The ExtremeX global climate model experiment: Investigating thermodynamic and dynamic processes contributing to weather and climate extremes

By Kathrin Wehrli, Fei Luo, Mathias Hauser, Hideo Shiogama, Daisuke Tokuda, Hyungjun Kim, Dim Coumou, Wilhelm May, Philippe Le Sager, Frank Selten, Olivia Martius, Robert Vautard, and Sonia I. Seneviratne

The reviewer's comments were included in the manuscript. Main changes include that all figures were redone, paying attention to a clearer presentation including longitude/latitude lines for map plots. Figure 1 now contains two subplots with a new plot showing the relaxation time scale of the nudging. A schematic showing the disentangling approaches was added in Figure 2. Glaciated areas were masked out in the plots showing evapotranspiration (Figure A5 in the revised manuscript) and for the plots showing the disentangling of the soil moisture vs. atmospheric circulation contribution to warm spells (Figure 8 in the revised manuscript). Also, a figure showing climatological biases for precipitation using MSWEP as reference was added (Figure A6 in the revised manuscript) to confirm the robustness of the results.
The terms "prescription" or "prescribe" are now used consistently for the constraining of soil moisture and "nudging" or "nudged" for the constraining of the atmospheric circulation. When referring to both methods the term "constrain" is used. This was adapted throughout the manuscript and also in Table 1. The number of simulation runs per experiment and model is now given in Table 1 and clearly stated in the writing.
The introduction and purpose of the study were expanded and more background on the predecessor papers was included. As suggested, the discussion of existing literature on the chosen heatwaves was also expanded. Further, the term U.S. heatwave 2012 was introduced instead of Midwestern heatwave 2012, as suggested by one of the reviewers. In the concluding remarks the role of the ocean and the suitability of the ExtremeX simulations in this context is discussed.
Further changes to the manuscript include small improvements to the wording, readability and elimination of typographical errors.

The data accompanying the paper is now available from:
https://data.iac.ethz.ch/Wehrli_et_al_2022_ExtremeX/

Please find below the point-by-point response to the reviewer comments

Reply to reviewer 1 (Paul Dirmeyer)

The modeling experiments described here shed light on the various roles of land versus atmosphere in extremes, going a step or two beyond what was done in the 1990s and 2000s in the "Koster style" studies of those days. It is interesting, adds to our scientific knowledge of climate variability, and should be published after revision. I do not wish to remain anonymous - Paul Dirmeyer

We thank Paul Dirmeyer for his thoughtful and positive evaluation of the manuscript and the helpful suggestions. In the following we will give answers to the comments in blue.

General comments:

1. Realizing this may be difficult without redesigning and rerunning the simulations, but I long to see a bit more separation in the various drivers, e.g., in the atmospheric component, could the roles of dynamics (circulation) versus physics (radiation, clouds, precipitation) be separated? At the land surface, could drivers acting through the energy balance terms versus water balance be quantified separately? Others have delved more into the process level (e.g., https://doi.org/10.1029/2012GL053703), and having models in hand for sensitivity studies enables many possibilities. Likely for "future work", but I wanted to bring up this question.
2. I greatly appreciate the message of the paper regarding the role of compensating errors and tuning. There remains among many in the model development community a strange hope that "fixing" one component of an Earth system model (e.g., upgrading the LSM) will somehow solve other problems. But often it just serves to expose those problems even more as the balance of errors has been disturbed. This paper also shines a bit more light on this issue.
3. Mainly in §5.1 but also conclusions: The conventional wisdom is that persistent anomalies in the atmospheric general circulation (which may have various causes themselves) establish conditions for heat waves and/or droughts, and then land-atmosphere feedbacks can exacerbate or prolong them. Is there any way to diagnose (confront or confirm) this idea from these experiments? Can the role of climate change on this evolutionary sequence be investigated here? These analyses are co-temporal and do not seem to account for the evolution over time of heat wave events, although you do consider persistence. It seems the two "approaches" (A) and (B) get at this somewhat (e.g. L343-344) but it is somewhat elusive.
4. There are a couple of recent papers that are quite germane to ExtremeX, particularly the notion that heat waves have a mix of land and atmosphere (which may ultimately be traced to remote ocean) drivers: https://doi.org/10.1029/2020AV000283, https://doi.org/10.1002/asl.948.

Thank you for the thoughts and ideas for the manuscript. We also appreciate the reference for the two additional papers in the fourth comment. We will mention these relevant and very recent results in our introduction. Regarding a further separation of drivers this is certainly a very relevant question and we agree that it would be great to have future studies going in this direction. We agree with the reviewer that this would require a new experiment design and the additional simulations would be pout of scope for the present work. Regarding a separation of the processes at the land surface we could think of experiments similar to those in Teng et al. (2019), where heating anomalies (from a dry simulation) are imposed in the atmospheric model.

