

Jena, 07 March 2016

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

Dear Editors and Reviewers,

we thank both reviewers for their positive evaluation of our manuscript and their constructive comments. We also appreciate the more critical comments that were mainly related to structure and clarity of the manuscript.

We respond to all comments in a point-by-point manner in individual responses, which are reproduced below and in few cases slightly expanded. Changes made to the manuscript are highlighted in an additional file.

We believe that the manuscript has improved significantly on the basis of the comments, in particular with respect to clarity, structure, balance of cited references, and in providing a concise message.

Please do not hesitate to contact us in case any further questions arise. We would highly appreciate if our manuscript could be considered suitable for publication in *Earth System Dynamics*.

Sincerely,

Sebastian Sippel on behalf of all authors

22 **Reply to Anonymous Referee #1 comment on Sippel et al., 2016, Earth System Dynamics**  
23 **Discussions, doi:10.5194/esd-2016-48**

24 The paper "Refining multi-model projections of temperature extremes by evaluation against  
25 land-atmosphere coupling diagnostics" uses temperature and ET to explore the land-  
26 atmosphere interactions during heat waves in models. It then uses the derived coupling metric  
27 to constrain the CMIP-5 models and show that the constrained ensemble provides better  
28 representation of heat waves. Overall the paper is interesting and should be published, however  
29 it is poorly written in that there is insufficient justification of the methodology, overuse of  
30 supplementary figures and a lack of discussion and analysis. This makes the paper difficult to  
31 follow and understand. The authors present too much material and they need to trim it down to  
32 provide a strong and concise message. Therefore, I am recommending that the paper needs  
33 major revisions before it is published. The specific comments that motivated this decision are  
34 given below.

35 **We thank the reviewer for the overall positive comments on our manuscript, and also**  
36 **acknowledge the more critical comments. We understand that the critical comments are**  
37 **mainly related to an indeed densely written manuscript, partly unsatisfactory description**  
38 **of figures/methods, and that the manuscript would benefit from a more concise message.**  
39 **In a revised version of the manuscript, we focus on the two main findings of this paper to**  
40 **provide a strong and concise message: Namely, we show that**

- 41 **1) a relatively large fraction of CMIP5 models (up to ~50% in some regions)**  
42 **misrepresent warm season land-atmosphere coupling in mid-latitude regions, i.e.**  
43 **the occurrence of water-limited regimes (coincidences of high T anomalies with**  
44 **low ET anomalies) are systematically overestimated relative to observations-**  
45 **based datasets;**
- 46 **2) the representation of land-atmosphere coupling (i.e. T-ET coincidences) in the**  
47 **models is closely related to biases in the simulation of temperature extremes;**  
48 **therefore the application of a "land-atmosphere coupling constraint" derived from**  
49 **our ensemble of observations-based datasets reduces biases in temperature**  
50 **extremes and variability in a physically consistent way.**

51 **Please note also that we have restructured the Abstract in that it now specifies the**  
52 **presentation of these two main findings in logical order. Also, we have reduced the**  
53 **number of Supplementary Figures to avoid unnecessary detail.**

54 **Furthermore, we improve the description of figures/methods, elaborate on explanations**  
55 **and discussions as outlined below in response to the reviewer's specific comments.**

56 Specific Comments:

57 Pg. 2, lines 3-7: This description makes sense but I cannot see the connection between this  
58 description and the “conceptual” Fig. 1a,b. It may be conceptual for the authors, but there needs  
59 to be more description about what each of the symbols in the figures means and how it connects  
60 to the description in the text. Specifically, what is the difference between a thick and thin arrow,  
61 curved versus straight, positive-negative sign, square, circle, rectangle and different colors.  
62 There is a lot going on in the figure and it is hard to know what is important and what is just  
63 there for aesthetic purposes.

