Answers to Reviewer #1 :

This article presents and discusses the performance of a large multi-physics ensemble configured with the WRF regional climate model system in representing two major heat wave episodes observed in recent years across the European continent. The paper attempts to objectively assess which configurations perform best in reproducing the heat wave characteristics with the purpose to identify a few best performing configurations which might be recommendable for application in studies on the role of heat wave in Europe under climate change conditions in summer.

The paper is reasonably well written, although I came across quite a few sentences which I found difficult to understand. In addition the manuscript also contains a number of assumptions or claims that go without solid argumentation (or argumentation at all) and are not backed up by references. Some of the figures are difficult to view, in particular the Figure 1 and the figures displaying scatter plots. But with some work that can easily be improved.

We thank the reviewer for the useful comments and suggestions. As concerns figures, they are now adapted to the suggestions of both reviewers.

A major concern, however, I have with the followed methodology is that the authors have chosen to leave out the land surface scheme from their considerations and to restrict their construction of a multi-physics ensemble to what they refer to as the atmospheric physical schemes. There are probably quite a few large-scale weather phenomena that are rather insensitive to the details of a land surface scheme, but for sure such approach does not hold for heat wave conditions across Europe. There is an overwhelmingly large amount of literature that points to the role of land surface processes and their effect on land-atmosphere exchanges in the weeks or months ahead of the onset of a heat wave episode across (a subregion in) Europe, so I really can't understand why the authors have chosen to pursue this approach. The more so as the authors themselves write in their concluding remarks." ... a limitation of this study is the use of only one land-surface scheme; the five selected WRF configurations may actually all compensate for systematic errors of the NOAH land surface scheme ..." etcetera.

This paper is designed as a methodological paper. Our intention was to test as many physic schemes as possible, including land-surface schemes. However, at the moment we performed the study 4 different land-surface schemes were available: NOAH, RUC, Pleim-Xiu and the 5-layer thermal diffusion scheme. The last one, non-physical, is designed for test cases and cannot be used in realistic situations. The Pleim-Xiu scheme comes with a dedicated set of other physical schemes and does not allow in most cases to combine different possibilities. We performed an ensemble of simulations using the RUC scheme, but we found that it provided extremely sensitive fluxes, with large latent heat fluxes at the beginning of season and extreme subsequent drying in summer months. Sensible heat fluxes also appear overestimated. Comparisons with several FLUXNET sites are now explicitly shown in Figure 1. So even if temperatures of heat waves would match observations in a few combinations of physics schemes, we would almost be sure that this would be for wrong reasons. We also experienced technical problems while running several of the RUC simulations. This made us decide to only use one land-surface scheme and focus on the atmospheric physics processes. We believe that this can be very useful to the many users of the WRF model to examine this sensitivity and to be aware of best-performing physics combinations using NOAH land surface as it is very widely used.

We added some sentences to the methodological section, and added a figure with the comparison RUC vs NOAH (new Figure 1).

In contrast to the use of a single land surface scheme there is a multitude of atmospheric physics schemes examined in this study that can be exchanged for one another, in particular there are six different boundary-layer/surface-layer schemes, but it is not at all made clear to the reader in what aspect they differ or how different they are. To a lesser extent similar considerations apply to the other physics schemes as well. Too many, unclear what their differences are, and in a way accidentally selected because these schemes happen to be implemented in the WRF system, which doesn't help in making this a "clean" ensemble, meaning that there is no way in which the various members of the ensemble can be neatly discerned from each other in some model physics phase space. In that respect, this ensemble is not so very different from the multi-physics ensemble approaches mentioned in the paper.

We would have liked to test all possible physics combinations, but this was not possible due to the large number of combinations. The selection was however based on a strategy. First we performed preliminary tests to see the behavior of the physics and to exclude least-advanced ones (such as the land-surface scheme without soil moisture). Then we looked into the different schemes to see the physics behind, in order to choose those that were most different from one another. For example in the cumulus options we used only one of the two Arakawa-Schubert Schemes, the 'new' one and left the 'old' one out of our study, and in the planet boundary schemes we tried to use different physics (K2 non-local; TKE; MF). We might not have explained in detail on what we based our decision. We added a part to the methodological section 'physics schemes' in order to explain better.

The methodology developed here is innovative. As compared to previous studies, our ensemble is very large, a lot of combinations were tested, and furthermore one originality is that our ensemble is dedicated to simulate heat waves, which has (in our knowledge) not been done before.

To conclude, in my opinion ignoring variations in the type of land surface scheme in building up a multi-physics ensemble makes the approach that is followed quite out of balance, in particular when such ensemble is meant to draw conclusions for the model ability to represent heat wave episodes. I would argue that the authors should first carry out the future study they announce in their "Concluding remarks" in which they intend to investigate the performance of a joint permutation of different atmospheric physics schemes and land surface schemes. On the basis of the results from that study they could then have solidly zoomed in on what they want to present in this study.

We hope to have answered most concerns, and justified our choice to focus on atmospheric processes only.

MAJOR COMMENTS:

1.) The title of the paper should already contain a reference to WRF to directly inform the reader what the paper is about. WRF is not synonymous nor equivalent to RCM, thus conclusions drawn for WRF can probably not immediately be generalized to an arbitrary RCM.

We propose to change the title into : 'An observation-constrained multi-physics WRF ensemble for simulating European mega-heatwaves.

2.) Following my general comment on the omission of having varied the land surface scheme which

makes the role of the NOAH LSM that is being used even more relevant I want to bring up some issues on the initialization of soil moisture. The authors write that soil moisture is directly taken from ERA-Interim. The authors also state that varying the initial date (1st of May against before 1st of May) results in variations in temperature outcomes of less than 0.5 C, at least for the August 2003 case. I would argue that this result does not at all imply that soil moisture is adequately initialized, it only indicates that there is little sensitivity regarding the date of initialization. My concerns are the following.

• ERA-Interim employs TESSEL as the land surface scheme which is rather different from the NOAH LSM. How was the actual mapping of TESSEL soil moisture onto NOAH soil moisture be carried out. Through interpolation of soil moisture (in relative volumetric units) or through interpolation of a soil moisture index taking account of the soil moisture field capacity and wilting point parameters in the respective schemes? The 2nd approach is obviously preferable. Please mention in the text what approach is followed in this respect.

• But even if they did, there are issues of soil water buffering capacity. Soil depths in both schemes may be difficult so, while properly mapping soil moisture content from TESSEL to NOAH, the resulting soil moisture columns might still be quite different. The authors should mention this point.

Initial soil moisture is obtained from interpolation of ERA-Interim data on the soil layers, accounting for the field capacity. We verified that soil moisture did not differ much in the upper layers from TESSEL, but we agree with the reviewer that initialization is an issue. However, as heat waves occur 2 months after initialization, memory from this initial condition is mostly lost. We performed initial condition experiments to select the starting time of the runs, and found that 1 May does not provide very different results from 1 April, to verify absence of spin-up sensitivity.

• Also, soil moisture must still be predominantly regarded as a model (or module) specific quantity. It is poorly constrained by observations and in the context of data-assimilation often treated as a free parameter that can be used (or abused) to reproduce observed near-surface parameters like temperature and relative humidity.

We agree with the reviewer. This is one of the reasons which also led us to focus on atmospheric processes, having this in mind.

So, on the basis of these grounds I would argue that there is considerable amount of uncertainty that can be attributed to the initial soil moisture profile used as the starting point for the set of WRF-simulations. I think the authors could benefit from this situation. Instead of taking a second and, even, third, land surface scheme, (still the most recommendable option, though) or altering the formulation of the land surface scheme, they might bring in the potential role of soil moisture on the evolution of the seasonal slice simulations by perturbing the initial soil moisture profile. For example, an option is to use plausible dry and wet perturbations of soil moisture initial profiles to examine the sustainability of the 5 or 55 best performing configurations that came out of their current exercise. (I don’t think it is necessary to redo all 216 configurations that have been done so far).

We did a new experiment as suggested by the reviewer. We made simulations with our 5 best performing configurations initializing the soil moisture differently. Instead of starting with the normal/original amount of soil moisture, we took out 20% of soil moisture (in all layers) in the first

experiment, and added 20% of moisture in the second one.

We found that drying the soil all along the column led to a general increase of all temperatures in the heat wave period, and a wetting led to a general decrease of all temperatures (regardless of physics schemes combinations). This confirms the sensitivity of temperatures to the land surface scheme. However, all perturbed runs still perform relatively well, suggesting that, despite the temperature shift, variability remains well simulated.

We added a description of this sensitivity test in the discussion section, and added an extra figure to the paper (Figure 8).

3.) The Russia-region is quite close to the eastern boundary of the EURO-CORDEX domain. This is potentially problematic in simulations of anomalous circulations like those that give rise to heat wave episodes, where there might be a conflict between the circulation sustained by the regional model and the forcing imposed at the near-by eastern boundary. Please, make clear in the text how you paid attention to this aspect. Did you choose a model domain that is actually much larger than the EURO- CORDEX domain? Did you broaden the boundary relaxation zone? Anything else?

We did not take this point into account. However for the calculation of the variables over Russia we excluded the pixels too close to the border of the domain.

MINOR COMMENTS: (in indicating the line number I refer to the line number value in the pdf-document of the discussion paper as outlined by GMD).

1.) It seems to me that the proper spelling of “heatwave” must be “heat wave”; similarly “mega-heatwave” should be spelled as “mega heat wave” (see Wiktionary).

In different papers we find different spelling of 'heat( )wave'. We chose to use the version without space, but now changed it into the version with the space, as suggested by the referee.

2.) Page 7862, line 2. “Climate models are often not evaluated ...” This is a too strong statement. Take e.g. Vautard et al. and Kotlarski et al. who evaluated models contributing to EURO-CORDEX.

We changed this sentence into: 'Climate models are not often evaluated...', to make the statement less strong.

3.) Page 7862, line 6: “sensitive” Sensitive to what?

The most sensitive physics (e.g. Convection, micro-physics etc.). This has been adapted in the text: with the goal of detecting the most sensitive physics.

4.) Page 7862, lines 7-9: these 55 combinations that can reproduce” can so because they satisfy an pre-imposed criterion. Please mention that the criterion is less than 1 degree bias during the heat wave episodes.

We changed the sentence into : 55 out of 216 simulations combining different atmospheric physical schemes have a temperature bias smaller than 1 degree during the heat wave episodes, the majority of the simulations showing a cold bias of on average 2-3 degrees Celsius.

5.) Page 7862, line 11: The statement “..and short wave radiation is slightly underestimated” seems to contradict the results discussed in the paper which clearly show that the model simulations in the

mean overestimate the observed radiation, and that this hold for the majority of the five best-performing configurations. Please, restate.

We changed 'underestimated' into 'overestimated'.

6.) Page 7863,lines 13-14: according to Weisheimer et al., 2011, the enhanced sophistication combined in land surface hydrology, convection and radiation proved key (their words) for a successful reforecast of the 2003 summer in Europe.

We adjusted the sentence and added that radiation needed to be adjusted as well, as is indeed described in the paper of Weisheimer.

7.) Page 7863, line 15: What is meant with “easily”? Please explain.

This has now been explained: “because model biases are mixed with sensitivity to initial conditions”

8.) Page 7863, lines 15-19: Sentence starting with “However ...” This a a very long and difficult sentence. Please, cut into pieces and clarify. E.g. what is meant with “ the effect of the representation of physical processes.”

We cut the sentence and rephrased in: However seasonal forecasting experiments do not easily allow the assessment of model physical processes underlying extreme temperatures during heat waves. This is partly due to their sensitivity to initial conditions. These may inhibit the effect of the representation of physical processes in reproducing the exact atmospheric circulation when starting simulations at the beginning of the season.

9.) Page 7863, line 23: “hindcast simulations” □ “evaluation experiments” in CORDEX-compliant terminology.

We changed 'hindcast simulations' into 'evaluation experiments' as suggested by the reviewer.

10.) Page 7863, line 26-27: Regarding the line “For some models ...for the 21st century”, I am wondering where it comes from. It is not a conclusion taken from Vautard et al. 2013, because that paper was only about ERA-Interim forced RCM- simulations. I am also wondering what the authors intend to say with this line. Because I do not immediately see how the size in bias can be connected with the projected temperature change in the coming century, be, please, more explicit in what you want to say.

We decided to remove this sentence because indeed it is a bit confusing.

11.) Page 7863, line 27- page7864, line 3. “Individual ... internal variability”. Again this is very long sentence, and I do not quite understand what you mean to say with the second part starting with “ because ...” Please, clarify.

We cut this sentence into different parts and adapted to make the second part more clear: 'Individual mega heat waves were reproduced by most models. However, it was difficult to infer whether these models could also simulate associated processes leading to the extreme heat waves. The exact same events with similar atmospheric flow and its persistence could not be reproduced due to internal variability of the models.

12.) Page 7864, line 28: please change “using the same model” into the “using the same model

system", because these groups are not using the precise same model.

This has been changed.

13.) Page 7864, line 29: What is meant with "democracy-driven". Please clarify.

We meant that instead of using all available models ('classical multi-model ensemble'), in some experiments different parameterizations of one model are used that are selected by different research groups. Using all these parameterizations from the different groups in one experiment, leads to a sort of democracy driven choice.

14.) Page 7868, lines 8-12: This a again a very long sentence. Please, break up in parts.

A part of the sentence is removed. It is now: 'From this final ranking, and in order to propose a reduced multi-physics ensemble of five combinations, we successively selected the highest-ranked schemes'.