As mentioned in the third comment the experiments are co-temporal. Hence, the ExtremeX setup is likely not ideal to confront or confirm whether circulation anomalies establish conditions for heatwaves or droughts and then land-surface feedbacks kick in by prolonging the events. Studies like Teng et al. (2019) and Martius et al. (2021) have shown that soil moisture anomalies can excite atmospheric circulation anomalies impacting the weather in other regions of the globe. Having experiments where the constraining of the soil moisture (or atmosphere) is confined to a certain period of time (and region) helps to isolate the processes and reduces other interactions. However, the soil moisture anomalies applied in the two studies mentioned would have to be created first which would likely be due to circulation anomalies. Going more deeply into this question would likely require dedicated case studies. Further, we think that the influence of climate change on the development of heatwaves/ droughts may be better investigated in fully coupled model simulations, potentially with a large ensemble to capture a sufficient number of events.

Reference:

Teng, H. Y., G. Branstator, A. B. Tawfik, and P. Callaghan, 2019: Circumglobal response to prescribed soil moisture over North America. J. Climate, 32, 4525–4545, https://doi.org/10.1175/JCLI-D-18-0823.1.

Martius, O., Wehrli, K., & Rohrer, M. (2021). Local and Remote Atmospheric Responses to Soil Moisture Anomalies in Australia, *Journal of Climate*, *34*(22), 9115-9131. Retrieved Dec 6, 2021, from https://journals.ametsoc.org/view/journals/clim/34/22/JCLI-D-21-0130.1.xml

Specific comments:

L75: Technical point: an ensemble of one is not an ensemble. It is just a single run.

This will be corrected.

Fig 1: It would be more clear to replot with the X-axis in a time dimension, e.g., label it as the e-folding (relaxation) time scale.

We think that for the manuscript it is more illustrative to keep the plot with the nudging intensity on the x-axis. However, we will add the formula used to compute the relaxation term to clear things up.

L111: Change "allows to isolate" to "allows isolation of".

This will be corrected.

L124-125: Which models nudged and which replace soil moisture states? And for those that nudged, what was the relaxation time scale?

Thank you for the question. It turned out that it was a misunderstanding among the modeling groups that in MIROC a soil moisture nudging was used. In fact, all models replace, hence prescribe, the simulated soil moisture. We will correct this throughout the manuscript, which will also simplify the terminology used.

L131: I think there was more than one version (combination of inputs) for the LandFlux-Eval data set for ET - which was used?

The mean from the merged ET synthesis product was used (hence their diagnostic, reanalysis and land surface model-based data sets). We will add this information to the description of the reference data sets.

§2.4: This would benefit from a schematic. Could you reproduce or recreate a figure based on Fig 1 of Wehrli et al. 2019? It would be very helpful. And doesn't differences in the results from approaches (A) and (B) shed light on the nonlinearities in the responses (evidence of feedbacks)?

A simplified figure based on the one in the 2019 paper will be added. Indeed, differences in the results following the two approaches show nonlinearities in the responses. We will add a sentence to mention this. The results from the two approaches were found to be qualitatively similar therefore we will not explore the differences.

L285-288: To this list should be added "unrepresented processes" in models, particularly those unresolved due to grid scale: non-hydrostatic atmospheric processes in coarse resolution models, unresolved mesoscale circulations, sub-grid surface heterogeneity.

We thank the reviewer for this suggestion and we will amend the list with processes unrepresented in models.

L288-289: Atmospheric modelers in particular are fixated on 500 hPa geopotential height errors as a metric of circulation fidelity.

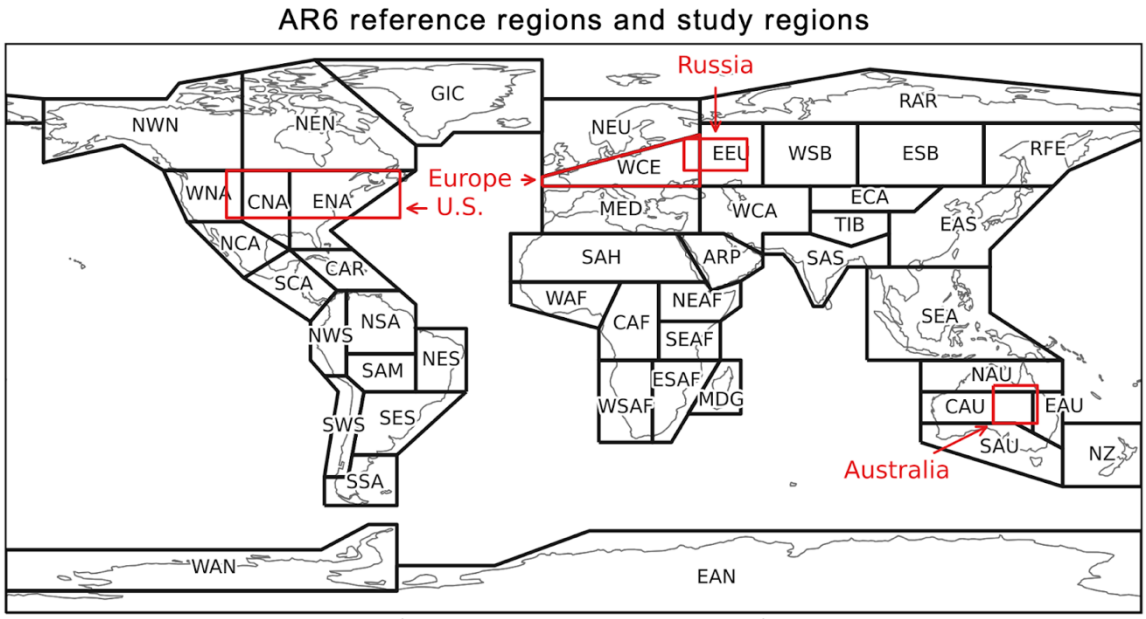
The predecessor papers Wehrli et al., 2018 and 2019 looked into the 500 hPa geopotential height for nudged CESM simulations. We will look into this also for the other models and describe the results for all three models in the manuscript.

