Reviewer #1:

The manuscript investigates the role of downscaling and bias correction to capture the climate change signal of multi-variate heat stress index, by comparing GCM and RCM simulations at different spatial resolutions. The corrected heat stress index (WBGT in the shade conditions) is calculated from air temperature and dew point temperature, which were separately corrected using two BC methods; a) ISIMIP (parametric quantile mapping) and b) empirical quantile mapping. The bias-correction methods applied in the manuscript are not newly developed techniques. However, the application on a multi-variate index and the evaluation of the corrected index are a needed task in the topic of bias-correction on climate model simulations. The overall manuscript is well written, and most of the figures included are clearly stated.

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:

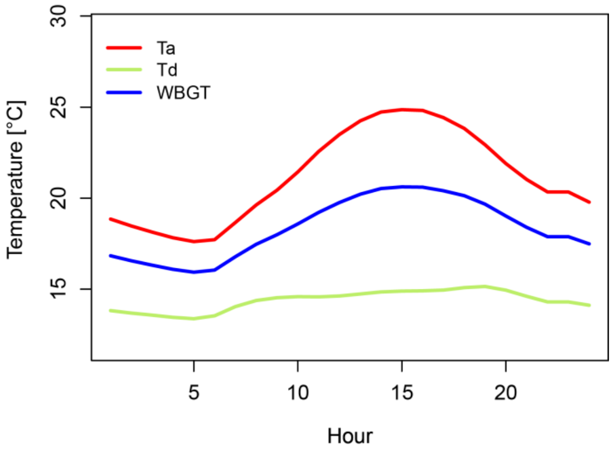
- Page 3, line 19: More explanation on "intensity-dependent biases" would help of the quantile mapping. Can you provide a reference for the term?

Thanks for the comment. A brief explanation and a reference have been included in the revised manuscript.

"Quantile mapping is, by construction, able to correct for intensity-dependent biases (i.e. biases that change throughout the distribution, Gobiet et al. 2015)"

- Page 5, line 3: I am curious about the reasoning of using daily 'mean' dew point temperature, instead of using daily maximum dew point temperature, to calculate the daily maximum WBGT.

Thanks for raising this point. The reason for approximating daily maximum heat stress using daily mean dew point temperature is the non-availability of data at hourly resolution in most observation-based and simulated datasets used in the current work (ideally, data at hourly or even higher temporal resolution should be used). Unlike relative humidity, which is anticorrelated with air temperature and changes strongly along the day, dew point temperature only slightly varies during the day. The following figure shows the diurnal cycle of air temperature (Ta), dew point temperature (Td) and WBGT for a typical summer day in Lugano (Switzerland), in the period 1981-2010. A similar result was found for other Swiss locations, where hourly data were available. In those cases dew point temperature shows a diurnal range of approximately 1-1.5°C.



Given the small diurnal cycle, daily mean values of dew point temperature in combination with daily maximum air temperature were used in the present work to approximate daily maximum heat stress. Also note that daily mean dew point temperature was obtained from daily mean temperature and relative humidity in models and observations (neither hourly nor minimum relative humidity data were available from either models or observations).

- Page 9, line 30: I like joint distributions of two input variables in Fig 5 to understand the characteristics of joint dependency for climate simulations better. However, it would be good to see some statistics like the correlation to show dependence between two input variables, maximum temperature and dew point temperature. In Fig 4d, it seems there exists a stronger negative correlation between two variables in the raw CCLM, compared to the correlation in Obs. If the negative relationship is stronger on extremes (e.g., above 95th percentile) of two variables, that might bring inaccurate bias adjustment in QM, leading to the underestimated negative biases?

Thanks for the comment. The following table summarizes the (Pearson) correlation coefficients between daily maximum temperature and daily mean dew point temperature, considering the full series (left) and the pairs of values that produce WBGT above the 95th percentile:

	Full series (JJA)			**Pairs producing WBGT>WBGTp95**		
OBS.	**0.54**			**-0.55**		
	RAW	ISIMIP	EQM	RAW	ISIMIP	EQM
GCM	0.16	0.2	0.16	-0.71	-0.61	-0.71
RACMO-044	0.49	0.52	0.49	-0.72	-0.65	-0.71
RACMO-011	**0.42**	**0.46**	**0.42**	**-0.43**	**-0.53**	**-0.45**
RCA-044	0.23	0.31	0.3	-0.74	-0.71	-0.8
RCA-011	0.27	0.35	0.34	-0.67	-0.62	-0.7
CCLM-044	0.08	0.16	0.14	-0.85	-0.81	-0.88
CCLM-011	**0.02**	**0.09**	**0.06**	**-0.82**	**-0.83**	**-0.88**

In general air and dew point temperatures present a positive, linear relation (r=0.54). However, extreme values of WBGT are produced (at this specific grid box close to Warsaw) under high values of air temperature and low values of dew point temperature, or vice versa (r=-0.55). RACMO stands out as the best of the three RCMs representing the intervariable relationships, for the full series (see also Fig.5, third row) and the highest WBGT (see also Fig.4a-c). However, for the GCM and CCLM the two full series do not correlate linearly (r is approximately 0) and they are too strongly anticorrelated for the extreme WBGT (see also Fig.5 last row and Fig.4d-f for CCLM).

The two bias correction methods do not tackle the temporal correlation and maintain the temporal structure of the raw data and the temporal correlation between air and dew point temperatures remain similar to the raw counterpart. A slightly stronger negative correlation for the pairs producing extreme WBGT is obtained for the CCLM simulations after QM. That means that high values of dew point temperature would then be linked to rather low air temperatures (stronger negative correlation than for the observations), which may imply lower values for extreme WBGT. This together with an overcorrection of the positive bias in extreme air temperatures (Fig.4d,f) might favour negative biases WBGTp99.

Some further explanations and the correlation coefficients from the table above have been included in Figs.3 (for pairs of variables producing WBGT>WBGTp95) and 4 (all pairs) in the revised manuscript.

- Page 11, line 24-25: I don't know how the conclusion is drawn. By comparing average Perkins scores?

This conclusion is drawn from Fig.6, where the spatial distribution of Perkins scores is depicted. In particular, the best results for GCM-QM are found in Fig.6c, with values close to 1 for all Europe. It is explicitly mentioned in the revised manuscript that this conclusion is shown in the mentioned plot.

- Page 15, line 6: If I understand correctly, you used a single ensemble (r1i1p1) of HadGEM2-ES. Do the biases relate to the biases across ensemble runs? If we use more ensemble members of the HadGEM2 simulation, do we expect the smaller biases?

All results are based on a single ensemble r1i1p1 of HadGEM-ES. When mentioning the need of large ensembles of simulations, we refer to ensembles built on different GCMs. A larger ensemble of GCM-RCMs (as in Casanueva et al. 2018 for climate projections of heat stress), can ease the quantification of the robustness and uncertainties in the projections. Using large ensembles, considering also other HadGEM2 runs, does not necessarily mean a reduction of model biases, but their own biases.

- Fig 1a: I am a bit confused. Are the CDFs of the (historical and future) RAW from RCM? Or GCM?

Thanks for the comment. This figure attempts to illustrate the generic bias correction procedure for any (regional or global) model. The numbers correspond to HadGEM-ES (i.e. GCM). The caption has been changed for the sake of clarity.