The last sentence was not longer applicable, as the 5 highest ranked configurations differ already by two schemes from each other.

15.) Page 7868, line 22: It is not entirely clear to me what criterion was used by the authors to determine the extreme configuration. Is it only based on "daily mean temperature"? Throughout the paper there is only one set of two extreme configurations (am I right?), which is used in Fig 1a-h and Fig S2. Or are there separate sets for France and Russia? I think it would be very helpful if you explicitly state how these extreme combinations are configured. (I might have missed it, but I couldn't find it spelled out).

The two configurations are simply chosen to show the consistency of 'warm' and 'cold' simulations. They are chosen based on daily mean temperature over France during the 2003 heat wave. They are not separate sets for different regions or years, because that would eliminate their goal: to show the consistency. We added this explication to the text: 'In Fig. 1, we select two extreme configurations (blue and red lines), based on daily mean temperature over France during the 2003 heat wave. Interestingly, they are extreme in all regions and years, indicating that each combination tends to induce a rather large systematic bias.'

16.) Page 7868, lines 23-24. The "large bias" mentioned in these lines is certainly not always large, specifically not for the extreme configuration on the warm side. Please mention.

We adapted the sentence to mention this point: 'In Fig. 1, we select two extreme configurations (blue and red lines). They are interestingly extreme in all three cases, indicating that each combination tends to induce a rather large systematic bias. This bias however, is different for the 'warm' and 'cold' extreme configuration'.

17.) Page 7869, lines 9-12: "The two selected extreme ... misrepresentation of the land water supply" What is meant with "land water supply"? Soil moisture content or evaporation/evapotranspiration from the land surface to the atmosphere? I find the argument presented in this line indeed quite suggestive.

We mean to say that the temperature bias in the two extreme configurations seems not to be due to too much or to less rain (water supply). For example if the 'cold' extreme would have had way too

much precipitation, it could have been a reason for the low temperatures, but this does not seem to be the case. To make this more clear we changed the sentence into: 'The two selected extreme combinations are reproducing precipitation overall without a major bias. This suggests that the temperature bias in these two extremes is not explicitly caused by misrepresentation of atmospheric water supply from precipitation.'

18.) Page 7869,' line 12: What is meant with "soil moisture", and also in Figs 2 and S3? Soil moisture of the top soil layer (how thick) or averaged over the whole soil column?

With soil moisture the moisture over the whole soil column is meant. To clarify this point we added it between brackets: 'However soil moisture (the soil moisture over the whole column) does show a strong relation to temperature biases in model simulations.'.

19.) Page 7869, lines 17-19: this sentence "This indicates ...summer precipitation" precisely underscores why there should have been at least two different land surface schemes included in this study.

We agree with the reviewer that to answer this question different land surface schemes could be used, but because of reasons mentioned above we chose not to do so.

20.) Page 7869, lines 20-21: "For solar radiation ... approximately 100W/m2 ..." Difference in solar radiation over France and Russia, or differences over one region within an physics-ensemble. Also, solar radiation over Russia is not shown in Fig 1g or 1h.

There is a mistake in the sentence. It should have been over France for 2003 and 2007 instead of over France and Russia. This is now adapted. Solar radiation data over Russia is very scarce, and so we were not able to compare the model simulations with radiation observations over this area, as is explained in the methodology section.

21.) Page 7869, lines 21-28: Apparently there is a discernible overestimation of solar radiation in the warmest extreme configuration which is not translated in an overestimation of near-surface temperature. So accordingly the authors suspect there is a cooling mechanism without mentioning what that mechanism would be. This is the interesting part. Is it compensated by an overly large reflected solar radiation (unlikely) or is it participated differently over sensible and latent heat flux than in nature, such that latent heat flux is overestimated. Yet, this is not giving rise to more precipitation (no large precipitation bias, see above), nor to more clouds (positive solar radiation bias), nor is it drying out the soil (because the excess latent heat flux continues, otherwise the partitioning of excess solar radiation would go into sensible heat flux giving rise to higher near-surface temperature.) Please, try to identify what this cooling mechanism could be.

We are not sure about the cooling mechanism. It might partly be the reflection of the solar radiation, but maybe more importantly and overestimation of latent heat flux (which does not necessarily need to lead to higher precipitation rates). However, it is almost impossible to be sure about this, due to the scarce observations of the land heat fluxes and clouds.

22.) Page 7870, line 3: Please rephrase "In order to identify the most sensitive schemes for the development of heatwaves ..." as "In order to identify the parameterizations (or parametric schemes or physics schemes) to which the development of heat waves is most sensitive ...". Schemes themselves are not sensitive! Check the remainder of your manuscript wrt the use of "sensitive".

We rephrased the sentence.

23.) Page 7871, lines 23-24: “The overestimation ...for other regions and years ...” I tend to disagree, I find the latent heat flux figure for Russia 2010 (Fig S5e) not very different from the result shown for France 2003 (Fig. 3d). I am wondering how the France-2007 time series for latent heat flux looks like in this respect. Is that comparison available?

Yes, it is available. The latent heat flux in France 2007 seems still to be overestimated, although maybe in a lesser degree (especially during late summer) than France 2003. However, we also looked at the Iberian Peninsula and Scandinavia, and especially in Scandinavia the latent heat flux seems not to be overestimated. We added an additional figure in the supplementary material to strengthen the statement.

24.) Page 7871, line 26: “cross-validation” □ “cross-comparison” (also page 7873, line 7, and in first line of the caption of Table 3)

These two cases are changed.

25.) Page 7872, line 7-11: The first sentence of the section “Concluding Remarks” is again a very long sentence. It is also not a a very adequate line. Why using the word “small” in front to set, you considered all available combinations in this context. Also the phrase “with a given accuracy thresholds for temperature, precipitation and shortwave radiation” is not clear to me. What kind of thresholds have been used for precipitation and short wave radiation?

We adapted the sentence to make it more clear: 'In this study we designed and analyzed a large multi-physics ensemble. It is made of all possible combinations of a set of different atmospheric physics parameterization schemes. They were evaluated on their ability to simulate the heat waves of 2003 and 2010 using the regional climate model WRF based on temperature, precipitation and shortwave radiation'.

26.) Page 7872, line 23-27: the conclusion might be that the performance of a configuration is related to its ability to adequately represent cloud parameters (cloud amount, liquid water, etc.) or cloud-radiative interactions. In that respect I am wondering which parameterization within WRF is representing the stratiform – or layer-clouds? Can you comment on that.

Stratiform/layer clouds are described in the micro physics schemes, together with other cloud parameters such as particle types. The interactions between clouds and radiation are mostly described in the radiation schemes. However, the performance of the configurations seem to be more sensitive to the convection schemes, where the convective clouds are described.

27.) Page 7873, line 4-6: replace “schemes” by “configurations” or “combinations” or “members”; “scheme” refers to a single parameterization, that is not what is meant here.

This is now replaced.

28.) Page 7873, lines 11-16: That points to the heart of the matter as I already mentioned under general comments.

We agree, but hope to have better explained our choice of using only one soil scheme.

29.) Page 7873, lines 17-26: Please mention explicitly that the conclusions from your investigation are only valid for heat wave conditions. There is no guarantee that the constrained ensemble is also better performing for e.g. wet summer conditions or winter conditions.

The ensemble was also constrained for the summer of 2007, which was a wet year. So although it is true that the ensemble was mostly trained for heat wave conditions, it also performs relatively well in a wet summer year. Winter conditions were not tested in this study, although primarily results of a next study indicate that winter temperatures (and precipitation) are also quite well simulated. We now mention however that the configurations were not tested on winter conditions.

30.) There are two schemes in Table 1 assigned with number (6), namely WRF-SM6 and Tiedtke. Is that correct?

Yes, this is correct. For all different physics (radiation, micro-physics, convection, planet-boundary, and surface physics), the schemes are numbered starting with '1', so the schemes can indeed have the same numbers.

31.) I would strongly recommend to split Fig 1 into three Figures, because it is very difficult to read. Fig 1i becomes Fig 1, Figs 1a-c become Figs 2a-c, use column- format like Figs 2a,c,e. Figs 1d-h become Fig 3, also column-format is preferred.

We changed the order of the figures. Figure 1 is splitted into several figures as was suggested by the reviewer.

32.) Figs 1d,e,f: Preferably use same y-axis range and start at 0.

We changed the y-axis range, as suggested by the reviewer.

33.) Figs 1g,h: Preferably use same y-axis range

We changed the y-axis range.

34.) Fig 2, but also Figs S3 and S4. It is quite hard to distinguish the points by their different colours. It would help to choose different plotting symbols as well.

We adapted the figures. Now different schemes are represented by different colors and different symbols.

OTHER POINTS:

1) Page 7862, line 7-8: “55 Out of ...” □ “55 out of ...”

Corrected

2) Page 7862, line 13: “4” □ “four”

Corrected

3) Page 7863, line 1: Use “evaluated” instead of “validated”

Corrected

4) Page 7863, line 26: leave out “Celsius”

Corrected

5) Page 7864, line 22: “with different set” □ “with different sets”

Corrected

6) Page 7865, line 11: “the number ... were limited”□ “the number ... was limited”

Corrected

7) Page 7867, line 1: “Tawari” □ “Tewari”

Corrected

8) Page 7868, line 11: “to keep”□“in order to favour”

The part of the sentence with ‘to keep’ has been removed.

9) Page 7869, line 1: “maximal” □”maximum”

Corrected

10) Page 7869, line 5: “during heatwaves years” □ “during heat wave years”

Corrected

11) Page 7869, line 6: “in a lesser extent” □”to a lesser extent”

Corrected

12) Page 7869, line 7: “findings found” □”findings reported”

Corrected

13) Page 7869, line 22: “under” □”below”; “the middle of the simulations” □ “the mean value (the median value?) derived from the simulations”

Corrected

14) Page 7870, line 2: “how temperature clusters” □ “how resulting temperatures are clustered”

Corrected

15) Page 7870, lines 18-19: “affect radiation before heatwaves” □“affect radiation prior to the onset of heat waves”

Corrected

16) Page 7870, line 23: “of Sect. 2” □”introduced in Sect.2”

Corrected

17) Page 7870, line 24: “model-data” □ “model-observation”

Corrected

18) Page 7871, lines 1-2: “The same is found ..” “The same is not only found ...”. Please also indicate for each statement the season and region. “for the years 2007 and 2010 in Russia” probably should be interpreted as “for the years 2007 in France, and 2010 in Russia”.

Corrected

19) Page 7871, line 21: “... are largely overestimating ...” ”... are found to considerably overestimate ...”

Corrected

20) Page 7871, line 23: “Tiedke” ”Tiedtke”

Corrected

21) Page 7871, line 25: “fairly simulated” “fairly well simulated”

Corrected

22) Page 7872, line 27: “before” “prior to”

Corrected

23) Page 7873, line 1: “feedback” “feed back”

Corrected

24) Page 7873,,line 4: “atmospheric schemes” “atmospheric physics schemes”

Corrected

25) Caption of Table 1: “Physic schemes” “Physics schemes”

Corrected

26) Caption of Figure 1: “Daily time series of temperature” “Time series of daily mean temperature”

Corrected

27) Caption Supplementary Figs 2: “2a-d” “2a-f”

Corrected

Answers reviewer 2:

* General comments:

This work analyses an unprecedented (to my knowledge) multi-physics ensemble consisting of 216 summer seasonal simulations, focusing on heat waves over France (2003) and Russia (2010). It provides a fully systematic approach towards the selection of an optimal sub-ensemble to represent heat waves. The paper presents fairly novel concepts and ideas and a new dataset which will probably feed subsequent work. Therefore I suggest the paper to be published after a minor revision taking care of the specific comments below, which mainly refer to further discuss some points and solve some doubts to improve the reproductibility of the results.

* Specific comments:

1) The authors intend to create an optimized WRF ensemble for heat waves (7873:910). What would be the use of such ensemble? The experimental setup used should be taken into account. The simulations shown were run for a few months nudged towards the observed flow. For climate change simulations, nested into a GCM, nudging the atmospheric flow could be a problem. Also, long-term simulations could build up biases not arising in a few months (e.g. related to soil moisture). For seasonal forecasting, the authors recognize problems (7863:15-19) to reproduce observed events due to the chaotic nature of the atmospheric circulation.

The ensemble could be used for climate change modeling studies. All 5 members of the reduced ensemble differ in physics, which could serve as uncertainty measure.

2) The 5-member sub-ensemble was only tested for heatwaves. The 2007 "normal" season is not shown in Figure 3 or any of those in the supplementary material (even though it is stated that they are in the Suppl. material in 7871:2). If these members are the best in any physically meaningful sense, they should also perform well in the "France 2007" case study. Is that the case?

Yes, they also perform well in 2007, although the spread of the whole ensemble (216 members) is smaller in 2007 than during the heatwave cases. We added some figures of 2007 in the supplementary material.

3) There are already examples in the literature of "sub-ensembles" breaking model democracy, in which the sub-ensemble outperforms the full ensemble. For instance, Herrera et al (2010) selected a sub-ensemble using mean precipitation and show that this sub-ensemble is also well fitted for extreme precipitation regimes. This result is close to the results found in this work (sub-ensembles selected for a heatwave work well for other regions or regular seasons –if this is the case–), and could be added to the discussion, given that it extended the idea to multiple models.

We added this study in the discussion section.

