found in DJF but is unlikely due to soil moisture as it spans all latitudes. For temperature, the most striking difference is an reduction of the bias over Southeast Europe, which may be driven by increased soil moisture in this region and possibly by effects of soil temperature. This suggests that a longer spin-up might potentially further reduce this model bias and thus that the spin-up may not be sufficient. For precipitation, the impact is small globally, in all seasons except DJF and, in other seasons, in Sahara, where % biases are very sensitive to small changes. DJF impacts are found throughout latitudes and are thus unlikely to be a soil moisture spin-up issue but may results from changes in circulation induced by temperature changes. In HadRM3P, similar results are found, with mostly an impact in DJF unlikely related to soil moisture.

These results highlight that a longer spin-up may be required in future uses of w@h2. In light of these results, we plan to update the w@h2 experimental setup to use spun-up conditions from longer simulations.

We have therefore included these figures in the Supplementary Information (Supplementary Figs. S5–S7 and S15–S17), and have added the following comments in the main text of the paper:

- Section 2.3: "The effect of the relatively short spin-up for soil variables on simulated temperature and precipitation is discussed in Sect. 3.1 for HadAM3P and Sect. 4.1 for HadRM3P".
- Section 3.1: "Finally, to assess whether the 1-year spin-up was sufficient to allow the soil variables to be spun-up, Supplementary Fig. S5 shows the difference between ensemble mean soil moisture (for each soil layer) in December between the 1st month and the 13th month of the analysed simulations (i.e., 13th and 25th month of simulation, respectively), scaled by the standard deviation of the second one. Apart from North Africa, the differences are confined to the 3rd (Central Asia) and 4th layer (many regions). This suggests that a longer spin-up may required in future experiments with w@h2. Fortunately, however, the upper 1m of the soil, corresponding to the root zone in most regions and therefore most critical for evapotranspiration, appears relatively well spun-up over Europe. It is not possible to assess whether an additional year would lead to further changes, as these are not available, and soil temperature is not examined here as this vari-

able has not been saved in our simulations. The impact on temperature biases is shown in Supplementary Fig. S6 and the largest impact is found in DJF but is unlikely due to soil moisture as it spans all latitudes. The most striking difference is a reduction of the bias over Southeast Europe and Central US, which may be driven by increased soil moisture in these regions with soil moisture-limited evapotranspiration regimes (Seneviratne et al., 2010) and possibly by effects of soil temperature. An impact is also found in MAM. This suggests that a longer spin-up might potentially further reduce the summer temperature warm model bias. For precipitation (Supplementary Fig. S7), the impact is small globally, in all seasons except DJF and, in other seasons, over Sahara (note that % biases are very sensitive to small changes in this region). DJF impacts are found throughout latitudes and are thus unlikely to be a soil moisture spin-up issue but may result from changes in circulation induced by temperature changes. These results highlight that a longer spin-up may be required in future uses of w@h2, which will be implemented for future w@h2 experiments."

- Section 4.1: "Finally, the impact of the short spin-up is evaluated as was done in Sect. 3.1 for HadAM3P. Fig. S15 shows the difference in soil moisture as in Fig. S5 (see Sect. 3.1). Over Europe, only Finland and Northwestern Russia display large differences in the upper 1 m of the soil. In the deepest layer, soil moisture is larger in the analyzed year than in the previous year over Southeastern Europe and in some other regions, but this deep layer is less critical to evapotranspiration and therefore to surface climate. Analysis of temperature and precipitation biases (Figs. S16 and S17) show that the hot MAM and JJA biases over Southeastern Europe are reduced with progressing spin-up, as expected from the increasing soil moisture and suggesting that a longer spin-up may further reduce this bias. Temperature biases in DJF and precipitation biases in all seasons are not related to soil moisture changes in a straightforward manner, and hence could be due to soil temperature, a variable not saved as an output in our simulations and therefore not analysed here.".
- Conclusion (Section 5): "A limitation of w@h2 as presented in this study is the relatively short spin-up (1 year). We find that a longer spin-up may further improve w@h2, in particular with respect to the representation of summer temperatures over Southeastern Europe. Future w@h2 experiments will therefore include a longer spin-up of 5–10 years, in order to allow for a full stabilization of soil moisture and soil temperature and to thereby take full advantage of the capability of the model."

Sec4.4: To be honest, I was somewhat surprised to see a section about reliability in this manuscript. Reliability is a very specific term with a precise definition in the verification of probabilistic forecasts, but I have never encountered it in the context

of climate simulations. On the other hand, the reliability of climate models is often discussed (e.g. in the IPCC AR) in the casual meaning of reliability as a synonym to trustworthiness. In this second definition of reliability, one often looks at how well the pdf of a quantity from a climate model matches the observed distribution. I wonder if this latter approach was what you had in mind when you started discussing reliability. Reliability and attribution diagrams as you present them now don't make much sense in the context of climate simulations, they should only be used for the verification of probabilistic forecasts. I therefore suggest you remove section 4.4 completely.

While we agree that it is slightly unusual to use reliability diagrams in this context, we do not agree that they don't make sense. We believe that Section 4.4 provides a useful quantification of the ability of weather@home2 to realistically simulate the response of climate to its drivers, which is very relevant for attribution. Given this and since Referee #2 showed interest in the reliability results, we have decided to keep this section. In addressing the comments from referee #2, we have complemented it with an new figure showing regional trends to support the interpretation of reliability diagrams, and have added some quantitative statements on model's reliability based on Weisheimer and Palmer (2014).



Figure R1: Soil moisture spin-up in HadAM3P. Difference between ensemble mean soil moisture in December between the end of the 1st year ("spin-up", 13th month from the generic restart) and the end of the 2nd year (25th months from the generic restart) in each simulation, normalized by the standard deviation (taken from the end of the 2nd year). Years 1961–1990 were used.

References

- Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture-climate interactions in a changing climate: A review, Earth-Sci. Rev., 99, 125–161, doi:10.1016/j.earscirev.2010.02.004, URL http://linkinghub.elsevier.com/retrieve/pii/S0012825210000139, 2010.
- 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 R2: Soil moisture spin-up in HadRM3P: same as Fig. R1 but for HadRM3P.



Figure R3: Spin-up effect on HadAM3P temperature biases: Biases in surface air temperature for HadAM3P in the spin-up run (w@h2spinup, left; a,d,g,j) and the 2nd year (w@h2, middle; b,e,h,k), and the difference in absolute biases (right; c,f,i,l, expressed as w@h2 minus w@h2spinup, i.e., negative values indicate an improvement with ongoing spin-up). Each row corresponds to a season (from top to bottom: DJF, MAM, JJA, SON).



Figure R4: Same as Fig. R3 but for precipitation, in %.



Figure R5: Spin-up effect on HadRM3P temperature biases: Same as Fig. R3 but for HadRM3P.



Figure R6: Same as Fig. R5 but for precipitation, in %.