Figs 5, 6 and associated text in §5.1: "Midwest" as a region name does not sit well in the global context, as it is a subregion of the U.S. In the other three cases, "Russia", "Europe" and "Australia" do not designate those entire areas, but a portion within each. Thus, "Midwest" should be replaced with "U.S."

The name of the region will be changed according to the reviewer's suggestion. For the events analysed in section 5.1 we will add that the U.S. heatwave was "also known as the Midwest heatwave" since the names of the events were chosen to match with existing literature.

For Russia (line 315) and the U.S. (line 336), how do these areas overlap or intersect the AR6 designated areas? Neither Fig A2 nor any of the other map plots in this manuscript show latitudes and longitudes, so it is difficult to compare by eye.

Figure A2 from the manuscript will be replaced by Figure R1 shown here. We will also update the figures to show longitude and latitudes where possible. Some figures already show rather small maps and we will make sure to not lose information or readability of the figures.



Abbrev.	Name	Abbrev.	Name	Abbrev.	Name
GIC	Greenland/Iceland	NEU	N.Europe	RFE	Russian-Far-East
NWN	N.W.North-America	WCE	West&Central-Europe	WCA	W.C.Asia
NEN	N.E.North-America	EEU	E.Europe	ECA	E.C.Asia
WNA	W.North-America	MED	Mediterranean	TIB	Tibetan-Plateau
CNA	C.North-America	SAH	Sahara	EAS	E.Asia
ENA	E.North-America	WAF	Western-Africa	ARP	Arabian-Peninsula
NCA	N.Central-America	CAF	Central-Africa	SAS	S.Asia
SCA	S.Central-America	NEAF	N.Eastern-Africa	SEA	S.E.Asia
CAR	Caribbean	SEAF	S.Eastern-Africa	NAU	N.Australia
NWS	N.W.South-America	WSAF	W.Southern-Africa	CAU	C.Australia
NSA	N.South-America	ESAF	E.Southern-Africa	EAU	E.Australia
NES	N.E.South-America	MDG	Madagascar	SAU	S.Australia
SAM	South-American-Monsoon	RAR	Russian-Arctic	NZ	New-Zealand
SWS	S.W.South-America	WSB	W.Siberia	EAN	E.Antarctica
SES	S.E.South-America	ESB	E.Siberia	WAN	W.Antarctica
SSA	S.South-America				

Figure R1: Reference regions of the IPCC AR6 as defined in Iturbide et al. (2020). Red outlines show the study regions considered in Section 5.1.

L353: You discuss results from MIROC, but what about the other two models?

The individual rations for MIROC were mentioned to highlight that both approaches lead to the conclusion that SM dominates over the atmospheric circulation contribution. However, we understand that it is confusing why the individual ratios for the other models are not mentioned and we will revise this paragraph.

Fig 7: There seems to a growing proportion of contribution from soil moisture as the anomaly periods grow longer (which would be reasonable, as locally soil moisture represents a slower manifold, a redder spectrum than tropospheric variables). It appears this could be easily quantified. Showing the area-weighted average of the metric in the figure (e.g., the SM-dominant percentage, averaged over unmasked areas only) in each panel would show a growing value with warm spell duration in each model, showing the growing relative importance of the land surface states for long-duration events (which would get at the "conventional wisdom" point above, to some degree).

We thank the reviewer for this suggestion and we will update the figure and description accordingly.

L419-420: Is this true? The atmospheric nudging is very weak in the lower troposphere, and other studies have shown the effect of land surface anomalies on the atmosphere is largely constrained to the boundary layer (e.g., https://doi.org/10.1175/1525-7541(2001)002%3C0329:AEOTSO%3E2.0.CO;2) except over elevated terrain where heating anomalies from the land surface can get into the upper troposphere directly (https://doi.org/10.5194/gmd-14-4465-2021).