4) The potential implications of the study for climate modeling (7873:17-) need to be discussed in a wider framework. The authors constrained the ensemble to a particular season, variables and error metrics. In this way, they were able to find an "optimized" set of configurations. However, It has long been recognized (Fernandez et al, 2007), that in a climate simulation an optimal configuration cannot be chosen. Biases and the best-performing configurations heavily depend on the season (Garcia-Diez et al, 2013), variable and even on the metric used (Jerez et al, 2013).

Yes this is true. We added this in the discussion section. Primary results from a next study suggest however that the small ensemble also performs relatively well in other seasons.

5) Moreover, observational uncertainty was not considered. It has been shown that the reference observations affect model rankings (Gomez-Navarro et al, 2012). This needs to be discussed at some point in the paper.

We added this in the methodological section, where we discuss the ranking method.

6) Jerez et al (2013) did not use WRF (7865:15). Other potential references here are Awan et al (2011) and Mooney et al (2012). Also, multi-physics ensembles did not start with WRF. There are a few other works with its predecessor, MM5.

We changed the reference of Jerez in Awan and Mooney, and added that earlier studies have been done with MM5.

7) Weisheimer et al (2011) did not use WRF (7870:15). Remove or rephrase.

We removed the reference.

8) The authors find probable (7872:18) that the inclusion of another land surface model would increase the ensemble spread. This statement can be accompanied by a cite to Mooney et al (2012), where they show strong differences when changing the LSM (see their Fig. 2b, e.g. Sim 9 vs. 11). It is not clear at all why the LSM sensitivity was left out of the study. There is plenty of literature (even cited by the authors) highlighting the role of soil-atmosphere interactions in the development of heat waves and the authors themselves recognize it (7869:16).

This paper is designed as a methodological paper. Our intention was to test as many physic schemes as possible, including land-surface schemes. However, at the moment we performed the study 4 different land-surface schemes were available: NOAH, RUC, Pleim-Xiu and the 5-layer thermal diffusion scheme. The last one, non-physical, is designed for test cases and cannot be used in realistic situations. The Pleim-Xiu scheme comes with a dedicated set of other physical schemes and does not allow in most cases to combine different possibilities. We performed an ensemble of simulations using the RUC scheme, but we found that it provided extremely sensitive fluxes, with large latent heat fluxes at the beginning of season and extreme subsequent drying in summer months. Sensible heat fluxes also appear overestimated. Comparisons with several FLUXNET sites are now explicitly shown in Figure 1. So even if temperatures of heat waves would match observations in a few combinations of physics schemes, we would almost be sure that this would be for wrong reasons. We also experienced technical problems while running several of the RUC simulations. This made us decide to only use one land-surface scheme and focus on the atmospheric physics processes. We believe that this can be very useful to the many users of the WRF model to examine this sensitivity and to be aware of best-performing physics combinations using NOAH land surface as it is very widely used.

We added some sentences to the methodological section, and added a figure with the comparison RUC vs NOAH (new Figure 1).

9) I don’t agree with the sentence (7870:12) "By contrast, heatwave temperatures do not seem very sensitive to the planetary boundary layer and surface layer physics schemes". Figure 2d seems

noisier than the rest because there are more PBL options tested. However, there is a clear, systematic temperature dependence on the PBL. If you imagine a regression line for each PBL scheme, all of them preserve the relationship (slope) with soil moisture, but the the heat wave average temperature is clearly different.

We adapted the text a little bit to strengthen the point mentioned by the reviewer. The sentence mentioning this point is now: 'Heat wave temperatures seem to be least sensitive to the planetary boundary layer and surface layer physics schemes'.

10) The caption of Figure 2 says correlation where scatterplot is meant. These particular plots show that there is indeed (negative) correlation, but the plots are scatterplots.

Correlation is changed into scatter plot.

11) X-axis labels in Figs 1 and 3 read "Time (DOY)". I assume it means Day Of the Year but, please, define. Also, in the panels with this axis, two vertical lines showing the heatwave period considered would help, given that different periods were chosen for each event. Also, if Fig 1bc shared the Y-axis with Fig 1a, they could be directly compared with each other. Currently, the normal year seems as hot as the 2003 heat wave, when in fact it is 5K colder.

DOY is now defined in figure 1a, and we changed the y-axes of figure 1 to be directly comparable to each other.

12) In Fig 3c-August, the cyan circle is missing (probably hidden behind other member). Using non-overlapping symbols would help. The same happens with the pink circle in Fig. 3B

Different symbols are now used so there is no longer full overlapping.

13) The resolution is stated to be 50km (approx 0.44deg). Was a Lambert grid projection in Kms used? or the Euro-Cordex standard 0.44 rotated lat-lon grid? Please, clarify.

We used the Euro-Cordex standart 0.44 rotated lat-lon. This is now added to the methodological section.

14) Observational data is not fully described. Which E-OBS version was used? Was any interpolation carried out in the analyses?

This is now clarified in the methodology section.

15) The pre-screening of the simulations considering only those within 1K of the E-OBS temperature might be problematic. RCMs have biases. With the method proposed, a fairly physically-consistent simulation could be disregarded, while a simulation unrealistically compensating temperature biases might get in. The latter can easily happen (Garcia-Diez et al. 2014).

It is true that the 1K temperature bias is rather arbitrary. But because we are really interested in heat waves, and especially the high temperatures, we chose to have this limit anyway. We agree on the fact that within the 55 simulations within the 1K limit, their might be configurations that compensate temperature biases. However, because the simulations are also tested on their ability to simulate well the precipitation and the radiation, we do not expect that this is the case for our 5 (or

even 10-15) best simulations.

16) The ranking metrics are not fully clear to me. Daily temperature differences are used (7867:26). But, which score was built out of them? RMSE? Why was temperature considered at a daily scale and precipitation at monthly scale?

For temperature we used the bias, not the RMSE. Because modeled daily precipitation is much noisier than daily temperature, we decided to used monthly values for this variable. Furthermore radiation data was only available on a monthly timescale.

17) For radiation data, was the model interpolated to the stations to compute the spatial averages? which interpolation method was used? How many radiation stations were available in each region?

Yes, the model data was interpolated to the station and spatial averages were computed. We used 'nearest neighbor' for the interpolation. We did only consider France for the radiation, because over Russia the observation data was too scarce. For France we used 21 stations for 2003 and 20 for 2007. This information is added in the methodology section.

18) Regarding the rejection of the members differing in only one scheme (7868:08): How many of these members were disregarded to get the top-5? What is the interest of "keep[ing] a large range of different realistic physics combinations between the simulations" (7868:11)? I see also an interest in the single-step ensemble members. In these ones, the differences can be traced to the single scheme that changed. For instance, "The two simulations that are largely overestimating latent heat flux" (7871:21) are those not using Tiedtke, but this could be just by chance, given that they also differ in other schemes and the schemes interact in a non-linear way (Awan et al. 2011).

Finally the 'top 5' members already differed with two schemes from each other, so this rejection was not longer necessary. Because we are looking for a variability of physics, we thought about using this rejection. We removed the sentences from the text.

19) How were the extreme configurations selected (7868:22)?

The two configurations are simply chosen to show the consistency of 'warm' and 'cold' simulations. They are chosen based on daily mean temperature over France during the 2003 heat wave. They are not separate sets for different regions or years. We added a sentence in the text to explain this better.

20) By "the middle of the simulations" (7869:22), I guess you mean the "median".

Yes, this has been changed.

21) At some point (7870:18), the effect of convective clouds on radiation is invoked. However, note that in WRF the interaction of radiation with sub-grid clouds has only recently been implemented (Alapaty et al. 2012) and included in WRF3.6 for certain combination of radiation and cumulus schemes. It was not included in the version used in this work (WRF 3.3.1).

We agree with the reviewer.

22) The discussion in 7871:18-25 seems to imply (although it is not explicitly stated) that the good performance of the Tiedtke scheme just during the heatwave is just by chance.

We did not mean to imply this, as we found that the Tiedtke scheme is performing quite well

overall. However, we cannot state that the other convection schemes do not simulate the latent heat flux very well, as this is not the case for other years. But we found that also in some other cases, the two of the five best configurations not using Tiedtke, are performing a little bit less well, for example precipitation over the Iberian Peninsula in 2003 (Suppl. Fig. 5c).

23) "We found a large spread" (7872:11) I would highlight, just at the beginning of this sentence "Even though the simulations were constrained by grid nudging,"

This is now added to the text. Thank you for the suggestion.

24) The journal recommendations suggest that "The model name and number should be included in [the title of] papers that deal with only one model". Replace RCM by WRF in the title.

We propose to change the title into : 'An observation-constrained multi-physics WRF ensemble for simulating European mega-heatwaves.

(beware I'm not a native speaker)

7862:17, "together with varied physics scheme." sounds odd to me. Please, rephrase.

This has been rephrased.

7866:04, "temperatures differ by less among one another than 0.5C" sounds odd.

Rephrased to: 'temperatures differ by less than 0.5C among one another'.

7867:01, "Tawary" should read "Tewary".

Corrected.

7869:26, "better" -> "well" (or "better than [what?]")

We rephrased: 'better than the coldest simulation'.

7872:27, missing period "heatwaves Changes"

Corrected.

An observation-constrained multi-physics WRF ensemble for simulating European mega heat waves

Annemiek I. Stegehuis[1], Robert Vautard[1], Philippe Ciais[1], Adriaan J. Teuling[2], Diego G. Miralles[3,4] and Martin Wild[5]

[1] LSCE/IPSL, Laboratoire CEA/CNRS/UVSQ, Gif-sur-Yvette, France

[2] Hydrology and Quantitative Water Management Group, Wageningen University, The Netherlands

[3] Department of Earth Sciences, VU University Amsterdam, Amsterdam, The Netherlands

[4] Laboratory of Hydrology and Water Management, Ghent University, Ghent, Belgium

[5] ETH Zurich, Zurich, Switzerland

Corresponding author : A.I. Stegehuis, LSCE/IPSL, Laboratoire CEA/CNRS/UVSQ, 91191 Gif-sur-Yvette CEDEX, France

Abstract

Climate models are not often evaluated or calibrated against observations of past climate extremes, resulting in poor performance during for instance heat wave conditions. Here we use the Weather Research and Forecasting (WRF) regional climate model with a large combination of different atmospheric physics schemes, with the goal of detecting the most sensitive physics and identifying those that appear most suitable for simulating the heat wave events of 2003 in Western Europe and 2010 in Russia. 55 out of 216 simulations combining different atmospheric physical schemes have a temperature bias smaller than 1 degree during the heat wave episodes, the majority of simulations showing a cold bias of on average 2-3°C. Conversely, precipitation is mostly overestimated prior to heat waves, and short wave radiation is slightly overestimated. Convection is found to be the most sensitive atmospheric physical process impacting simulated heat wave temperature, across four different convection schemes in the simulation ensemble. Based on these comparisons, we design a reduced ensemble of five well performing and diverse scheme combinations, which may be used in the future to perform heat wave analysis and to investigate the impact of climate change in summer in Europe. Future studies could include the sensitivity to land surface processes controlling soil moisture, through the use of varied land surface models together with varied physics schemes.

1. Introduction

An increasing number of simulations and studies project a higher frequency of several types of extreme weather events in the future (e.g. Schär et al., 2004; Meehl et al., 2004; Della-Marta et al., 2007; Beniston et al., 2007; Kuglistsch et al., 2010; Fischer and Schär, 2010; Seneviratne et al., 2012; Orlowsky and Seneviratne, 2012). Since summer heat waves are among the most problematic of such phenomena - threatening society and ecosystems - climate models used for future projections must provide accurate simulations of these phenomena, or at least their uncertainties should be documented. Even if climate models have been evaluated using observed weather in past decades, it is unclear whether they will be able to simulate extreme heat waves in future climates

that may not have analogues in the historical record. At least, models should be able to reproduce the conditions measured during recent extreme heat wave cases, some of them having been shown to be unprecedented when considering the climate over the past five or six centuries (Chuine et al., 2004; Luterbacher et al., 2010; García-Herrera et al., 2010; Barriopedro et al., 2011; Tingley and Huybers, 2013).

Given the importance of forecasting summer heat waves well in advance, many studies have analyzed their predictability, which remains poor in seasonal forecasts. For instance the 2003 European heat wave was not simulated realistically (neither timing nor intensity) by the operational European Centre for Medium-Range Weather Forecasts (ECMWF) system, but improvements were clear with the use of a new soil, convection and radiation schemes (e.g. Weisheimer et al., 2011; Dole et al. 2011; Koster et al. 2010; van den Hurk et al. 2012). However seasonal forecasting experiments do not easily allow the assessment of model physical processes underlying extreme temperatures during heat waves because model biases are mixed with sensitivity to initial conditions. These may inhibit the effect of the representation of physical processes in reproducing the exact atmospheric circulation when starting simulations at the beginning of the season.

From a statistical perspective, extreme temperatures have been found to be reasonably well represented in global simulations of the current climate (IPCC, 2013), as well as in regional simulations (Nikulin et al., 2010). In recent regional modeling evaluation experiments, using an ensemble of state-of-the-art regional models guided by re-analysis at the boundaries of a European domain, summer extreme seasonal temperatures were shown to be simulated with biases in the range of a few degrees (Vautard et al., 2013). Individual mega heat waves (2003 in Western Europe, 2010 in Russia) were reproduced by most models. However, it was difficult to infer whether these models could also simulate associated processes leading to the extreme heat waves. The exact same events with similar atmospheric flow and its persistence could not be reproduced due to internal variability of the models.

