General comments

Comment 1: Lin et al. used dynamical downscaling to analyse heatwaves based on simulations carried out with regional climate models (RCMs) from the Euro-CORDEX programme. A particularly relevant topic of this paper is that the authors investigated if there is any added value in the representation of heat waves in the RCMs compared to the driving GCMs. It is an interesting topic definitely worth pursuing.

Reply: Thanks!

Comment 2: A general remark is that all researchers discussing evaluation and use of GCM results on regional scales ought to read the paper by Deser et al (2012; DOI:10.1038/nclimate1562), and citing it in a study like this thus should be required. The findings of Deser et al. suggest that the small number of GCMs selected here is insufficient for a proper analysis of future outlooks and model evaluation, due to pronounced chaotic regional variability on decadal scales.

Reply: We agree that a large number of models is required for a proper analysis of future outlooks and model evaluation. It is worth mentioning that we are not trying to make any full-scale assessments of future heat waves in Europe but rather investigate how they differ between GCMs and RCMs by comparing the GCM-RCM chains. Thus, the problem is slightly less problematic. A problem from the downscaling side, though, is the lack of large ensembles of GCM-RCM chains that could be assessed. We conducted the same analyses based on another GCM-RCM combination (with the same size but different GCM/RCM members) and derived similar results, which can alternatively support our conclusions.

In the revised version, we added the following text in an additional subsection of Discussion (L363–368):

> *A full simulation matrix without gaps facilitates a fair comparison after aggregating along either the GCM or the RCM dimension. This requisite limits the size of RCM simulation matrix available for analysis. The limited size of the GCM-RCM simulation matrix could be a shortcoming influencing the robustness of the results in the study, especially for the uncertainty analysis where the uncertainty is described by the spread (maximum − minimum) across three or four ensemble members. In fact, we conducted the same analyses upon another GCM-RCM simulation matrix (with the same size but different GCM/RCM members) and derived similar results (not shown), which can alternatively support the conclusions herein.*

Comment 3: The regional climate modelling community also still seems to exhibit a 'silo thinking' behaviour, and in order to try to make som progress in the general thinking about downscaling, I would urge that this paper by Lin et al. also includes work based on empirical-statistical downscaling (ESD). Many papers on RCMs ignore ESD, which becomes invisible and under-appreciated, and this unfortunately seems to create an attitude that RCMs suffice - hence many of the climate services in Europe do not consider ESD. I suspect most people working with RCMs don't read the literature on ESD, but I think there are benefits from consolidating the two approaches - in particular when it comes to the evaluation of RCMs. There are also a few examples of ESD applied to heatwave statistics that merit a mention in the context of this paper (e.g. DOI:10.5194/ascmo-4-37-2018). Nevertheless, ignoring ESD is a weakness, although Lin et al. give a good summary of the limitations of RCMs. RCMs and ESD make use of different sets of assumptions and have different strengths and weaknesses independent of each other, and hence a combination of the two makes the results more robust.

Reply: We appreciate that the reviewer thoughtfully consider the different methods for downscaling, of which the ESD is worth and should be assessed in the context of assessments of climate change. However, this study is not focused on different downscaling approaches but rather the behaviors (e.g., signal modification and uncertainty transformation) within the GCM-RCM chains (see also our reply to Comment 2). Thus, we decided not to extend our analysis by including datasets based on ESD. Having said this, we did add the following statement concerning the use of both ESD and dynamical downscaling for deriving regional climate information (L40–44):

> *As a remedy, to improve the quality of the simulated climate and add value compared to GCMs, empirical statistical downscaling employing a wide range of approaches (Benestad et al., 2008, 2018; Hertig et al., 2019; Soares et al., 2019) and dynamical downscaling using a regional climate model (RCM) (Torma et al., 2015; Rummukainen, 2016; Strandberg and Lind, 2021) are used, each with its relative strengths. This study focuses on the dynamical downscaling.*

Comment 4: Often the most severe effects of heatwaves are connected with night-time temperatures not cooling off. It is therefore also of interest to use a heatwave index based on daily minimum temperatures and not the daily maximum. The most

pronounced temperature trends also are those of the nights.