As the reviewer mentioned, the effect of land surface anomalies on the atmosphere in general is local and constrained to the boundary layer. We will rephrase the lines in the manuscript to make clear that the present study does not disagree with this statement. In the second study mentioned nudging of the horizontal wind was used to initialize the model before perturbing the land surface temperature. In the present study horizontal wind is nudged for every model time step during the whole simulation period. Indeed, we found that there is only negligible variability between ensemble members due to the setup of the atmospheric nudging. This is not only true for horizontal winds in the free atmosphere as shown for CESM in Wehrli et al. (2018) but also for land surface conditions. We illustrate this in Figure R2 by showing the daily maximum temperature anomaly compared to the 1982-2008 climatology for the Russian heatwave as in Figure 5a) but for the AF_SI experiment instead of AF_SF.

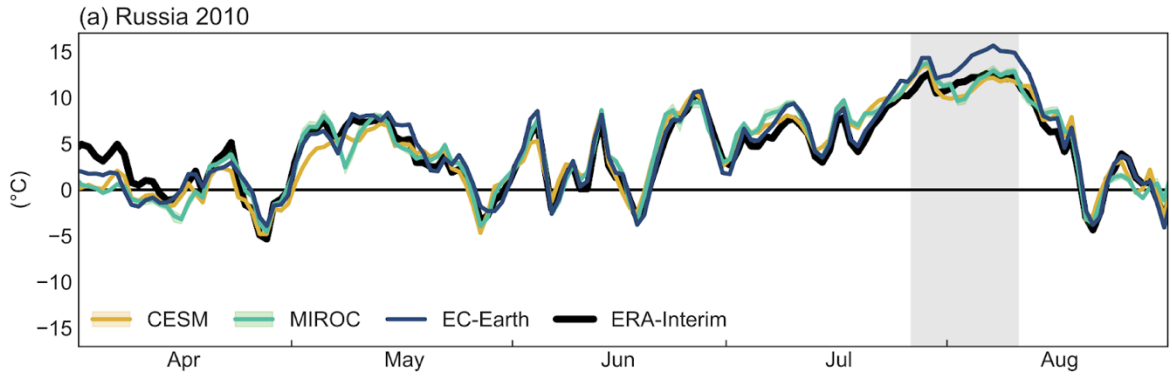


Figure R2: Daily maximum temperature anomaly compared to the 1982–2008 climatology for the nudging (AF_SI) experiment for the three models and for ERA-Interim (black line). The 15-day event period is highlighted in light grey. The shading shows the full ensemble spread and lines the ensemble mean (or single simulation for EC-Earth). The thick black line shows the values from ERA-Interim.

For CESM and MIROC five members for AF_SI were available and for EC-Earth only one. Figure R2 shows that the variability between the ensemble members is very small for both CESM and MIROC. For April to August (time period shown) daily standard deviation between ensemble members varies from 0.02°C to 0.19°C (0.07°C averaged over the whole time period) for CESM and 0.07°C to 0.47°C (0.22°C on average) for MIROC. The AF_SI experiment also captures the temporal evolution of TX anomaly similarly well as AF_SF. In Figure R3 the daily maximum temperature anomaly for all experiments is shown. For CESM (top) the AF_SI and AF_SF experiments barely differ while they do for MIROC (middle) and EC-Earth (bottom). This agrees with the findings from Figure 2 in the manuscript that AF_SI and AF_SF for CESM have a very similar climatology (and hence RMSE), which is not found for the other two models. This is due to the differences in how soil moisture was prescribed in the models and we will discuss this point in the manuscript.