A comprehensive assessment of simulations of recent mega heat waves has only been the object of a limited number of such studies. Process-oriented studies of high extreme temperatures over Europe have focused on land-atmosphere feedbacks (e.g. Seneviratne et al., 2006 and 2010; Fischer et al., 2007; Teuling et al., 2009; Stegehuis et al., 2013; Miralles et al., 2014) because, beyond atmospheric synoptic circulation, these feedbacks are known to play an important role in summer heat waves. However, the sensitivity of simulated heat wave conditions to physical processes in models has not yet been explored in a systematic way. This could be important because error compensation among processes that involve land-atmosphere interactions, radiation and clouds may cause high temperatures for the wrong reasons (Lenderink et al., 2007).

The goal of the present study is threefold. First we examine the ability of a regional climate model, the Weather Research and Forecast (WRF, Skamarock et al., 2008), to simulate recent European mega heat waves, with a number of different model configurations. Analysis of these experiments then allows understanding which physical parameterizations are prone to reproduce the build-up of extreme temperatures, and thus the need for carefully constraining them in order to simulate these events properly. Finally, using observational constraints of temperature, precipitation and radiation, we select a reduced ensemble of WRF configurations that best simulates European heat waves, with different sets of physical schemes combinations. This constrained multi-physics ensemble aims therefore at spanning a range of possible physical parameterizations in extreme heat wave cases while keeping simulations close to observations.

Our multi-physics regional ensemble approach contrasts with the classical multi-model ensembles that are constructed by the availability of model simulations in coordinated experiments (see e.g. Déqué et al., 2007 and references therein) or combinations of parameterizations selected by different groups using the same model system (García-Díez et al., 2014). In the latter “democracy-driven” ensemble, the lack of overall design strategy may lead the uncertainty estimation to be biased and the models to be farther from observations. In addition, the real cause of model spread is difficult to understand because of interacting physical processes and their biases. Regional

perturbed-physics or multi-physics ensembles could help understand and constrain uncertainties more effectively, but so far they have been seldom explored. García-Díez et al. (2014) showed that even a small multi-physics ensemble confronted to several climate variable observations can help diagnose mean biases of a RCM. Bellprat et al. (2012) showed that a well-constrained perturbed physics ensemble may encompass the observations. Their perturbed physics ensemble was designed by varying the values of a number of free parameters, and selecting only the configurations that were closest to the observations; however, the number of combinations of different physical parameterization schemes was limited to a total of eight different configurations.

The WRF model offers several parameterization schemes for most physical processes, and is thus suitable for a multi-physics approach. In fact, a WRF multi-physics approach has been used in several studies (e.g. García-Díez et al., 2011; Evans et al., 2012; Awan et al., 2011; Mooney et al., 2013), also with its predecessor MM5, but not specifically to simulate extreme heat waves.

Here we run an ensemble of 216 combinations of WRF physical parameterizations, and compare each simulation with a set of observations of relevant variables in order to select a reduced set of 5 combinations that best represent European summer mega heat waves. The evaluation is made over the extreme 2003 and 2010 events. The ensemble is also evaluated for a more regular summer (2007) in order to test the model configurations under non-heat wave conditions.

2. Methods

Simulations and general model setup

We use the WRF version 3.3.1 and simulate the three summers (2003, 2007, 2010) using an ensemble of physics scheme combinations. We first test the time necessary to initialize the soil moisture on a limited number of cases. Soil conditions are initialized using the ERA-Interim (Dee et al., 2011) soil moisture and temperatures; thereafter soil moisture and air temperature are calculated as prognostic variables by WRF. For the August 2003 case, we find that temperatures differ by less

than 0.5°C among one another when starting experiments before May 1st. Thus in the current study, each simulation is run from the beginning of May to the end of August for the years 2003, 2007 and 2010. The regional domain considered is the EURO-CORDEX domain (Jacob et al., 2014; Vautard et al., 2013) and the low-resolution setup of 50 km x 50 km (~0.44 degree on a rotated lat-lon grid) is used – note that Vautard et al. (2013) recently concluded that a higher spatial resolution did not provide a substantial improvement in heat wave simulations. We use a vertical resolution with 32 levels for WRF. Boundary conditions come from ERA-Interim (as well as initial snow cover, soil moisture and temperature). In order to focus on physical processes in the boundary layer and the soil-atmosphere interface, and to avoid chaotic evolution of large-scale atmospheric circulation, we constrain the model wind fields with ERA-Interim re-analyses above Model Level #15 (about 3000m), similar to previous studies (Vautard et al., 2014), using grid nudging, with a relaxation coefficient of 5.10^{-5} s^{-1}, corresponding to a relaxation time about equivalent to the input frequency (every six hours) (Omrani et al., 2013). Temperature and water vapor were not constrained, to let feedbacks fully develop.

Physics schemes

We test 216 combinations of physics schemes. We consider different physics of the planetary boundary layer and surface layer (PBL; 6 schemes), microphysics (MP; 3 schemes), radiation (RA; 3 schemes) and of convection (CU; 4 schemes). For each type of scheme, a few options were selected among the ensemble of possibilities offered in WRF. The selection was made to avoid variants of the same scheme, and to maximize the difference of temperature and precipitation outputs in preliminary experiments. At the time of study and model development stage, different land-surface schemes were available in WRF: 5-layer Thermal Diffusion Scheme (Dudhia, 1996), NOAH (Tewari et al., 2004), Rapid Update Cycle (RUC) (Benjamin et al., 2004) and Pleim-Xiu (Gilliam & Pleim, 2010). We decided however to only use one, the NOAH land surface scheme, in order to focus our study on atmospheric processes while limiting the number of simulations, and because the NOAH scheme is the most widely used in WRF applications. This was also motivated

by the poor performance and extreme sensitivity of the RUC land surface scheme for the land latent and sensible heat flux as compared with local observations in 2003. It simulates strong latent heat fluxes in the beginning of the season and an extreme drying at the end, while sensible heat flux is overestimated. The NOAH scheme seemed more stable in the tests that were done for capturing both latent and sensible heat fluxes during the 2003 heat wave at selected flux tower sites in Western Europe (Figure 1). Furthermore the Pleim-Xiu scheme is especially recommended for retrospective air quality simulations, and is developed with a specific surface layer scheme as coupled configuration (Gilliam & Pleim, 2010). The last possible option is the 5-layer thermal diffusion scheme (Dudhia, 1996) which predicts ground and soil temperatures but no soil moisture, and is therefore also not suitable for our study. Table 1 describes the physical schemes that were combined to simulate the weather over the three summer seasons.

Observational data

In order to evaluate the ensemble and to rank and select its best performing simulations we use gridded observed daily temperature and precipitation from E-OBS with a 0.25 degree resolution (version 7.0) (Haylock et al., 2008). Bilinear interpolation is used to regrid E-OBS data and the model output to the same grid. Furthermore we use station data of monthly global radiation from the Global Energy Balance Archive (GEBA) network (Wild et al., 2009). For France 2003 the data of 21 stations were available, for 2007 this number was 20. Observations over Russia were too scarce, and were not considered. Model data are interpolated to these stations using the nearest neighbor method. In addition, in order to check land-atmosphere fluxes and the partitioning of net radiation into sensible and latent heat fluxes, we use the satellite observation-driven estimates of daily latent heat fluxes from GLEAM (Miralles et al., 2011). Since the latter is not a direct measurement we do not use them to validate and rank the model configurations. Furthermore latent- and sensible heat flux measurements are used from three FLUXNET sites from the Carbo-Extreme database (Neustift/Stubai – Austria (Wohlfahrt et al., 2010); Tharandt-Anchor station – Germany (Grünwald & Bernhofer, 2007); and Soroe-LilleBogeskov – Denmark (Pilegaard et al., 2009)), for

the evaluation of the land surface schemes.

Evaluation and ranking of model simulations

For ranking, we set up several measures of model skill, based on the differences between observed and simulated spatial averages over two domains: France for 2003 and 2007 (5W–5E & 44N–50N), and one in Russia for 2007 and 2010 (25E–60E & 50N–60N) (Fig. 2). A first scheme selection is made based on the skill to reproduce air temperature dynamics, since this is the primary impacted variable and observations are reliable. Because we are interested in heat waves, we select only those simulations that are within a 1 K regional average difference between simulated and observed temperature, for heat wave periods; these periods are defined as August 1st-15th for France (in 2003), and July 1st till August 15th for Russia (in 2010). The 1 K threshold is arbitrary but is used to avoid processing a large number of simulations that have unrealistic temperatures. Only 55 of the 216 simulations meet this criterion and are further considered. Then, the ranking of the retained simulations is done based on: (i) the daily temperature difference between simulations and observations during the heat wave periods (as above for 2003 and 2010), and during the period 1st-31st August for the normal year 2007, (ii) the root mean square error of monthly precipitation and radiation for the months July, June and August. The GEBA data set only contains scarce radiation observations over Russia, and therefore we could not consider this region for ranking models against incoming shortwave radiation. As a final step, an overall ranking is proposed by averaging the ranks obtained from the three variables (temperature, precipitation and radiation). From this final ranking, and in order to select an elite of multi-physics combinations, we selected the top-5 highest-ranked configurations. Note that observational uncertainty is not considered in this study, which is shown to be able to impact model ranking over Spain (Gomez-Navarro et al., 2012).

3. Results

3.1. Large systematic errors found during heat wave periods

Figure 3 shows the large temperature range spanned by the 216 ensemble members for the spatial

average over the heat wave areas. The min-max range between ensemble members is up to 5°C during heat wave periods (Figure 3). Locally at 50 km resolution, the difference between the warmest and the coldest simulation during a heat wave is larger, reaching more than 10°C in 2003 (Figure 3d). In 2007, when summer temperatures were not extreme, the range is about twice as small. Only a few simulations match the observed high temperatures (Figure 3a-c). In Fig. 3a, we select two extreme configurations (blue and red lines), based on daily mean temperature over France during the 2003 heat wave. Interestingly, they are extreme in all regions and years, indicating that each combination tends to induce a rather large systematic bias. This bias however, is different for the 'warm' and the 'cold' configuration. It seems not to be due to a misrepresentation of the diurnal cycle, since they remain when analyzing time series of maximum and minimum daily temperatures independently (see supplementary Figures 1a-f). However, minimum temperatures show a less consistent bias than maximum daily temperatures. A systematic temperature underestimation by WRF simulations over Europe has also been found in other multi-physics ensemble studies over Europe (e.g. Awan et al., 2011; García-Díez et al., 2011, 2014).

For monthly precipitation we obtain a large range of simulated values, with most configurations overestimating monthly summer rainfall (JJA) during heat waves years, and to a lesser extent during the wetter 2007 season (Fig. 4a-c). This is in line with the findings reported by Warrach-Sagi et al. (2013) and Awan et al. (2011), and with the overestimation of precipitation by many EURO-CORDEX models shown by Kotlarski et al. (2014). The two selected extreme combinations (based on temperature, as explained above) are reproducing precipitation overall without a major bias. This suggests that the temperature bias in these two extreme simulations is not explicitly caused by a misrepresentation of the atmospheric water supply from precipitation. However soil moisture (the soil moisture over the whole column) does show a strong relation to temperature biases in model simulations. Figure 5a-d shows soil moisture at the end of July versus temperature in August 2003 for each model configuration. Configurations with low soil moisture level are associated with higher temperatures and vice versa, confirming the role of land-atmosphere feedbacks during heat

waves, already pointed out by previous studies. This indicates that the evapotranspiration from spring to summer depleting soil moisture can be a critical process during summer for the development of heat waves, and that this process is not simply related to summer precipitation.

For solar radiation, the mean differences between our simulations over France 2003 and 2007 reaches approximately 100 Wm^{-2} (Fig. 6a,b). Observations for France (black dots) are found below the median value of the simulations so a slight overestimation of the ensemble is obtained. The first (warmest) extreme configuration (red dot) is associated with an overestimated radiation of 10-50 Wm^{-2} while the other (coldest, blue dot) extreme configuration exhibits an underestimated radiation by about the same amount. Since the warmest simulation agrees better with temperature observations than the coldest simulation, one may therefore suspect that it contains a cooling mechanism that partly compensates for the overestimated solar radiation.

3.2. Sensitivity of temperatures to physical parameterizations and sources of spread

In order to identify the physics schemes to which the development of heat waves is most sensitive, we examine how resulting temperatures are clustered as a function of the scheme used. We find that the spread between all simulations – both in terms of temperature and soil moisture – is mostly due to the differences in convection scheme (clustering of dots with the same color in Fig. 5a). For instance the Tiedtke scheme (blue dots) systematically leads to higher temperatures and lower soil moisture, while the Kain-Fritsch scheme (green dots) leads to wetter soils and lower temperatures, inhibiting heat waves. Microphysics and radiation schemes are also contributing to the spread of simulated temperature and soil moisture values (Fig. 5b-c), although their effect is less marked than for convection. Heat wave temperatures and soil moisture seem to be least sensitive to the planetary boundary layer and surface layer physics schemes. The sensitivity of the convection scheme in WRF has already been mentioned in previous studies (Jankov et al., 2005; Awan et al., 2011;; Vautard et al., 2013; García-Díez et al., 2014). Note that the soil moisture simulated in early August 2003 is better correlated with preceding radiation than with precipitation (compare Supplementary

Figures 2 and 3), indicating that the way clouds, and particularly convective clouds, affect radiation prior to the onset of heat waves is a major driver of the spread for the development of heat waves, higher radiation leading to drier soils and higher temperatures during heat waves.