Reply: This is an interesting suggestion and is worthwhile pursuing; however, it is out of scope of this paper. To make this point more visible, we added the following text in the revised Discussion section (L369–374):

> *The HWMId can also be applied to other temperature variables, but with different processes/impacts involved. For example, the HWMId applied to daily minimum temperature serve a measurement of heat wave magnitude taking into account also the nighttime cooling effect (Russo et al., 2015). As another example, Apparent Heat Wave Index (AHWI, Russo et al., 2017), the HWMId applied to daily apparent temperature, considers also the impact of air humidity on human beings. Such variants of HWMId are being considered for future studies.*

Comment 5: It would be interesting to see the statistical distribution of yearly HWMId values - are they normally distributed? (E.g. is the central limit theorem valid for this statistic aggregated over Europe?) One way to evaluate the models is to compare their statistical distributions (e.g. Kolmogorov-Smirnov Test).

Reply: This is an interesting comment, which was also mentioned by Reviewer#3. We are also grateful for providing the method to test statistical distribution. Below you will find my answers to your questions:

a) Are yearly HWMId values normally distributed? The answer is no. As reflected in some figures in the manuscript (Fig. 1b and Fig. 9 for spatial; Fig. 2b and Fig. 10 for temporal), the distribution is somehow skewed.

b) Is the central limit theorem valid for this statistic aggregated over Europe? The answer is yes. In probability theory, the central limit theorem (CLT) establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution **even if the original variables themselves are not normally distributed**.

Comment 6: I was a bit surprised by Fig.1 that seems to indicate more heatwave activity in the Nordic countries and less further south on the continent. This also seems to be the case for EOBS and ERAINT - does that mean that perhaps HWMId doesn't represent the typical heatwave reported by the news headlines? It's defined in terms of local variability (IQR) and autocorrelation - and not on any threshold value, as far as I read this paper. At least, this warrants some comments.

Reply: We agree that the HWMId index is somehow complex and assisting materials are needed to fully understand what it represents. The answer to "does HWMId represent the typical heatwave reported by the news headlines?" is yes, for which please refer to Russo et al. (2015, Fig. 2 therein). Moreover, we have added the following text (and associated figures) to the supplementary materials for better understanding the HWMId:

> *The HWMId (Russo et al., 2015) used in the study is a fairly established indicator for classifying heat waves, taking into account both the duration and intensity. The detailed definition of HWMId can be found in Sect. 2.1, as well as Russo et al. (2015). It is worth exploring how it works by presenting examples though. Figures S1 and S2 show the spatial distribution of observed (E-OBS) European HWMId values and the detected heat waves at four selected grid points for Years 2007 and 1989, respectively. Following the definition, heat waves are detected when daily maximum temperature exceeds the daily threshold at least three consecutive days, where the threshold for a given day is 90th percentile of daily maximum temperature within the 31-day window centered at this day and within the reference period (1989–2008 for the examples shown here, same as the results in Sect. 3.1). The HWMId value at each grid point is equal to the maximum red area above the climatological 25th percentile of maximum temperature in Fig. S1b (or Fig. S2b) normalized by the climatological interquartile range (IQR) within the reference period, where the 25th percentile and the IQR are constant at a given grid point. As such, the following points are noted:*
>
> *A) The yearly HWMId values shown on the map may not necessarily represent the spatial distribution of magnitudes of a single heat wave (e.g., occurred in the same period) but the maximum magnitudes of the grid points respectively. For example, Fig. S1b (or Fig. S2b) shows that the heat wave with maximum red areas appeared at different time of the year for the four selected grid points.*
>
> *B) The yearly HWMId values are in most cases associated with summertime heat waves due to the higher temperature (Fig. S1b and S2b). However, it does not rule out the case that the heat waves with the maximum magnitudes appear in wintertime; an example is given at grid point "SE" in 1989 (Fig. S2b).*
>
> *C) The HWMId can certainly identify those outrageous heat waves reported by the news headlines according to Russo et al. (2015, Fig. 2 therein). Here, the heat wave occurred over southern Europe in 2017 summer*

is clearly visualized in Fig. S1a. However, as an index that is reference period-based, the HWMId might be problematical when quantifying magnitudes of moderate heat waves within the reference period, because a higher threshold can be expected for a location with more heat wave activities. Figure S3 shows that the grid point "SE" compared to points in western Europe has a shorter distance between daily 90th and 95th percentiles (within a 31-day window) of daily maximum temperature within 1989–2008, where the former was used as the threshold for the HWMId within Sect. 3.1, over the long summer season (May–September), indicating very likely less heat wave activities at southeastern Europe. Considering that the yearly HWMId values are in most cases associated with summertime heat waves (i.e., Point B), therefore, the result shown in Fig. S1a explains the spatial pattern of HWMId with values in western coastal areas higher than those in eastern areas (Fig. 1a).