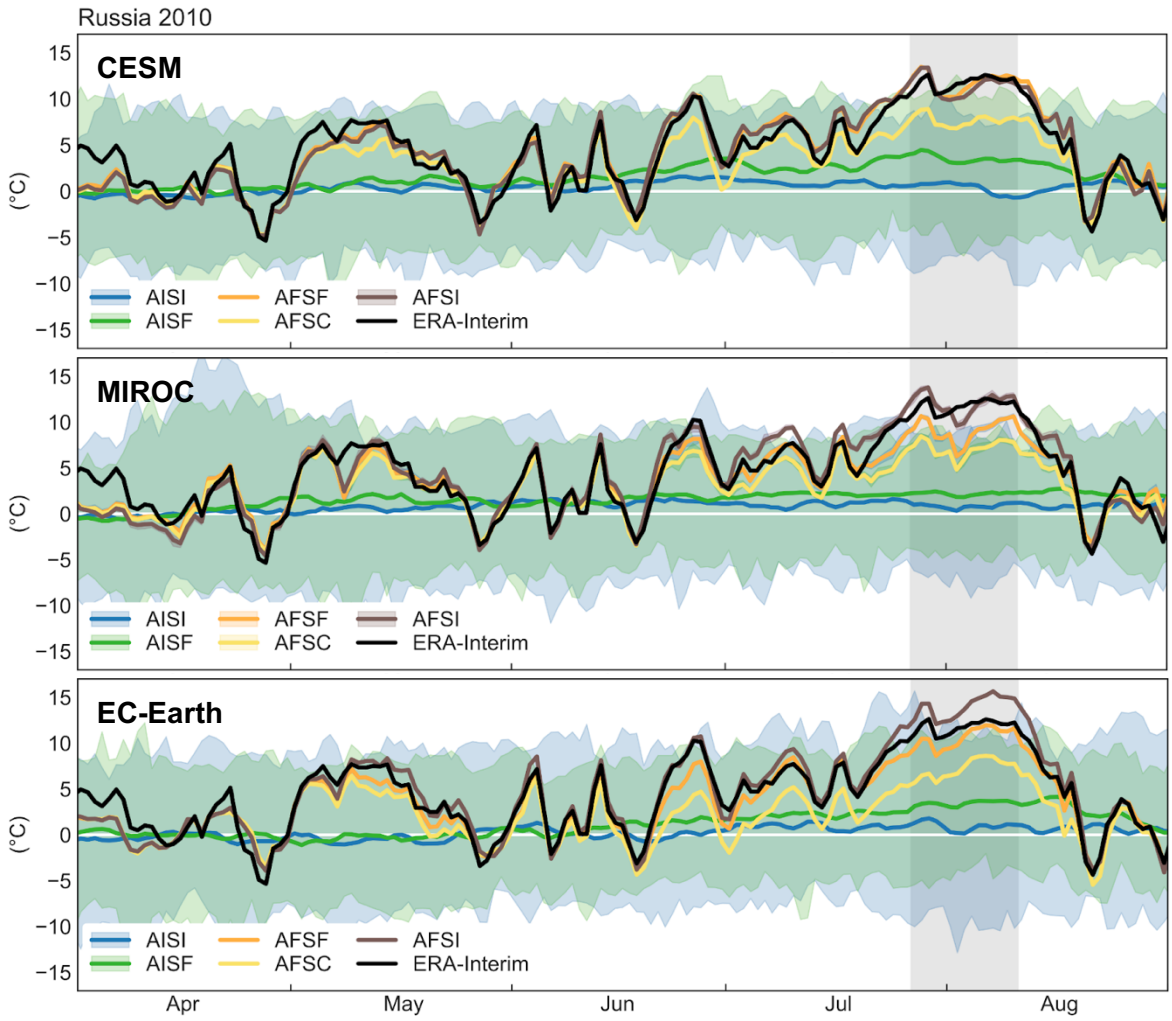


Figure R3: Daily maximum temperature anomaly compared to the 1982–2008 climatology for all experiments per model. Shown is the year 2010 and values are averaged for the region considered for the Russian heatwave. From top down: CESM, MIROC, EC-Earth. The shading shows the full ensemble spread and lines the ensemble mean (or single simulation). The black line shows the values from ERA-Interim.

All map figures: Since soil moisture as a climate driver has no meaning over (under) permanent ice, glacial areas like Greenland should be masked from the maps.

Yes, we agree and figures will be redone to mask out glaciated areas.

Code and data availability: This is not consistent with COPDESS / FAIR data standards to which EGU journals adhere. Public data and/or code repositories should be used and indicated with permanent hyperlinks.

We will make the relevant fields for the figures shown available with the revised manuscript.

Reply to reviewer 2

This paper presents the ExtremeX set of climate model experiments, where in three Earth System Models the moisture and atmospheric circulation are systematically constrained (nudged) towards observation-based values, either separately or jointly. Mean surface temperature and precipitation biases across these different experiments are evaluated, and it is found that these biases do not generally become smaller as the models are more constrained. The ExtremeX experiments are then applied to quantify the degree to which four recent strong heatwaves can be attributed to (i) sea-surface-temperature anomalies, (ii) atmospheric circulation anomalies, (iii) soil-moisture anomalies and (iv) recent climate change (from a 1979-2008 reference period to the time of the four events occurring within 2010-2015). The attribution method is then also applied to a wider set of warm spells during 2010-2015. It is found that most of the heatwaves and warm spells studied are predominantly due to circulation anomalies, with soil-moisture anomalies playing a secondary but important role, especially in subtropical and tropical regions. Contributions from sea-surface-temperature anomalies and recent climate change are typically much smaller than the other two.

The findings of this study are interesting, and it is nice to see a co-ordinated experiment across three models which lends robustness to the results, which will inform future model applications such as seasonal forecasting. I therefore recommend this study for publication in ESD subject to the comments provided below. While the presentation is generally clear, some additional investment in the introduction will make the paper more easily accessible to a wider audience. I also think that the model evaluation section would benefit from a concrete example (case study) in addition to the more general discussion provided so far. The role of the ocean in the ExtremeX setup also needs to be clarified.

We thank the reviewer for the detailed lecture and the helpful comments and suggestions to improve the manuscript. In the following we will address the points raised by the reviewer point-by-point. Answers to the comments made are given in blue below.