3.3. A constrained reduced ensemble of best simulations

Focusing only on the 55 selected simulations that differ less than 1°C from the observations during the heat waves, we apply the ranking method introduced in Section 2 based on temperature, precipitation and radiation model-observation comparison metrics. The 5 highest ranked simulations are given in Table 2 and are actually the numbers 1-5 in Supplementary Table 1. Figure 7a confirms the ranking by showing that these simulations also perform well in terms of temperature, during the months prior to the heat wave. The same is furthermore found for the years 2007 in France (Supp. Fig. 5) and 2010 in Russia (Supp. Fig. 4), and also for other regions such as the Iberian Peninsula and Scandinavia (Supp. Fig. 6a,d). The selected simulations however performed less well for precipitation over France in 2003 (Fig. 7b), but do not show a large overestimation of precipitation either. Precipitation over Russia for the 5 highest-ranked simulations does show good performance (Supp. Fig. 4b), as well as for other European regions (Supp. Fig. 6). The mean radiation of the ensemble of the five best simulations is closer to the GEBA observations than in the case of the original ensemble (Fig. 7c).

Nonetheless, the better match of the reduced ensemble of the five highest-ranked simulations to the observations of temperature, precipitation and radiation is to a very large degree unsurprising: the selection was based on the fit to observations. However, it is still satisfactory to see that some simulations are capable of matching all three variables. Conversely, we also compare simulations against another key variable that was not used for evaluating and ranking simulations, namely the latent heat flux (Figure 7d). Albeit somehow reduced compared to the full-ensemble spread, the spread of the five best simulations for the latent heat flux remains large over the whole period, on average between 50 and 120 Wm^{-2} (observed values are around 75 Wm^{-2}). However, during the

2003 heat wave over France three of the five best simulations exhibit a close resemblance to the latent heat observations (approximately 5-10 Wm^{-2}) (Fig. 7d). The two simulations that are found to considerable overestimate latent heat flux by approximately 30-40 Wm^{-2} (as compared to GLEAM) are those that use a different convection scheme than the Tiedtke scheme. The overestimation of latent heat fluxes in these schemes is however not generalized for other regions and years (Suppl. Fig. 4c, 5d, 6c,f-h), for which the latent heat flux was fairly well simulated within the range of uncertainty of GLEAM.

A cross-comparison for the years 2003 and 2010, that is, using only the 2010 heat wave to select schemes and verify the performance of the selected schemes over 2003 and vice versa, yields some promising results. Table 3 shows the average ranking of the best (5, 10, 15, 20 and 25) simulations. When only using one heat wave to select the best configurations, they all lie in the top-ranked half, and even higher in the ranking in the case of the 2010 heat wave over Russia being used to select the best configurations. This suggests that the selection based upon one heat wave in one region should also provide better simulations for other heat waves or heat waves in other areas, i.e. that the bias of a member of the WRF ensemble is not local, but at least regional at the scale of Western Europe.

4. Concluding remarks

In this study we designed and analyzed a large multi-physics ensemble with the WRF model. It is made of all possible combinations of a set of different atmospheric physics parameterization schemes. They were evaluated for their ability to simulate the European heat waves of 2003 and 2010 using the regional climate model WRF based on temperature, precipitation and shortwave radiation. Even though the simulations were constrained by grid nudging, we found a large spread between the different physics for the simulations for temperature, precipitation and incoming shortwave radiation, three variables we used to create an overall configuration ranking. Most simulations systematically underestimate temperature and overestimate precipitation during heat

waves, a model pattern that was already found in previous studies dealing with much smaller ensembles (e.g. Awan et al., 2011; García-Díez et al., 2011; Warrach-Sagi et al., 2013). The spread among ensemble members is amplified during the two extreme heat waves of study. Since we only considered a single land surface scheme, it is probable that the ensemble spread would largely increase when incorporating the uncertainty associated with modeling land surface processes. Nevertheless, considering only atmospheric processes, the magnitude of the spread still reaches 5°C during the peak of the heat waves.

We also showed that among atmospheric process parameterizations, the choice of a convection scheme appears to dominate the ensemble spread. We found indications that the large differences between convection schemes seem to occur mostly through radiation, and therefore the way convective clouds affect the surface energy and water budget prior to and during heat waves. Changes in incoming radiation cause changes in evapotranspiration and therefore soil moisture, which may subsequently feed back on air temperature.

From this ensemble, we selected a small sub-ensemble with the five best combinations of atmospheric physics schemes based on the fit to observations. These combinations capture well the temperature dynamics during the mega heat waves of France and Russia, and they perform better than other combinations in other regions of Europe. In addition, they are consistent with independent latent heat flux data used for cross-validation. This indicates that the constraints set for the selection reduce the uncertainty across the whole European continent and points towards the creation of an optimized ensemble of WRF configurations specific for heat waves, with reduced error compensations. A sub-ensemble that outperforms a larger ensemble was also found by Herrera et al. (2010). The sub-ensemble based on mean precipitation showed better results for extreme precipitation as well.

However a limitation of this study is the use of only one land-surface scheme; the five selected WRF configurations may actually all be affected by systematic errors of the NOAH land surface

scheme. The importance of the selected land surface scheme is further confirmed by the larger spread of the "best" ensemble for latent heat (in Wm-2) than for shortwave radiation. In order to mimic radically different land surface processes, a sensitivity test where initial soil moisture was artificially increased and decreased by 20% all along the soil column was conducted. Results confirm the sensitivity of the temperature simulations to soil moisture, a variable partly controlled by the land surface scheme (Figure 8). The full answer to this question is left for a future study in which different atmospheric schemes and surface schemes will be jointly permuted.

Although our ensemble is trained on only summer conditions, our results have several implications for climate modeling. First, the constrained WRF ensemble may be used in future studies of climate change; each of the five members may exhibit a different sensitivity to future climate change conditions, leading to a constrained exploration of the uncertainty. Then it is important to notice that our study pinpoints the need to carefully design or adjust the convection scheme for a proper representation of the summer climate during heat waves. This is particularly important in order to evaluate the impacts of climate change on ecosystems, health, carbon cycle, water and cooling capacity of thermal energy plants, since heat waves in the mid latitudes are expected to be of the most impacting phenomena in a human altered climate. Therefore, impact studies can be designed based on the selected configurations.

Acknowledgments

AIS acknowledges CEA for funding as well as of the GHG-Europe FP7 project. AJT acknowledges financial support from The Netherlands Organisation for Scientific Research through Veni grant 016.111.002. P.C. acknowledges support of the ERC-SYG project P-IMBALANCE. The authors acknowledge K. Pilegaard, A. Ibrom, C. Bernhofer, G. Wohlfahrt and CarboEurope for sharing FLUXNET data.

References

Awan, N. K., H. Truhetz & A. Gobiet (2011) Parameterization-induced error characteristics of MM5 and WRF operated in climate mode over the Alpine region: an ensemble-based analysis. *Journal of Climate*, 24, 3107-3123, doi:10.1175/2011JCLI3674.1.

Barriopedro, D., E. M. Fischer, J. Luterbacher, R. Trigo & R. Garcia-Herrera (2011) The hot summer of 2010: redrawing the temperature record map of Europe. *Science*, 332, 220-224, doi: 10.1126/science.1201224.

Beljaars, A.C.M. (1994) The parameterization of surface fluxes in large-scale models under free convection. *Quart. J. Roy. Meteor. Soc.*, **121**, 255–270.

Bellprat, O., S. Kotlarski, D. Luthi & C. Schär (2012) Exploring perturbed physics ensembles in a regional climate model. *Journal of Climate*, 25, 4582-4599, doi: 10.1175/JCLI-D-11-00275.1.

Beniston, M., D. B. Stephenson, O. B. Christensen, C. A. T. Ferro, C. Frei, S. Goyette, K. Halsnaes, T. Holt, K. Jylha, B. Koffi, J. Palutikof, R. Scholl, T. Semmler & K. Woth (2007) Future extreme events in European climate: an exploration of regional climate model projections. *Clim. Change*, 81, 71-95 , doi: 10.1007/s10584-006-9226-z.

Benjamin, S. G., G. A. Grell, J. M. Brown & T. G. Smirnova (2004) Mesoscale weather prediction with RUC hybrid isentropic-terrain-following coordinate model. *Mon. Wea. Rev*, 132, 473-494.

Chou, M.–D. & M. J. Suarez (1999) A solar radiation parameterization for atmospheric studies. *NASA Tech. Memo 104606* **40, Greenbelt, Maryland.**

Chuine, I., P. Yiou, N. Viovy, B. Seguin, V. Daux & E. L. Ladurie (2004) Historical phenology: grape ripening as a past climate indicator. *Nature*, 432, 289-290, doi: 10.1038/432289a.

Collins, W. D., P. J. Rasch, B. A. Boville, J. J. Hack, J. R. McCaa, D. L. Williamson, J. T. Kiehl, B. Brieglib, C. Bitz, S.-J. Lin, M. Zhang & Y. Dai (2004) Description of the NCAR Community Atmosphere Model (CAM 3.0). NCAR Tech. Note NCAR/TN–464+STR. 214 pp.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot,

N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, B. K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thépaut & F. Vitart (2011) The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, 137, 553-597.

Della-Marta, P.M., M. R. Haylock, J. Luterbacher & H. Wanner (2007) Doubled length of western European summer heat waves since 1880. *J. Geophys. Res.*, 112, D15103, doi:10.1029/2007JD008510.

Déqué, M, D. P. Rowell, D. Luthi, F. Giorgi, J. H. Christensen, B. Rockel, D. Jacob, E. Kjellstrom, M. de Castro & B. van den Hurk (2007) An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. *Climatic Change*, 81, 53-70, doi:10.1007/s10584-006-9228-x.

Dole, R., M. Hoerling, J. Perlwitz, J. Eischeid, P. Pegion, T. Zhang, X.-W. Quan, T. Y. Xu & D. Murray (2011) Was there a basis for anticipating the 2010 Russian heat wave? *Geophys. Res. Lett.*, 38, L06702, doi:10.1029/2010GL046582.

Dudhia, J. (1996) A multi-layer soil temperature model for MM5. Sixth Annual PSU/NCAR Mesoscale Model Users' Workshop. Boulder CO, July 1996, 49-50.

Evans, J. P., M. Ekstrom & F. Ji (2012) Evaluating the performance of a WRF physics ensemble over South-East Australia. *Clim. Dyn.*, 39, 1241-1258, doi:10.1007/s00382-011-1244-5.

Fischer, E. M., S. I. Seneviratne, D. Luthi & C. Schär (2007) Contribution of land-atmosphere coupling to recent European summer heat waves. *Geophy. Res. Lett.*, 34, L06707, doi:10.1029/2006GL029068.

Fischer, E. M. & C. Schär (2010) Consistent geographical patterns of changes in high-impact European heat waves. *Nat. Geosci.*, 3, 398-403.

García-Díez, M., J. Fernández, L. Fita & C. Yague (2011) Seasonal dependence of WRF model

biases and sensitivity to PBL schemes over Europe. *Q. J. R. Meteorol. Soc.*, 139, 501-514, doi:10.1002/qj.1976.

García-Díez, M., J. Fernández & R. Vautard (2014) An RCM multi-physics ensemble over Europe : Multi-variable evaluation to avoid error compensation. *Clim. Dyn.*, submitted.

García-herrera, R., J. Diaz, R. M. Trigo, J. Luterbacher & E. M. Fischer (2010) A review of the European summer heat wave of 2003. *Critical Reviews in Environmental Science and Technology*, 40, 267-306, doi: 10.1080/10643380802238137.

Gomez-Navarro, J. J., J. P. Montávez, S. Jerez, P. Jimenez-Guerrero & E. Zorita (2012) What is the role of the observational dataset in the evaluation and scoring of climate models? Geophys. Res. Lett. 39:L24701.

Grell, G. A. & D. Devenyi (2013) A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, 29, 1693, doi:10.1029/2002GL015311.

Grünwald, T. & C. Bernhofer (2007) A decade of carbon, water and energy flux measurements of an old spruce forest at the Anchor Station Tharandt. *Tellus*, 59B, 387–396.

Han, J., & H. Pan (2011) Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, 26, 520–533.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones & M. New (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *J. Geophys. Res.*, 113, D20119, doi: 10.1029/2008JD010201.

Herrera, S., L. Fita, J. Fernández & J. M. Gutierrez (2010) "Evaluation of the mean and extreme precipitation regimes from the ENSEMBLES regional climate multimodel" J. Geophys. Res. 115:D21117.

Hong, S.–Y., & J.–O. J. Lim (2006a) The WRF single–moment 6–class microphysics scheme (WSM6). J. Korean Meteor. Soc., 42, 129–151.

Hong, S.-Y., Y. Noh, J. Dudhia (2006b) A new vertical diffusion package with an explicit treatment

of entrainment processes. Mon. Wea. Rev., 134, 2318–2341.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough & W. D. Collins (2008) Radiative forcing by long–lived greenhouse gases: Calculations with the AER radiative transfer models. J. Geophys. Res., 113, D13103, doi:10.1029/2008JD009944.

IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp, doi:10.1017/CBO9781107415324.

Jacob, D., J. Petersen, B. Eggert, A. Alias, O. B. Christensen, L. M. Bouwer, A. Braun, A. Colette, M. Deque, G. Georgievski, E. Georgopoulou, A. Gobiet, L. Menut, G. Nikulin, A. Haensler, N. Hempelmann, C. Jones, K. Keuler, S. Kovats, N. Kroner, S. Kotlarski, A. Kriegsmann, E. Martin, E. Van Meijgaard, C. Moseley, S. Pfeifer, S. Preuschmann, C. Radermacher, K. Radtke, D. Rechid, M. Rounsevell, P. Samuelsson, S. Somot, J. F. Soussana, C. Teichmann, R. Valentini, R. Vautard, B. Weber & P. Yiou (2014) EURO-CORDEX: new-high-resolution climate change projections for European impact research. *Regional Environmental Change,*14, 563-578.