Comment 7: Does the result that all RCMs show less agreement with E-OBS in RMSE and r compared to that of ERA-Interim suggest that these RCMs don't add value to that of the global model? Or could it be differences in heat fluxes, cloudiness and topography of the driving and nested models? Perhaps the model domain is so large that the RCMs generate their own dynamics within the interior of their lateral boundaries? Or have they involved spectral nudging to avoid that? See e.g. DOI:10.1007/s00382-022-06219-y (it's also a useful paper to discuss in this context). These questions certainly merit some discussion. The results are nevertheless useful and interesting as they suggest that differences between the RCMs matter.

Reply: The issue of spectral nudging has been discussed vividly in the literature and it is indeed a powerful method for better capturing single events that are governed by the lateral boundary conditions. For the ERA-Interim-driven runs this could have been a meaningful thing to do. A problem in this context is that most RCMs are run without spectral nudging so there is no option of doing such an analysis for practical reasons. Also, as the GCMs do not necessarily represent these events in a realistic way it is interesting to analyse how a free-running RCMs modify the results.

Larger biases w.r.t E-OBS in the RCMs than in ERA-Interim imply that the RCMs do modify the results and that it infers its own biases. At the same time, however, it does not by default mean that it does not add value w.r.t. the GCMs. The better performance shown compared to the GCMs (Table 3 and 4) indicates that they do add value. Related to the size of the domain it is clearly so that the RCMs have some freedom to create their own climate, and that this freedom is generally larger in summer (when we see these heat waves) as the region is less well "ventilated" or "flowed through" by the westerlies that are weaker in these situations (especially, when there are strong high pressure situations - this also speaks against spectral nudging as a solution as it will still be "interior" processes that dominates).

Comment 8: I'm not sure that I understand Table 4 and the use of MBE, RMSE and correlation for results derived from GCMs since we don't expect the GCMs to be synonymous with the real world and hence no correlation with observed heatwaves. The only way to evaluate the downscaled results from GCMs is through statistical properties such as statistical distributions and parameters. But perhaps Table 4 shows the correlation in space rather than over time? If so, this ought to be explained more explicitly and clearly. Also if the appearance of the number of heatwaves more or less follows a random process, then we'd expect that it over a given period will follow a Poisson distribution - this can be assumed to be true for both models and the real world. Then the number of observed heatwaves can be compared to a statistical distribution of corresponding number of heatwaves based on the model ensemble by assuming a Poisson distribution (this works if the ensemble is considerably greater than 30 independent runs). Is this possible, or does the HWMId statistic suffice? Also, so-called 'common Empirical Orthogonal Functions' can be used to compare spatial structures and the covariance structures in different data sets - it's an elegant maths-based approach that is surprisingly uncommon. However, this is more general and not specific for a small selection of extreme events. But regarding my comment on Fig 1, I'm a bit unsure what HWMId really represents. Perhaps it also may be of relevance here to mention that one indicator of trends in extremes, including an increasing severity of heatwaves, can also involve an analysis of record-breaking events. There is some literature on this subject connected to climate change.

Reply: Statistics in Table 4 (as well as 2 and 5) are of spatial characteristics, which has been explicitly stated in the revised table caption. Sorry for the opaque manner in the previous version of table caption. We also thanks for the various approaches provided for evaluating the models. Regarding what the HWMId really represents, please see our response to Comment 6.

Comment 9: The most rapid warming in northern Europe is during winter, but maximum daily temperatures are highest in summer, and it's only summer that defines HWMId? (L348)

Reply: The HWMId analyzed of each year represents the maximum heat wave magnitude throughout the year. Please see our response to Comment 6 (especially Point B of supplementary text).

Comment 10: The point about 'cascade of uncertainty' is a myth and forgets that each step of analysis also introduces new information (or constraints) in addition to uncertainty. It's only sensible with several model stages as long as we introduce more information than uncertainty for each step (see e.g. DOI:10.1038/NCLIMATE3393). In fact, downscaling can be considered as an act of adding new information to that already provided by GCMs: information about how local geography influences the local climate (as in this case) and information about how local climates depend on the ambient large-scale conditions and teleconnections that the GCMs skillfully reproduce.

