Reviewer #2:

This study examines simulations of a climate indicator over Europe with implications for human health (heat stress index, Wet Bulb Globe Temperature (WGBT)). Bias corrected simulations from both Global and Regional Climate Models (GCMs and RCMs) are compared with the goal of determining the added value provided by the RCM in this scenario as well as more complex BC methods (QM vs ISIMIP). One novel aspect of this study in particular is the fact that the WBGT is multi-variate as it is based on both temperature and dew point temperature, which adds considerable complexity in the context of assessing the value of bias correction methods due to intervariable relationships. Overall, the manuscript is very clear, concise, and provides some evidence to support its conclusions, in particular that the chosen RCMs added little value with respect to the GCM after bias correction. The authors have properly acknowledged some major caveats to this conclusion, including (1) Only 1 GCM was used in the comparison between RCMs and (2) Regridding the high-resolution RCM simulations to the much coarser reference dataset may reduce any potential added value they would have otherwise provided. These open up several avenues for future work.

Response: We thank the reviewer for the comments and the time devoted to our paper. Please, see below our point-by-point responses and the changes highlighted in the new version of the manuscript.

Specific Comments:
- Page 5, Line 31: Given the issues you had to account for due to the 360-day calendar in HadGEM-ES, why did you select it for this study over other CMIP5 GCMs which have more standard calendars?

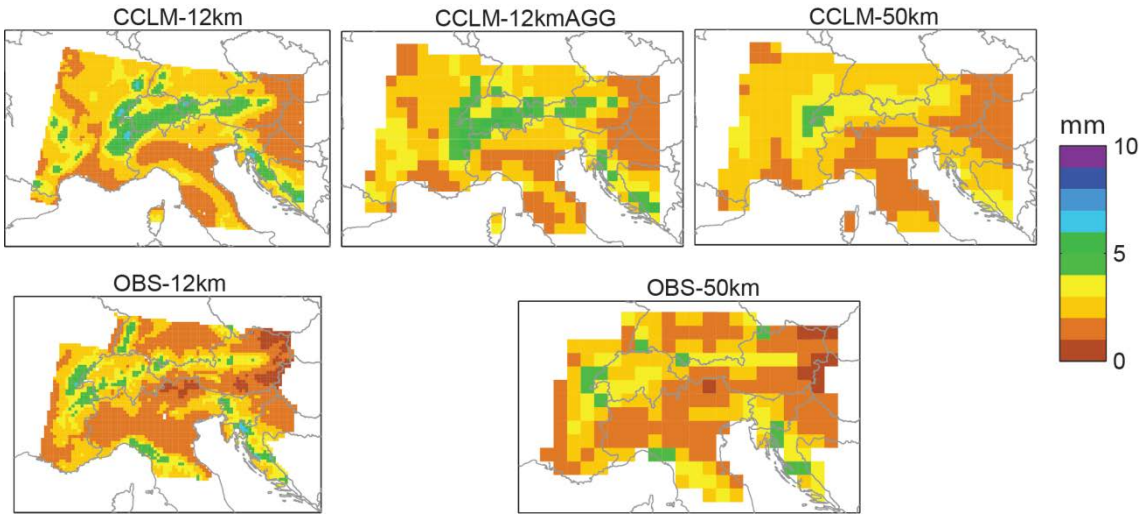
The motivation of the present work started after Kjellstrom et al. 2017, who estimated population heat exposure and impacts on working people at a global scale with GCM data from the ISIMIP project. Only two out of the 4-5 ISIMIP-corrected GCMs were used in the cited work, as representative of the range of different models used by IPCC for global temperature change (GTC) estimates; HadGEM2 producing GTC results close to the upper limit of models and GFDL producing results close to the lower limit. Thus our aim was to assess the robustness of those results based on GCM data and the ISIMIP correction. Unfortunately, GFDL was only dynamically downscaled through the RCA4 (SMHI) regional climate model within EURO-CORDEX, whereas HadGEM2 provided the boundary and initial conditions for three RCMs (at two spatial resolutions and for three scenarios, see Table 1), therefore only the latter was considered. In general, however, we believe that the filling in of a few missing values in order to account for the full Gregorian calendar does not distort the results obtained.