Janjic, Z. I. (1994) The Step–Mountain Eta Coordinate Model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945.

Janjic, Z. I., (2002) Nonsingular implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso model. *NCEP Office Note No. 437*, 61 pp., National Centers for Environmental Prediction, College Park, MD.

Jankov, I., W. A. Gallus, M. Segal, B. Shaw & S. E. Koch (2005) The impact of different WRF model physical parameterizations and their interactions on warm season WCS rainfall. *Weather and Forecasting*, 20, 1048-1060, doi:10.1175/WAF888.1.

Kain, J. S. (2004) The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, 43, 170–181.

Koster, R. D., S. P. P. Mahanama, T. J. Yamada, G. Balsamo, A. A. Berg, M. Boisserie, P. A. Dirmeyer, F. J. Doblas-Reyes, G. Drewitt, C. T. Gordon, Z. Guo, J. H. Jeong, D. M. Lawrence, W. S. Lee, Z. Li, L. Luo, S. Malyshev, W. J. Merryfield, S. I. Seneviratne, T. Stanelle, B. J. J. M. Van den Hurk, F. Vitart & E. F. Wood (2010) Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, 37, L02402, doi: 10.1029/2009GL041677.

Kotlarski, S., K. Keuler, O. B. Christensen, A. Colette, M. Déqué, A. Gobiet, K. Goergen, D. Jacob, D. Lüthi, E. van Meijgaard, G. Nikulin, C. Schär, C. Teichmann, R. Vautard, K. Warrach-Sagi & V. Wulfmeyer (2014) Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model Dev.*, 7, 1297-1333, doi:10.5194/gmd-7-1297-2014, 2014.

Kuglitsch, F. G., T. Toreti, E. Xoplaki, P. M. Della-marta, C. S. Zerefos, M. Turkes & J. Luterbacher (2010) Heat wave changes in the eastern Mediterranean since 1960. *Geophys. Res. Lett.*, 37, L04802, doi: 10.1029/2009GL041841.

Lenderink, G., A. van Ulden, B. van den Hurk, E. van Meijgaard, 2007, Summertime inter-annual temperature variability in an ensemble of regional model simulations: analysis of the surface energy budget. *Climatic Change*, 81:233–247

Luterbacher, J., S. J. Koenig, J. Franke, G. Van der Schrier, E. Zorita, A. Moberg, J. Jacobeit, P. M. Della-marta, M. Kuttel, E. Xoplaki, D. Wheeler, T. Rutishauser, M. Stossel, H. Wanner, R. Brazdil, P Dobrovolny, D. Camuffo, C. Bertolin, A. Van Engelen, F. J. Gonzalez-Rouco, R. Wilson, C. Pfister, D. Limanowka, O. Nordli, L. Leijonhufvud, J. Soderberg, R. Allan, M. Barriendos, R. Glaser, D. Riemann, Z. Hao & C. S. Zerefos (2010) Circulation dynamics and its influence on European and Mediterranean January-April climate over the past half

millennium: results and insights from instrumental data, documentary evidence and coupled climate models. *Climate change*, 101, 201-234, doi: 10.1007/s10584-009-9782-0.

Meehl, G.A. & C. Tebaldi (2003) More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, 305, 994-997, doi:10.1126/science.1098704.

Miralles, D. G., T. R. H. Holmes, R. A. M. De Jeu, J. H. Gash, A. G. C. A Meesters & A. J. Dolman (2011) Global land-surface evaporation estimated from satellite-based observations. *Hydrol. Earth Syst. Sci.*, 15, 453-469, doi:10.5194/hess-15-453-2011.

Miralles, D. G., A. J. Teuling, C. C. van Heerwaarden & J. Vilà-Guerau de Arellano (2014) Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geosci*., 7, 345-349, doi:10.1038/ngeo2141.

Mooney, P. A., F. J. Mulligan & R. Fealy (2013) "Evaluation of the Sensitivity of the Weather Research and Forecasting Model to Parameterization Schemes for Regional Climates of Europe over the Period 1990–95", J. Clim. 26:1002-1017.

Morrison, H., G. Thompson & V. Tatarskii (2009) Impact of Cloud Microphysics on the Development of Trailing Stratiform Precipitation in a Simulated Squall Line: Comparison of One– and Two–Moment Schemes. *Mon. Wea. Rev.*, **137**, 991–1007.

Nakanishi, M. & H. Niino (2006) An improved Mellor–Yamada level 3 model: its numerical stability and application to a regional prediction of advecting fog. *Bound. Layer Meteor.*, **119**, 397–407.

Nakanishi, M. & H. Niino (2009) Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteor. Soc. Japan*, **87**, 895–912.

Nikulin, G., E. Kjellstrom, U. Hansson, G. Strandberg and A. Ullerstig (2010) Evaluation and future projections of temperature, precipitation and wind extremes over Europe in an ensemble of regional climate simulations. *Tellus A*, 63, 41-55.

Omrani, H., P. Dobrinski, P & T. Dubos (2013) Optimal nudging strategies in regional climate

modelling: investigation in a Big-Brother experiment over the European and Mediterranean regions. *Climate Dynamics*, 41, 2451-2470.

Orlowsky, B. & S. I. Seneviratne (2012) Global changes in extreme events: regional and seasonal dimension. *Climatic Change*, 110, 669-696, doi:10.1007/s10584-011-0122-9.

Pilegaard, K., A. Ibrom, M. S. Courtney, P. Hummerlshøj & N. O. Jensen (2011) Increasing net CO_2 uptake by a Danish beech forest during the period from 1996 to 2009. *Agricultural and Forest Meteorology*, 151, 934-946.

Pleim, J. E. (2007) A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part I: Model Description and Testing. *J. Appl. Meteor. Climatol.*, **46**, 1383–1395.

Schär, C., P. L. Vidale, D. Luthi, C. Frei, C. Haberli, M. A. Liniger & C. Appenzeller (2004) The role of increasing temperature variability in European summer heatwaves. *Nature*, 427**,** 332-336, doi:10.1038/nature02300.

Seneviratne, S. I., D. Luthi, M. Litschi & C. Schär (2006) Land-atmosphere coupling and climate change in Europe. *Nature*, 443, 205-209, doi: 10.1038/nature05095.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky & A. J. Teuling (2010) Investigating soil moisture-climate interactions in a changing climate: a review. *Earth Sci. Rev.*, 99, 125-161.

Seneviratne, S. I., N. Nicholls, D. Easterling, C. M. Goodess, S. Kanae, J. Kossin, Y. Luo, J. Marengo, K. McInnes, M. Rahimi, M. Reichstein, A. Sorteberg, C. Vera, and X. Zhang, 2012: Changes in climate extremes and their impacts on the natural physical environment. In: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* [Field, C. B., V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor, and P. M. Midgley (eds.)]. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC). Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 109-230.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y. Huang, W.

Wang & J. G. Powers (2008) A description of the Advanced Research WRF version 3. NCAR Tech. Note 1–125, http://nldr.library.ucar.edu/collections/TECH-NOTE-000-000-000-855.

Stegehuis, A., R. Vautard, P. Ciais, A. J. Teuling, M. Jung & P. Yiou (2013) Summer temperatures in Europe and land heat fluxes in observation-based data and regional climate model simulations. *Clim. Dyn.*, 41, 455-477, doi:10.1007/s00382-012-1559-x.

Sukoriansky, S., B. Galperin, & V. Perov (2005) Application of a new spectral model of stratified turbulence to the atmospheric boundary layer over sea ice. *Bound.–Layer Meteor.*, **117**, 231–257.

Teuling, A. J., M. Hirschi, A. Ohmura, M. Wild, M. Reichstein, P. Ciais, N. Buchmann, C. Ammann, L. Montagnani, A. D. Richardson, G. Wohlfahrt & S. I. Seneviratne (2009) A regional perspective on trends in continental evaporation. *Geophys. Res. Lett.*, 36, L02404, doi: 10.1029/2008GL036584.

Tewari, M., F. Chen, W. Wang, J. Dudhia, M. A. LeMone, K. Mitchell, M. Ek, G. Gayno, J. Wegiel & R. H. Cuenca (2004) Implementation and verification of the unified NOAH land surface model in the WRF model. *20th conference on weather analysis and forecasting/16th conference on numerical weather prediction*, pp. 11–15. Seattle, WA, American Meteorological Society.

Thompson, G., P. R. Field, R. M. Rasmussen & W. D. Hall (2008) Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115.

Tiedtke, M., (1989) A comprehensive mass flux scheme for cumulus parameterization in large–scale models. *Mon. Wea. Rev.*, 117, 1779–1800.

Tingley, M. P. & P. Huybers (2013) Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature*, 496, 201-205, doi: 10.1038/nature11969.

Van den Hurk, B., F. Doblas-Reyes, G. Balsamo, R. D. Koster, S. I. Seneviratne & H. Camargo (2012) Soil moisture effects on seasonal temperature and precipitation forecast scores in

Europe. *Clim. Dyn.*, 38, 349-362, doi: 10.1007/s00382-010-0956-2.

Vautard, R., A. Gobiet, D. Jacob, M. Belda, A. Colette, M. Deque, J. Fernandez, M. García-Díez, K. Goergen, I. Guttler, T. Halenka, T. Karacostas, E. Katragkou, K. Keuler, S. Kotlarski, S. Mayer, E. Van Meijgaard, G. Nikulin, M. Patarcic, J. Scinocca, S. Sobolowski, M. Suklitsch, C. Teichmann, K. Warrach-Sagi, V. Wulfmeyer & P. Yiou (2013) The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. *Clim. Dyn.*, 41, 2555-2575, doi:10.1007/s00382-013-1714-z.

Vautard, R., F. Thias, I. Tobin, F.-M. Breon, J.-G. Devezeaux de Lavergne, A. Colette, P. Yiou & P. M. Ruti (2014) Regional climate model simulations indicate limited climatic impacts by operational and planned European wind farms. *Nature Communications*, 3196, doi:10.1038/ncomms4196.

Warrach-Sagi, K., T. Schwitalla, V. Wulfmeyer & H. S. Bauer (2013) Evaluation of a climate simulation in Europe based on the WRF-NOAH model system: precipitation in Germany. *Clim. Dyn.*, 41, 755-774, doi:10.1007/s00382-013-1727-7.

Weisheimer, A., F. J. Doblas-Reyes, T. Jung & T. N. Palmer (2011) On the predictability of the extreme summer 2003 over Europe. *Geophys. Res. Lett.*, 38, L05704, doi: 10.1029/2010GL046455.

Wild, M., B. Trussel, A. Ohmura, C. N. Long, G. Konig-Langlo, E. G. Dutton & A. Tsvetkov (2009) Global dimming and brightening: An update beyond 2000. *J. Geophys. Res.*, 114, D00D13, doi: 10.1029/2008JD011382.

Wohlfahrt, G., S. Pilloni, L. Hörtnagl & A. Hamerle (2010) Estimating carbon dioxide fluxes from temperature mountain grasslands using broad-band vegetation indices. *Biogeosciences*, 7, 683-694.

Zhang, C., Y. Wang & K. Hamilton (2011) Improved representation of boundary layer clouds over the southeast pacific in ARW–WRF using a modified Tiedtke cumulus parameterization

scheme. *Mon. Wea. Rev.*, 139, 3489–3513.

Table and figure captions

Table 1. Physics schemes used in this study (with references). All possible permutations are made, yielding a total of 216 simulations. The numbers in the table refer to the number the schemes have in the Weather Research and Forecasting (WRF) model.

Table 2. The five best performing combinations of physics in ranked from the first to the fifth best.

Table 3. Cross-comparison between France 2003 and Russia 2010. The (5, 10, 15, 20 and 25) best simulations, when only using one heat wave to select the best configurations and vice versa, are taken and compared with their ranking for the other heat wave. If there would be no correlation between the two years, the average ranking would lay approximately at half of the total number of simulations for both years that lay within a first selection of 1K (column 8). In bold the rankings that are lower than this number. Because observations of radiation are lacking over Russia, we tested France with and without including radiation in the ranking.

Figure 1. Time series of daily land heat fluxes in 2003 from May to the end of August on three different FLUXNET sites, with latent heat flux (LH) on the first row, sensible heat flux (SH) on the second row, and evaporative fraction (EF – latent heat flux divided by the sum of latent and sensible heat flux) on the last row. The three columns represent three sites, with Neustift/Stubai (Austria – ATneu 47N, 11E) in the first column, Tharandt (Germany – DETha, 51N, 4E) in the second, and Soroe-LilleBogeskov (Denmark – DKsor, 66N, 11E) in the third column. Vegetation types on the three sites are respectively grassland (GRA), evergreen needleleaf forest (ENF), and deciduous broadleaf forest (DBF). In grey all 216 simulations with the NOAH scheme. Observational data is shown in black (FLUXNET). The solid light blue line is one configuration with NOAH, while the blue dots represent the same configuration but with RUC instead of NOAH.

Figure 2. Domains used in this study: France, Iberian Peninsula, Russia and Scandinavia.

Figure 3. Time series of daily mean temperature over France in 2003 (a) and 2007 (b) and Russia in 2010 (c). Every simulation is shown in gray and observations of E-OBS in black. The blue and red lines are the coldest and the warmest simulations over France during the heat wave. These lines have the same set of physics in all the figures (3, 4, 5). Figure d shows the simulated temperature min-max range during the heatwave of 2003 (1-15 August). The range is calculated as the difference between the warmest and the coldest simulation during the heat wave period between the 216 members of the ensemble.