General comments

1) Introduction

Having read the whole paper, and then re-read the introduction, I can follow it much better, but I think some additional explanations (and clearer signposting of contents that is already provided) would make the introduction easier to follow, especially for other readers like me who are not necessarily familiar with the predecessor papers of this study. More specifically, I recommend paying attention to the following points:

- Some key references are provided, for example in the first two paragraphs and the lead author's own papers (line 48), but the main findings of these previous studies should be discussed in greater detail, as well as remaining knowledge gaps and which of these gaps this study aims to close.
- The focus and objectives of this study should be made clearer, especially which sort of extremes are to be studied. Line 41 rather vaguely mentions "extreme weather and climate events", whereas in the research questions it then transpires that the interest is in heatwaves/warm spells, although location, extent and duration remain unspecified. Part of my initial confusion seems to be due to the fact that there are two main purposes of this study, namely to (i) introduce the ExtremeX experiments (which I understand have a range of different possible applications) and to (ii) identify the drivers of heatwaves and warm spells, which is the specific application in this study. This distinction should be made clearer.

- Briefly motivate how to get from the conceptual distinction of dynamic and thermodynamic processes to setting up model experiments with constrained soil moisture/atmospheric circulation.

We thank the reviewer for sharing his experience from reading the introduction and the very helpful recommendations on how to improve the understanding and readability. We will follow the advice given. Specifically, we will explain the results of the Wehrli et al., 2018 and 2019 predecessor papers as briefly but also as completely as possible. Further, we will make clear from the beginning that the purpose of the study is to introduce the experiments and that we will apply them to study drivers of four recent heatwaves and to identify globally for which locations warm spells are generally dominated by processes at the land surface or by atmospheric circulation. We will also motivate the constrained experiments on the paragraph on line 36 saying:

"The processes driving a specific extreme event and their relative importance can be studied in observation-based studies using multiple linear regression (e.g. Arblaster et al., 2014; Wang et al., 2016) or forecast sensitivity experiments (e.g. Hope et al., 2016; Petch et al., 2020). In climate model simulations the role of the drivers can be studied by constraining the processes in the ocean, the atmosphere or at the land surface, which allows to study the drivers in isolation (e.g. Fischer et al., 2007; Hauser et al., 2016; Jaeger and Seneviratne, 2011). In this study, …"

References:

Arblaster, J. M., Lim, E.-P., Hendon, H. H., Trewin, B. C., Wheeler, M. C., Liu, G., & Braganza, K.(2014). Understanding Australia's hottest September on record. Bulletin of the American Meteorological Society, 95, 37–41.

Wang, G., Hope, P., Lim, E.-P., Hendon, H. H., & Arblaster, J. M. (2016). Three methods for the attribution of extreme weather and climate events (*018*): Bureau of Meteorology.

Hope, P., Wang, G., Lim, E.-P., Hendon, H. H., & Arblaster, J. M. (2016). What caused the record-breaking heat across Australia in October 2015? Bulletin of the American Meteorological Society, 97(12), S122–S126.

Petch, JC, Short, CJ, Best, MJ, *et al.* Sensitivity of the 2018 UK summer heatwave to local sea temperatures and soil moisture. *Atmos Sci Lett*. 2020; 21:e948. https://doi.org/10.1002/asl.948

Fischer, E. M., Seneviratne, S. I., Vidale, P. L., Lüthi, D., & Schär, C. (2007). Soil moisture-atmosphere interactions during the 2003 European Summer Heat Wave. Journal of Climate, 20(20), 5081–5099.

Hauser, M., Orth, R., & Seneviratne, S. I. (2016). Role of soil moisture versus recent climate change for the 2010 heat wave in western Russia. Geophysical Research Letters, 43, 2819–2826. https://doi.org/10.1002/2016GL068036

Jaeger, E. B., & Seneviratne, S. I. (2011). Impact of soil moisture-atmosphere coupling on European climate extremes and trends in a regional climate model. Climate Dynamics, 36(9), 1919–1939.

2) Validation of experiments

In section 4.2 (roughly Lines 258-283), a general discussion is provided of the issues that can arise in the constrained experiments based on tuned fully interactive models. I don't disagree with this discussion, but it is a little unsatisfactory as it stands, and I think an example (case study), possibly in a new subsection 4.3, could help to illustrate some of these issues more clearly. A case in point already highlighted by the authors are the large summer precipitation biases seen in the MIROC5 AFSI experiment (Fig. 4) without, I believe, correspondingly large biases in clouds or evapotranspiration (Figures A3, A4). I suggest

analysing this further, for example by evaluating the moisture budget (including circulation and transport) of the different experiments in a suitable study region. A possible example is WCA, where, remarkably, the precipitation bias changes sign and the RMSE increases from 0.53 to 4.3 mm/day from AISI to AFSI.