Reply: Thanks for providing the insights into RCMs' added value w.r.t. GCMs. Indeed, we have indicated in the manuscript that the concept of "cascade of uncertainty" is questionable by referring to other literature and to results of our study indicating that RCMs add information (and sometimes value). In the revised version, we integrated with more arguments including your insights and have the following text (L283–298):

> *Some studies (e.g., Schiermeier, 2010; Kerr, 2011) show that uncertainty may increase when downscaling GCMs with RCMs as biases from the GCMs are conveyed to the RCMs and RCMs additionally add their own biases, referred to as the 'cascade of uncertainty' (e.g., Wilby and Dessai, 2010). However, many other studies (e.g., Torma et al., 2015; Di Luca et al., 2016; Rummukainen, 2016; Sørland et al., 2018; Strandberg and Lind, 2021) indicate that RCMs can also add value upon the driving GCM simulations. This study demonstrates the added value for heat wave magnitudes that have so far not been studied as extensively as for other aspects of climate and climate change, reflected in adding more detailed geographical patterns and pulling the results closer to the observations (Fig. 3; Table 4 and 5). Such added value confirms the usefulness of RCMs for downscaling coarse-scale GCM simulations, because downscaling can be considered as an act of adding useful information to that already provided by GCMs: information about how local geography influences the local climate (as in this case) and information about how local climates depend on the ambient large-scale conditions and teleconnections that the GCMs skillfully reproduce. Moreover, RCMs may also more realistically represent some atmospheric processes relative to the GCMs. Our analysis of the ensemble spread along the GCM dimension, reflecting uncertainty associated with driving data, reveals that the RCMs alter the spatial HWMId pattern from their driving GCM simulations, and that the alteration is different between the RCMs (Fig. 5 and Table S3). This, on the other hand, suggests that the uncertainties of GCMs in simulating heat wave magnitudes would be transformed by RCMs in a complex manner, rather than simply inherited, due to the nonlinear nature of model dynamics and physics, thus rejecting the concept of 'cascade of uncertainty'.*

Comment 11: In summary, the tiny sample of GCMs in this study severely limits the application of these results and there were some points which were unclear and needed elaboration, as pointed out above. One way to improve this is to extend the ensemble of GCMs to the whole of CMIP5 (CMIP6?), and then compare those three selected here in this study with the larger set of GCMs. There are also some issues that merit more discussion, as mentioned above. I also think it's useful to discuss other definitions of heatwaves than HWMId, even if this paper focuses on just this fairly established indicator. Furthermore, it's important to consider ways to connect these results with what can be delivered by ESD (e.g. much larger ensembles than Euro-CORDEX), and in general I suggest that papers on downscaling that ignore one of these strategies do not merit publication.

Reply: In the revised version, we explicitly stated the focus of the study and why we need a full simulation matrix without gaps (L90–93), as follows:

> *The large number of GCM-RCM combinations available from EURO-CORDEX allows us to examine RCMs' behaviors (e.g., signal modification and uncertainty transformation) within the downscaling process for simulating heat wave magnitudes. For that we used only a subset of the available ensemble to gain a full GCM-RCM matrix without gaps, to ensure a fair comparison after aggregating along either the GCM dimension or the RCM dimension.*

Regarding the size issue of model members, the suggestion of use of ESD, and other definitions of heatwaves than HWMId, please see our response to Comments 2–4.

Details:

Comment 12: L52 "hace" is misspelt.

Reply: It is actually "have"; corrected.

Comment 13: Fig. 2 caption: 'Scott's rule' needs a reference.

Reply: Following the suggestion of Reviewer#2, we have removed the violin-plot and thus there is nothing to do with "Scott's rule" anymore.

Comment 14: L.188: Missing "there" in "shows a similar pattern to the ensemble mean (first row of Fig. 5) but exists considerable differences in the spread (second row Fig. 5) of the RCM ensembles"?

Reply: The sentence has been rephrased as:

> *Aggregating on the GCM dimension (i.e., calculating mean and spread for each RCM with different GCMs), the ensemble means (first row of Fig. 5) reveal a similar spatial pattern, whereas the spreads (second row of Fig. 5) show considerable differences in the spatial pattern.*