Figure 4. Monthly precipitation over France in 2003 (a) and 2007 (b) and Russia 2010 (c). The boxplots show the extremes, 25^{th}, 50^{th}, and 75^{th} percentiles. The blue and red dots are the coldest and the warmest simulations over France during the heat wave (as in figure 3).

Figure 5. Scatter plot of soil moisture content at July 31, and temperature in August. Every point is one simulation. Different colors and symbols represent different physics for convection (CU) (a), microphysics (MP) (b), radiation (RA) (c) and planetary boundary layer-surface (PBL-SF) (d).

Figure 6. Monthly radiation over France in 2003 (a) and 2007 (b); no radiation data being available in Russia for 2010. The boxplots show the extremes, 25^{th}, 50^{th}, and 75^{th} percentiles. The blue and red dots are the coldest and the warmest simulations over France during the heat wave (as in figure 3).

Figure 7. Daily time series of temperature (a) and latent heat flux (c); monthly time series of precipitation (b) and incoming shortwave radiation (d). Observations are shown in black, and the

five best performing runs in colors. Gray lines indicate other simulations. All figures are a spatial average over France during summer 2003.

Figure 8. Sensitivity test of the initialization of soil moisture. Difference between the ‘control’ simulation and the perturbed ones (minus (red) and plus (blue) 20% initial soil moisture) of the five highest ranked configurations. The darkest lines are the best simulations (1), and descending colour shade agrees with descending ranking (1-5).

Table 1

Microphysics (MP)	PBL+Surface (PBL-SF)	Radiation (RA)	Convection (CU)	Soil
6) WRF-SM6 (Hong et al. 2006a)	1-1) Yonsei Uni-MM5 (Hong et al. 2006b; Beljaars, 1994)	3) CAM (Collins et al. 2004)	1) Kain-Fritsch (Kain 2004)	2) NOAH (Tewari et al. 2004)
8) New Thompson (Thompson et al. 2008)	2-2) MYJ-ETA (Janjic et al. 1994; Janjic, 2002)	4) RRTMG (Iacono et al. 2008)	3) Grell-Devenyi (Grell & Devenyi, 2012)	
10) Morrison DM (Morrison et al. 2009)	4-4) QNSE-QNSE (Sukoriansky et al. 2005)	5) Goddard (Chou & Suarez, 1999)	6) Tiedtke (Tiedtke 1989; Zhang et al. 2011)	
	5-2) MYNN-ETA (Nakanishi & Niino, 2006, 2009; Janjic, 2002)		14) New SAS (Han & Pan, 2011)	
	5-5) MYNN-MYNN (Nakanishi & Niino, 2006, 2009)			
	7-1) ACM2-MM5 (Pleim 2007;			

	Beljaars, 1994)			

Table 2

Microphysics	PBL-Surface	Radiation	Convection	Soil	Rank
Morrison DM	Yonsei Uni-MM5	RRTMG	Tiedtke	NOAH	1
WRF-SM6	MYNN-MYNN	RRTMG	Grell-Devenyi	NOAH	2
WRF-SM6	ACM2-MM5	Goddard	Tiedtke	NOAH	3
New Thompson	MYNN-MYNN	RRTMG	New SAS	NOAH	4
New Thompson	ACM2-MM5	RRTMG	Tiedtke	NOAH	5

Table 3

		Average ranking of 5, 10, 15, 20 and 25 best simulations					
		5	10	15	20	25	Number of simulations within 1ºC
With radiation	Average rank Fr-Ru	**22.6**	**21.8**	**25.3**	**23.1**	**27.5**	104
With radiation	Average rank Ru-Fr	**15.75**	**15.2**	**14.7**	**13**	39.3	58
Without radiation	Average rank Fr-Ru	53	**37**	**28.4**	**27.6**	**25.5**	104
Without radiation	Average rank Ru-Fr	**20.25**	**16.8**	**18.1**	**17**	**19.9**	58

Figure 1

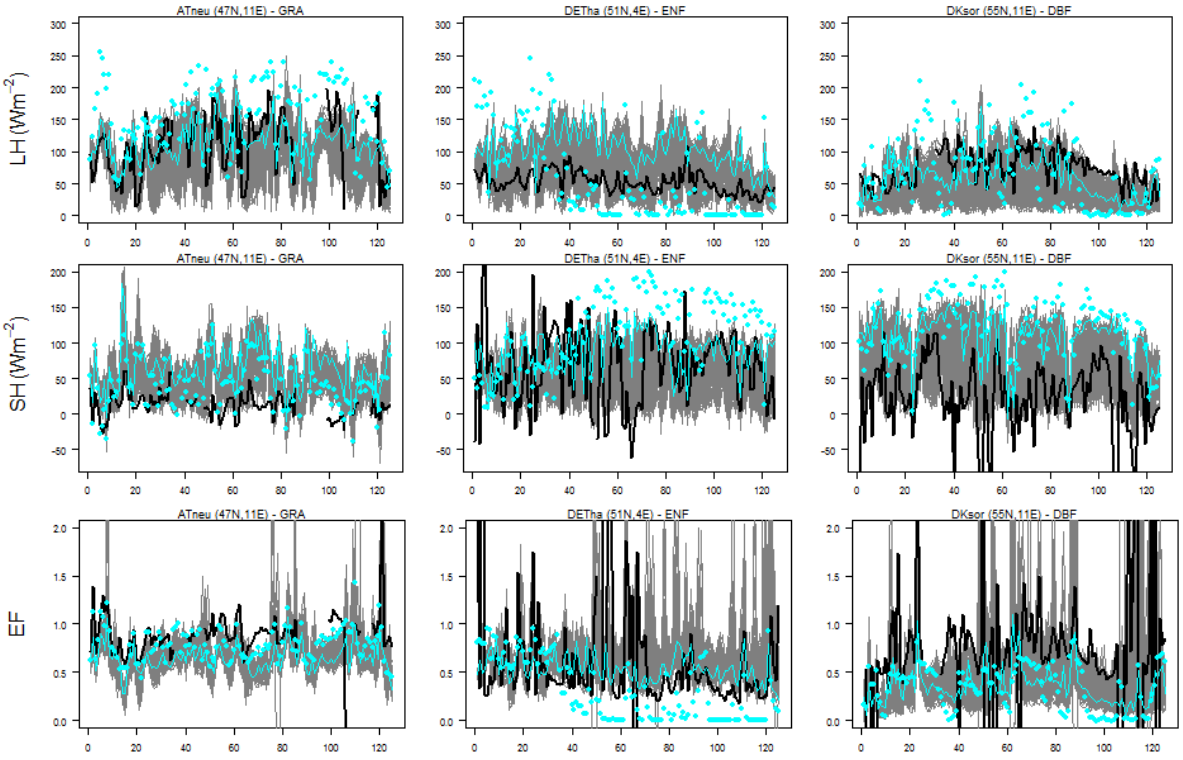

ATneu (47N,11E) - GRA
DETha (51N,4E) - ENF
DKsor (55N,11E) - DBF
LH (Wm^{-2})
SH (Wm^{-2})
EF

Figure 2

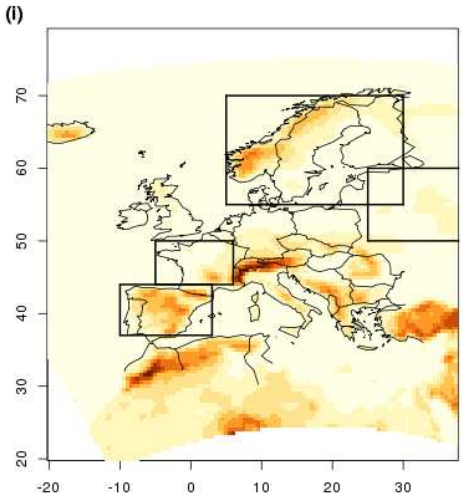


Figure 3

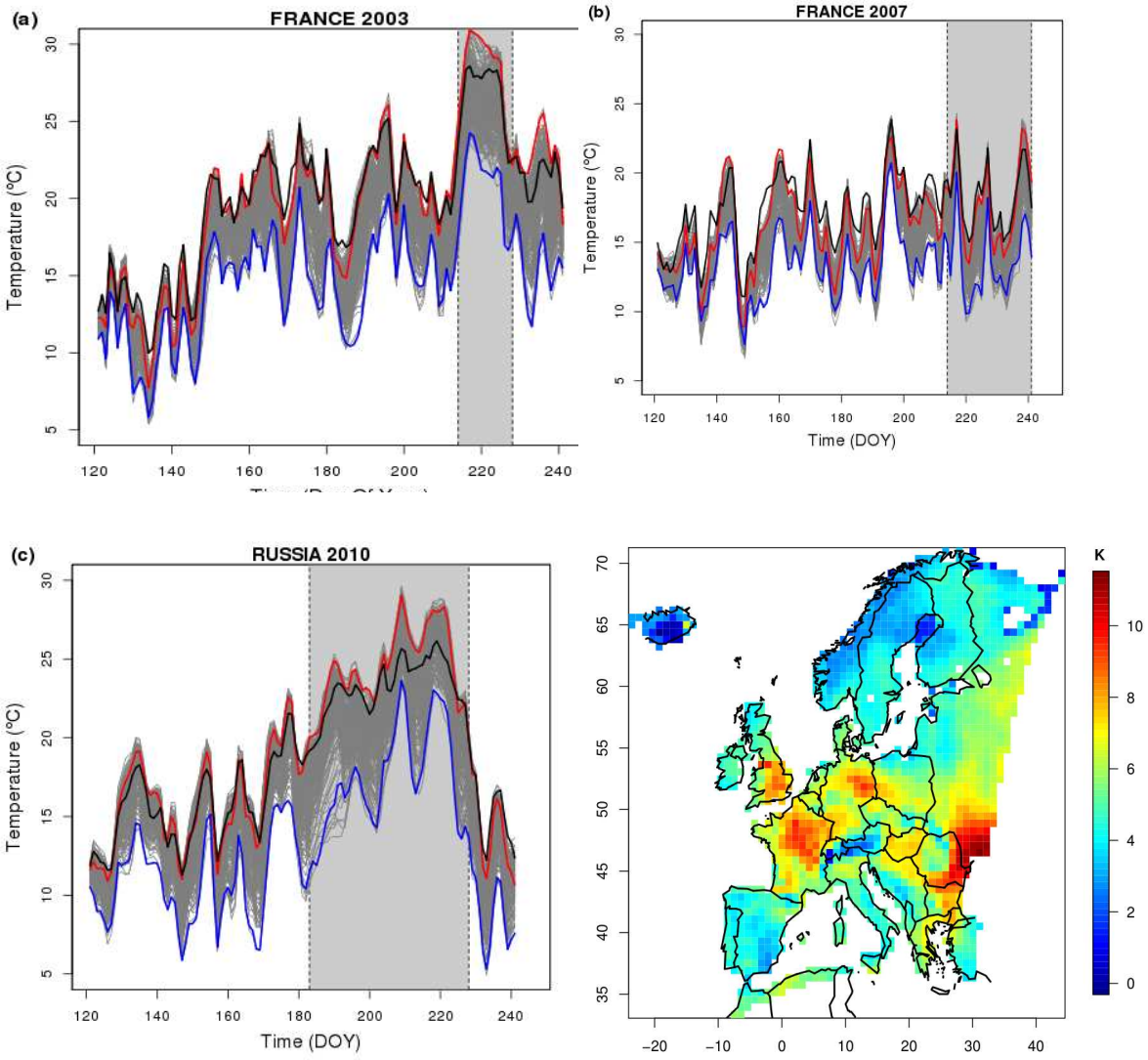

(a)
FRANCE 2003
Temperature (°C)
(b)
FRANCE 2007
Temperature (°C)
Time (DOY)
(c)
RUSSIA 2010
Temperature (°C)
Time (DOY)
K

Figure 4a-c

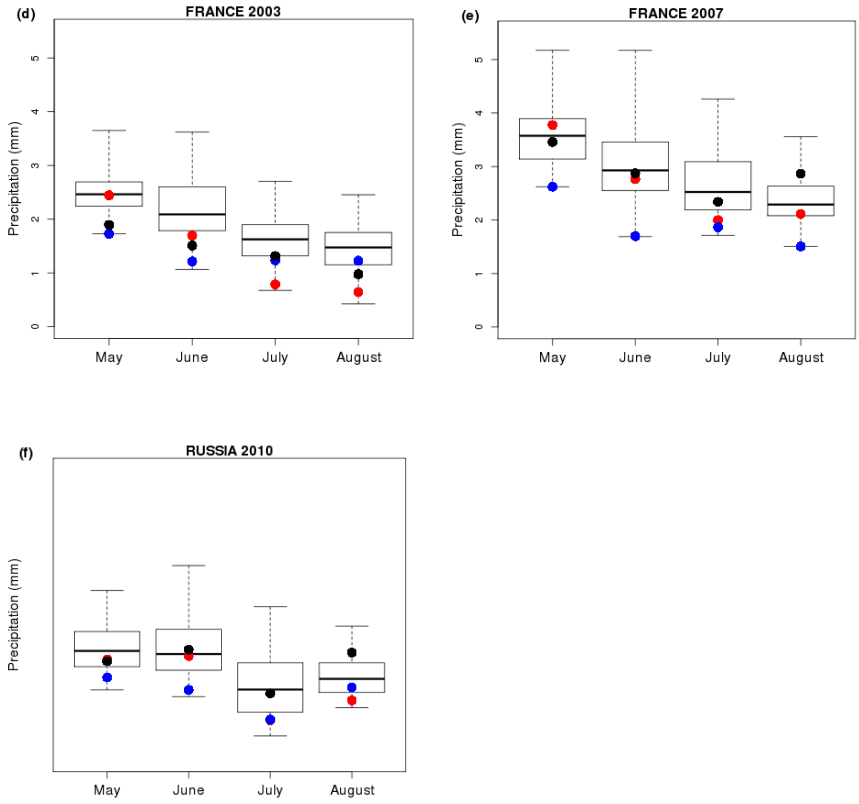

(d)
FRANCE 2003
Precipitation (mm)
0
1
2
3
4
5
May
June
July
August
(e)
FRANCE 2007
Precipitation (mm)
0
1
2
3
4
5
May
June
July
August
(f)
RUSSIA 2010
Precipitation (mm)
May
June
July
August