Thank you for the comment. This is a really interesting suggestion. Unfortunately, we do not have all necessary variables from all models for the moisture budget evaluation. As we write in the manuscript it is known that MIROC5 shows large biases of the atmospheric circulation for example in the North Atlantic stormtrack (Zappa et al., 2013). These issues have been and are being targeted in model development. Since the tuning of the model parameterisations compensates for the deficiencies in the circulation we think that the analysis of the moisture budget will not suffice to explain the issues and biases seen. The aim of the presented study is to introduce the ExtremeX models and experiments, the constraining methodologies used and to provide an example for an application of the framework. Hence, we think an analysis going more deeply into a discussion of the origin of biases would be out of the scope of this paper. However, we feel like the question brought up by the reviewer would be a great starting point for a future study dedicated to understanding model biases and making recommendations for model improvement.

3) Role of the ocean

I am unclear about the role of the ocean in the ExtremeX setup and for the results of this study. This is illustrated in the conclusions: In Line 413, the authors say “Thus, the presented set of experiments can be used for extreme event analysis as long as the atmospheric circulation and/ or soil moisture are major drivers of the event.” This means that the role of the ocean must be small – a working assumption, or limitation of the approach. However, in Line 431 it is asserted that “The ocean was not found to have a substantial role in driving any of the events considered” – this reads like a result of this study and may be seen to be incompatible with the earlier statement. Please explain this more clearly.

We thank the reviewer for pointing this out. The experiments can also be used if the ocean is an important driver of the event under consideration. Some of the analysis, like for example the separation in circulation vs. soil moisture driven in Figure 7 would not make a lot of sense in that case. The experiments are also not ideal, if the focus is on the role of the ocean because the ocean is prescribed. If experiments with interactive ocean were available, the ocean contribution could be computed more accurately, which we would recommend for mainly ocean-driven events. In that case the experiment setup would require an additional ensemble of 100 simulations like AI_SI. We will rewrite the sentences on L413 and L431 to explain this.

Minor comments

4) Abstract

The last sentence about where soil moisture effects are important raises the expectation of a similar sentence for the circulation effects.

We will rephrase the last sentence to say: “Soil moisture effects are particularly important in the tropics, the monsoon areas and the Great Plains of the United States, whereas atmospheric circulation effects are major drivers in other mid- and high-latitude regions.”

5) Line 19

What does “consistent” here refer to? Extreme and mean model biases? Or maybe the range of CMIP5 models?

Consistent refers to consistent across models. We will rearrange the sentence to make this more clear:

“For climate models used in the fifth phase of the Coupled Model Intercomparison Project (CMIP5) consistent biases can be found across models in the mean climatology of the lower atmosphere and land surface, for example temperature and precipitation …”

6) Line 48

“… by validating the forcing of the atmosphere and the land for the near-surface climatology.” I did not understand this (before reading the paper). Please rephrase.

We will change this sentence to: “The presented work expands on previous work in Wehrli et al. (2018, 2019) by quantifying biases of the near-surface climatology for different constraining experiments and three models.”

7) Line 52

Specify “overall model biases”.

We will rephrase this to say “climatological model biases”.

8) Table 1

Provide the number of ensemble members as three comma-separated values using a specified order of models, e.g., “5,5,1” for 5 members in CESM1.2, EC-EARTH3, and 1 member in MIROC5.

Thank you for the suggestion, we will do this in the revised manuscript.

9) Line 94

I suggest listing/explaining the different terminologies once upfront (forcing, constraining, nudging, relaxing) and then to stick to one choice for the remainder of the paper. “Constraining” seems to work well.

It turned out that due to a misunderstanding between the modeling groups the assumption was that different methodologies were used to constrain soil moisture. In fact, soil moisture was prescribed in all models. We will therefore use the terms “soil moisture prescription” and “atmospheric circulation nudging” throughout the manuscript and explain the terminology at the beginning of Section 2.2.

10) Line 139

Regarding the “additivity” – can this, or has this, been tested? Clarify briefly.

The additivity assumption has not been tested for the disentangling method presented here. However, it was based on the study by Kröner et al. (2017) where it was shown that it can be assumed that the contribution of the thermodynamic effect due to global warming, the lapse-rate effect and the large-scale circulation (as well as remaining effects) to the summer climate in Europe are additive. The assumption was tested for other seasons but not for other regions in that study. We think that disentangling method A and B giving similar results for all models is a further indication showing that in a first order assumption the contributions can be treated as additive. We will mention this in the manuscript.

Reference:

Kröner, N., Kotlarski, S., Fischer, E. *et al.* Separating climate change signals into thermodynamic, lapse-rate and circulation effects: theory and application to the European summer climate. *Clim Dyn* **48,** 3425–3440 (2017). https://doi.org/10.1007/s00382-016-3276-3

11) Line 141

Replace “analyses investigate” by “disentangling method determines”.

We will change the sentence as suggested.

12) Line 172

Replace “The target data set” by “The prescribed target soil moisture” (if true).

We will change the sentence as suggested.

13) Line 192

Explain (or omit) “non-operational”.