Kjellstrom, T., Freyberg, C., Lemke, B., Otto, M. and Briggs, D.: Estimating population heat exposure and impacts on working people in conjunction with climate change. International Journal of Biometeorology, 01 3, 62, pp. 291-306, 2018.

- Page 6, Lines 4-6: Could you also be more specific in regards to what beneficial features aren't smoothed out from the high-resolution simulations after regridding?

Thanks for the comment. That part of the sentence refers to the aspects mentioned before. We mean that the added value of the high resolution on certain processes might still be evident after regridding/smoothing. The sentence has been rewritten to:
"As a consequence, there will be aspects of the added value of the high-resolution EUR-11 experiments (related to better-resolved, fine-scale processes; Prein, et al., 2015) that can be smoothed out, but **they** may still be present after remapping them onto a coarse resolution (Casanueva et al. 2016)."

As an example, the following figure shows daily mean precipitation (period 1989-2008) for the Alpine region as represented by a single RCM (CCLM) at the 12km original resolution, regridded onto the 50km (12kmAGG) and the original simulation at 50km. Compared to the latter, the aggregated 12km version shows more details that are also present in the full 12km version. See more details of this added value analysis in Casanueva et al. 2016.

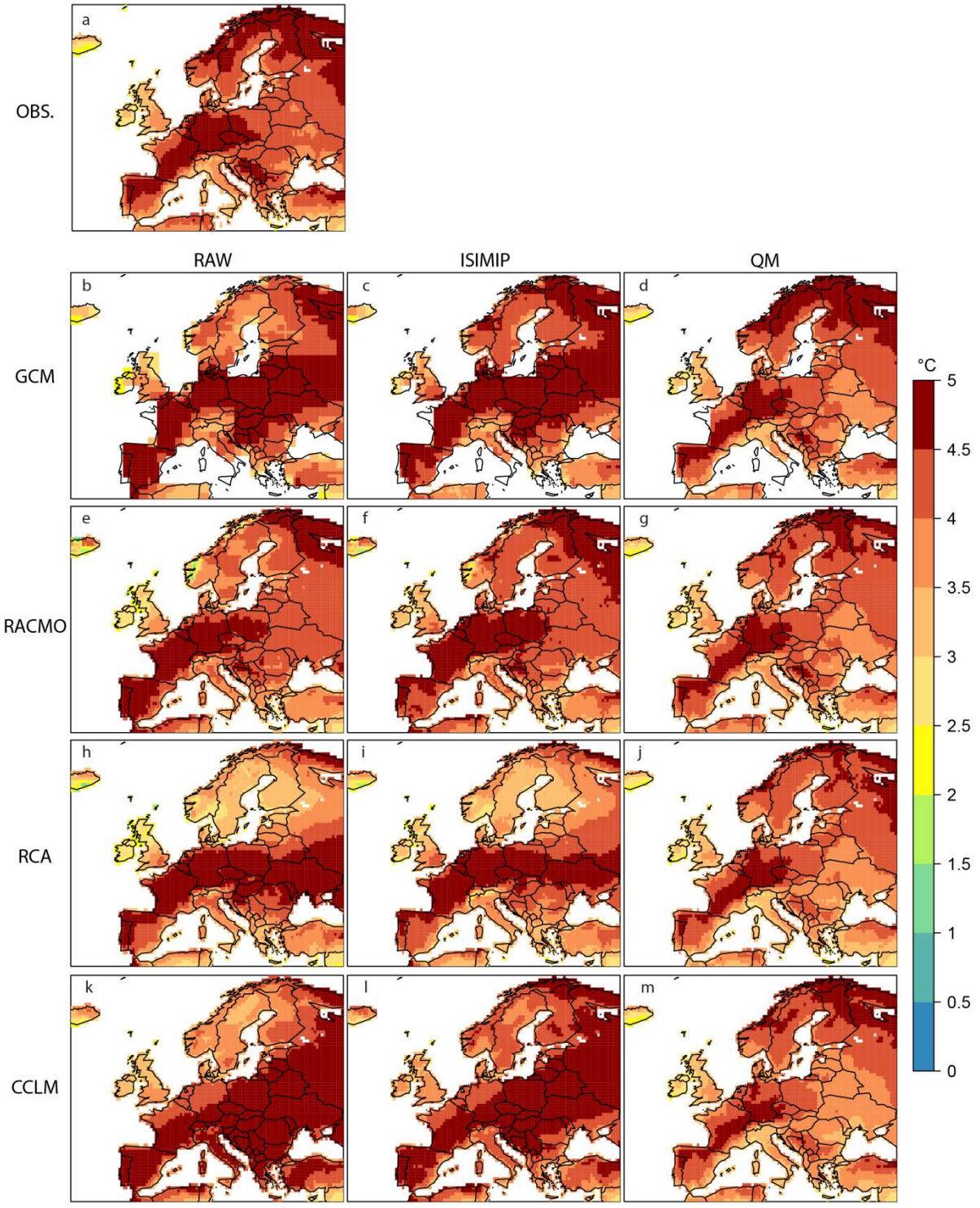


Casanueva, A., Kotlarski, S., Herrera, S., Fernández, J., Gutiérrez, J. M., Boberg, F., Colette, A., Christensen, O. B., Goergen, K., Jacob, D., Keuler, K., Nikulin, G., Teichmann, C. and Vautard, R.: Daily precipitation statistics in a EURO-CORDEX RCM ensemble: added value of raw and bias-corrected high-resolution simulations. Climate Dynamics, 47, pp. 719-737, 2016.