Figure 5a-d

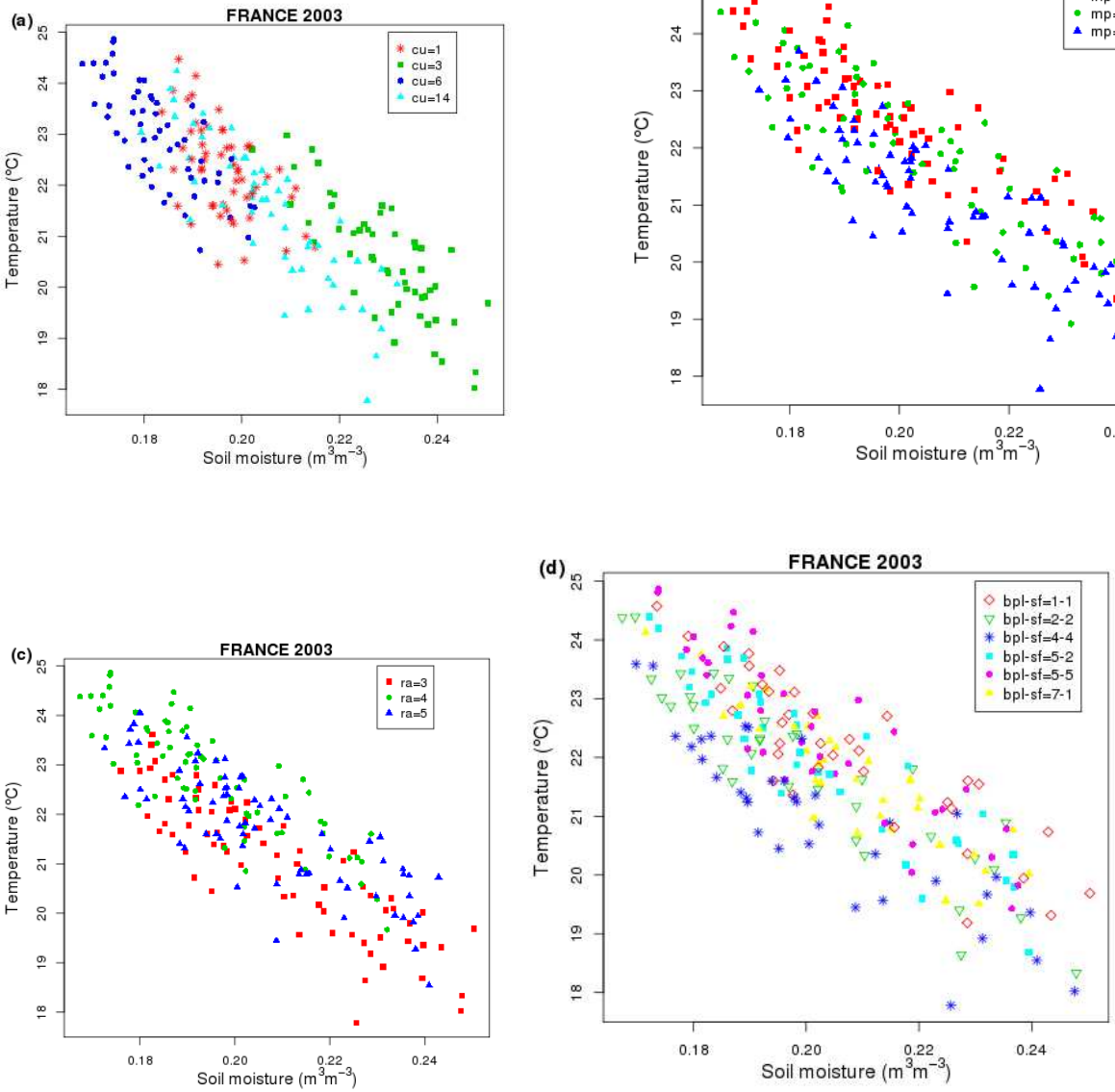

(a)
FRANCE 2003
cu=1
cu=3
cu=6
cu=14
Temperature (°C)
Soil moisture (m3m−3)
(b)
FRANCE 2003
mp=6
mp=8
mp=10
Temperature (°C)
Soil moisture (m3m−3)
(c)
FRANCE 2003
ra=3
ra=4
ra=5
Temperature (°C)
Soil moisture (m3m−3)
(d)
FRANCE 2003
bpl-sf=1-1
bpl-sf=2-2
bpl-sf=4-4
bpl-sf=5-2
bpl-sf=5-5
bpl-sf=7-1
Temperature (°C)
Soil moisture (m3m−3)

Figure 6a-b

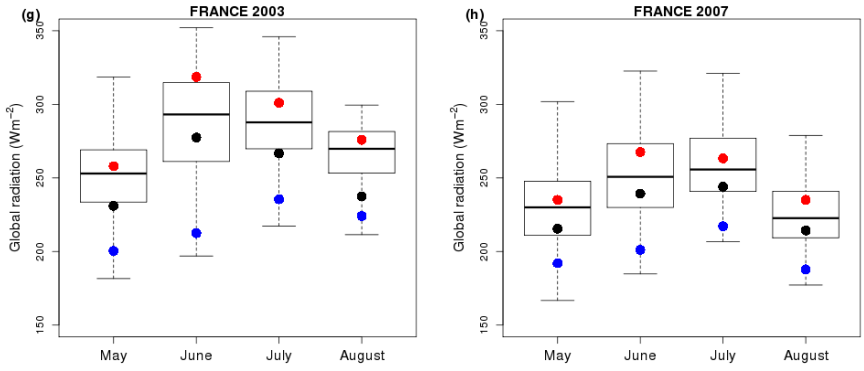

(g)
FRANCE 2003
Global radiation (Wm⁻²)
150
200
250
300
350
May
June
July
August
(h)
FRANCE 2007
Global radiation (Wm⁻²)
150
200
250
300
350
May
June
July
August

Figure 7a-d

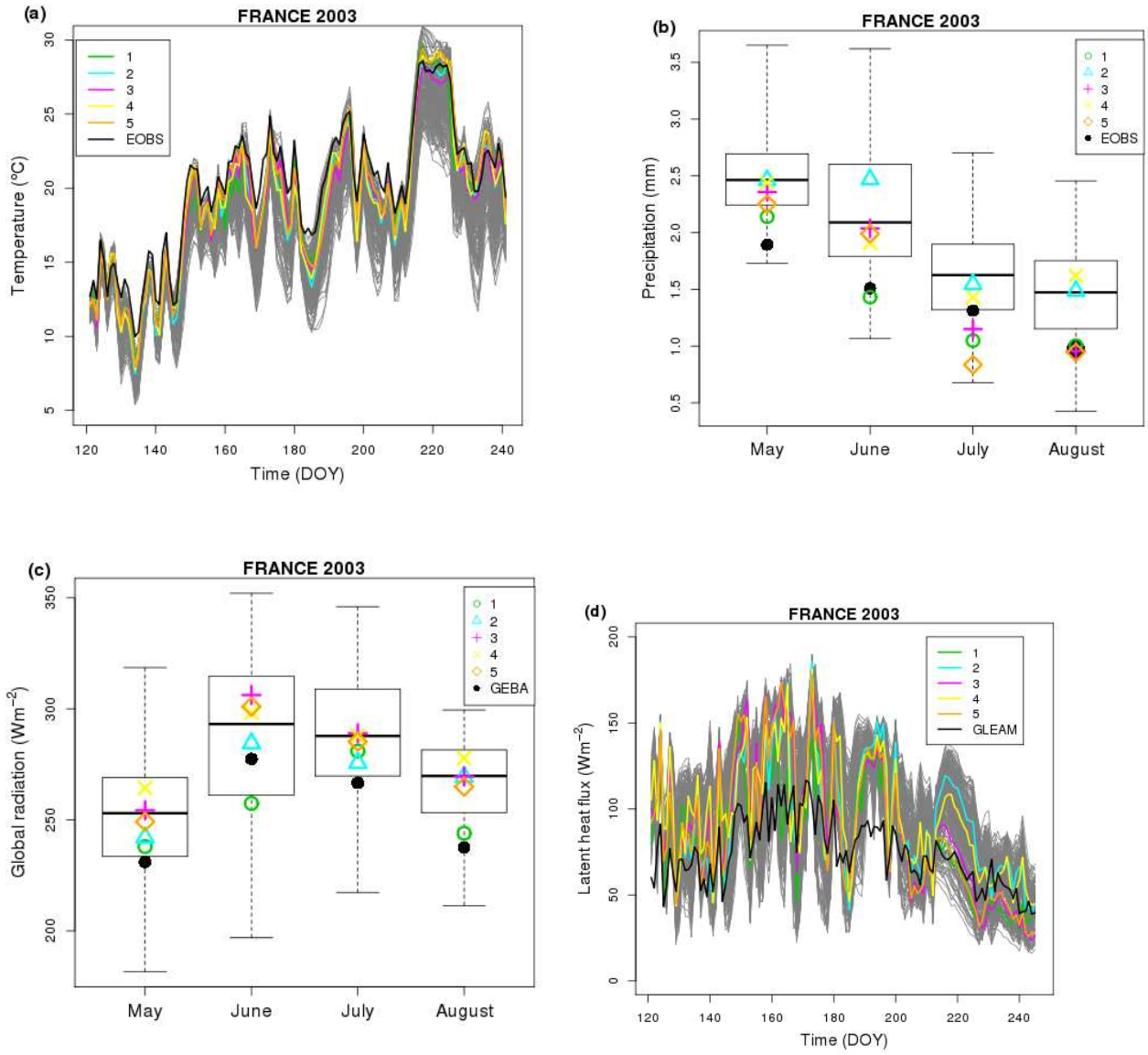
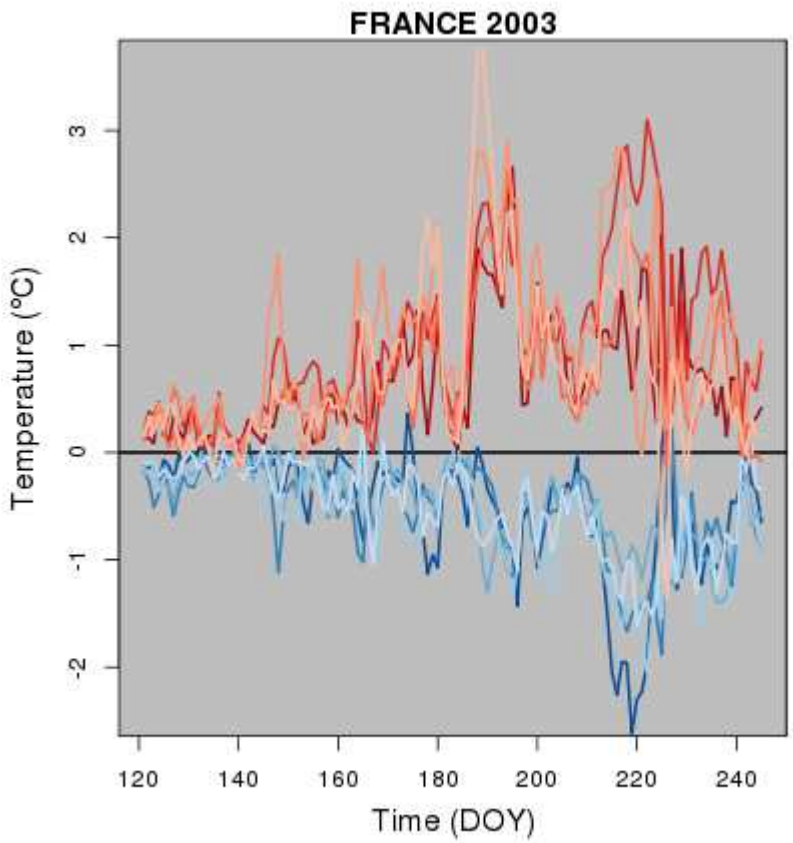

(a)
FRANCE 2003
Temperature (°C)
Time (DOY)
1
2
3
4
5
EOBS
(b)
FRANCE 2003
Precipitation (mm)
May
June
July
August
1
2
3
4
5
EOBS
(c)
FRANCE 2003
Global radiation (Wm^{-2})
May
June
July
August
1
2
3
4
5
GEBA
(d)
FRANCE 2003
Latent heat flux (Wm^{-2})
Time (DOY)
1
2
3
4
5
GLEAM

Figure 8


FRANCE 2003
Temperature (°C)
3
2
1
0
-1
-2
120
140
160
180
200
220
240
Time (DOY)