We will omit the expression. What is meant is that the model is not used for making actual (operational) weather forecasts or seasonal predictions.

14) Line 206

Replace “toward observations” by “toward reanalysis” (if true). Make this distinction throughout.

“Toward reanalysis” is correct, we will change that in the manuscript.

15) Figure 2

This figure nicely summarises the performance for different experiments and regions!

Thank you!

16) Figure 4

Explain the grey areas in the caption.

Ocean grid points and Antarctica are masked out in Figures 3 and 4. We will add this to the captions.

17) Line 292

“… nudging the atmospheric large-scale circulation and constraining the soil moisture results …” – The *and* seems key here as there can be substantial biases in the experiments where circulation and soil moisture are constrained individually. Please discuss if this is expected to impact the disentangling method.

There are substantial (climatological) biases in all experiments as we do not perform any bias-correction. Constraining the circulation or soil moisture individually can lead to larger biases than in the unconstrained setup as for example for precipitation in MIROC for the AF_SI experiment. In section 5 we are interested in temperature anomalies during specific events or warm spells in general. These anomalies are always computed with respect to the

climatology of each experiment and model individually. Hence, climatological biases do not come into play here and do not impact the disentangling method. The section focuses on the magnitude of the anomaly and the temporal evolution during specific events. In fact, the nudging only experiment (AF_SI) already compares very well to temperature anomalies from ERA-Interim and even the constrained soil moisture experiment tends to show positive anomalies during the events examined. To make this clearer we will add figures like Figure R3 in the response to Paul Dirmeyer to the appendix.

18) Line 319

Previous work has suggested an important role of anomalous sea surface temperatures for the 2010 Russia heatwave (Trenberth and Fasullo 2012). This study finds that "CESM is the only model which shows a negative ocean contribution of around −7%, whereas the role of the ocean is negligible in the other models".

Does this mean that this study contradicts Trenberth and Fasullo 2012? Is there further evidence for or against in the literature?

Such context with the existing literature should be briefly discussed – also in the conclusions and for the other three events (see also comment 1).

There is other literature supporting a weak role of the ocean to the Russian heatwave like Dole et al. (2011) and Hauser et al. (2016). On the other hand, the Trenberth and Fasullo (2012) study is supported for example by the study of Martius et al. (2013) who link SST anomalies to atmospheric circulation conditions over the Asian continent leading to the Pakistan floods and the Russian heatwave. In that sense our results do contradict these studies. However, this has certainly also to do with differences in experimental setup as the findings here are based on simulations with a prescribed ocean. We will include more discussion of existing literature and comparison to our findings in the revised manuscript.

References:

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

Hauser, M., Orth, R., & Seneviratne, S. I. (2016). Role of soil moisture versus recent climate change for the 2010 heat wave in western Russia. Geophysical Research Letters, 43, 2819–2826. https://doi.org/10.1002/2016GL068036

Martius, O., Sodemann, H., Joos, H., Pfahl, S., Winschall, A., Croci-Maspoli, M., Graf, M., Madonna, E., Mueller, B., Schemm, S., Sedláček, J., Sprenger, M. and Wernli, H. (2013), The role of upper-level dynamics and surface processes for the Pakistan flood of July 2010. Q.J.R. Meteorol. Soc., 139: 1780-1797. https://doi.org/10.1002/qj.2082

19) Line 376

"The spells are analysed by taking the same dates in the experiments." – I don't understand this.

Warm spells are identified and categorised based on the ERA-Interim reanalysis. Then the same date (calendar year and days of the year) is analysed in the experiments. Hence, if at a certain location a warm spell (during the local warm season) lasts from August 12 to August 17 of a given year, temperature anomalies in the experiments during August 12 to August 17 of the same year will be used. We will clarify this in the manuscript.

20) Figure 7

What limits this application to events that last longer than ~2 weeks? Is this simply a question of sampling/ensemble sizes, or an inherent limitation of the disentangling method? Please discuss this briefly.

The choice of the categories used for warm spell lengths was motivated both by sampling size but also other considerations. The lower bound of three days was chosen due to the common definition of heatwaves lasting at least three days or longer. The separation between 5 and 6 days for categories 1 and 2 was made subjectively to separate events lasting a couple of days from events lasting roughly a week but it was also made to obtain a similar sample size. The last category for events lasting two weeks and more was introduced to have an additional separation for very long-lasting events like for example the Russian heatwave. Due to the small sample size (and some regions not showing events of this length) introducing more categories for even longer events would not make sense for the global analysis carried out here.

21) Figure 7

Say in the caption how the local warm season is defined.

The warm season is defined as the hottest consecutive three months (from ERA-Interim) for each grid point. We will explain this in the figure caption.

Reference

Trenberth, K. E., & Fasullo, J. T. (2012). Climate extremes and climate change: The Russian heat wave and other climate extremes of 2010. Journal of Geophysical Research: Atmospheres, 117(D17), n/a-n/a. https://doi.org/10.1029/2012JD018020