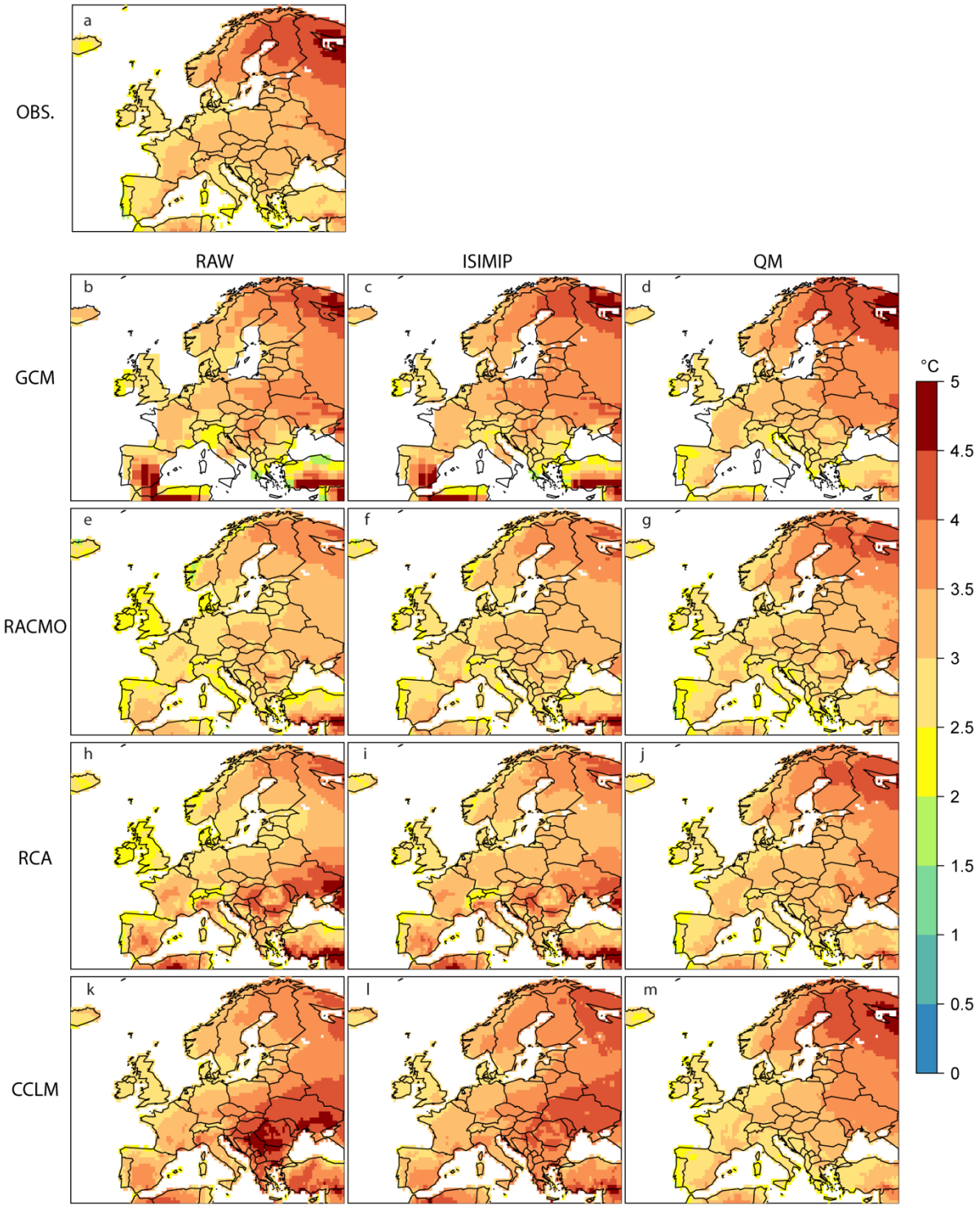
- Page 11, Lines 20-25: Some interpretations which explain these results would be nice to have here, in particular to explain the lower skill in Scandinavia for the RCMs. It might be helpful to see some additional maps showing the standard deviations of daily max temperature and daily mean dewpoint temperature.

Thanks for the comment. The two figures below show the standard deviation of the two variables. The areas with larger biases in the standard deviation agree with those with smaller Perkins scores in Fig.6, pointing out that deficiencies in the temporal variability of the individual variables might be responsible for some of the deficiencies in the intervariable relationships. It is also noticeable that the lower skill in Scandinavia after QM corresponds to lower variability (standard deviation) than observed in the two variables.
These two plots have been added to the Supplementary Material and the text has been completed as follows:

"High Perkins scores are found especially along the Atlantic coast. QM improves on ISIMIP in large areas, although low scores are found in Scandinavia (0.7-0.8) for the RCMs. The spatial distribution of the scores agrees qualitatively with biases in the temporal variability of maximum and dew point temperatures (Figs.S4-S5). This is a first order indication that the misrepresentation of the temporal variability of the individual variables might be responsible for most of the deficiencies in the intervariable relationships. Raw model data overestimate the temporal variability especially in Eastern Europe, leading to Perkins scores lower than 0.6. In other areas, such as Scandinavia, the models underestimate the temporal variability of the two input variables, and thus present the lowest scores even after QM."



Standard deviation of the distribution of daily maximum temperatures (1981-2010, JJA). Results for the observations (a), GCM (b-d) and three RCMs-EUR11 (e-m). Raw and bias-corrected data are depicted in columns.



As the previous figure, but for daily mean dew point temperature.

- Page 14, Lines 27-28: This would be a bit beyond the scope of this paper, but given that the RCMs chosen in this study are still coarse enough to require many parameterizations, I would be interested in seeing future work examine the robustness of this conclusion for convection permitting models.

Thanks for the comment. That is certainly an interesting point for future work. A sentence on this has been included in the discussion:
"Future works including convection-permitting simulations could help to assess the robustness of these results".