64 **We agree with the reviewer's critique, and provide a detailed description of the purpose**  
65 **of all graphical elements in the revised manuscript. Specifically, we have removed curved**  
66 **arrows, misleading colors and shapes in the revised figure (except that the yellow**  
67 **triangle indicates an external trigger of feedbacks). Plus and minus signs indicate positive**  
68 **and negative impacts, respectively. We hope all this becomes clear in the revised Figure**  
69 **and explanation.**

70 Pg 4, line 12: Why use the old version of the NCEP reanalysis and not the CFSR? Why not use  
71 datasets with consistent temperature, evaporation datasets like MERRA and CFSR. These  
72 datasets have both temperature and evaporation. Why would you expect ET from one dataset to  
73 be correlated with T from another? There needs to be more discussion on this.

74 **One of the main goals of our study was to explicitly use as many ET datasets as possible**  
75 **(including diagnostic/empirical datasets derived from e.g. remote sensing or flux**  
76 **measurements), because there remains a very large spread between individual ET**  
77 **products (see e.g. Mueller et al., 2013). Therefore, using only "consistent" T-ET**  
78 **combinations from a specific reanalyses would exclude diagnostic datasets and would not**  
79 **sample the large spread between the different ET products. However, despite this all T-ET**  
80 **dataset combinations are originally derived from real-world observations (or land**  
81 **models driven by observations), and therefore can be expected to represent relevant**  
82 **features in the observations, especially on large spatio-temporal scales (here monthly**  
83 **data, mostly on SREX-region level). Please note also that the coincidence methodology is**  
84 **robust to noise in that it is not sensitive to any numerical values or outliers in the data**  
85 **(but e.g. covariance or correlation-based metrics would be sensitive to numerical values),**

86 but only considers threshold exceedances (see e.g. Donges et al. (2016) for an in-depth  
87 description of the method and Schleussner et al. (2016) for an application).

88 Most importantly, however, we have investigated the T-ET consistency issue in detail, and  
89 find that the spread in T-datasets is basically negligible in comparison to the large  
90 uncertainties and discrepancies in ET-datasets (please see figures 1 - 3 below):

91 Figure 1 below is reproduced from Fig. 2 in the main manuscript and shows the  
92 distribution of (a) VACb ("energy-limitation") and (b) VACc ("water-limitation") in CMIP5  
93 models and various combinations of T-ET datasets. The ET datasets are derived through  
94 diagnostic means (circles), land surface model simulations (diamonds), and reanalyses  
95 (triangles). If we assume that an "inconsistent" combination of T and ET datasets (i.e.  
96 from different sources) would not be feasible for the present analysis (as suggested by the  
97 Reviewer), then we would expect these "inconsistent" T-ET combinations to be shifted  
98 towards independent noise without correlations (i.e. symmetrically around the pink  
99 range). However, to the contrary, we find that the observations-based dataset  
100 combinations are indeed not strongly affected by noise, as they lie systematically above  
101 (Fig. 1a) or below (Fig. 1b) the noise level. More importantly, however, we have tested  
102 explicitly for the effect of choosing different temperature datasets (see Figure 2 below).  
103 Here, we find that changing the temperature observations that are to be combined with  
104 ET datasets has only a minor effect on the location of the respective grey or black dot in  
105 the VACb-VACc diagram (e.g. three black triangles are the Median Reanalyses ET dataset  
106 combined with CRU-temperature, ERAI-temperature, and NCEP-temperature,  
107 respectively). This indicates that by far the largest source of uncertainty stems from the  
108 choice of ET dataset, and the uncertainty that stems from the choice of temperature  
109 dataset is (almost) negligible. Also, in Fig. 2 we also find that the observations-based T-ET  
110 combinations are systematically shifted from the "random range" (around 0.3 in both  
111 VACc/VACb) to favour the more water-limited (high VACb) regimes (except one  
112 diagnostic ET dataset that in many cases tends to lie around random and uncorrelated  
113 data). Finally, a direct comparison between T datasets reveals that on a monthly time  
114 scale these datasets are almost perfectly correlated (Figure 3, top). In contrast,  
115 correlations between different ET datasets are low (Figure 3, bottom), which indeed  
116 confirms that the uncertainty induced by the choice of T-dataset is almost negligible given  
117 the large uncertainties induced by the spread between different ET datasets.

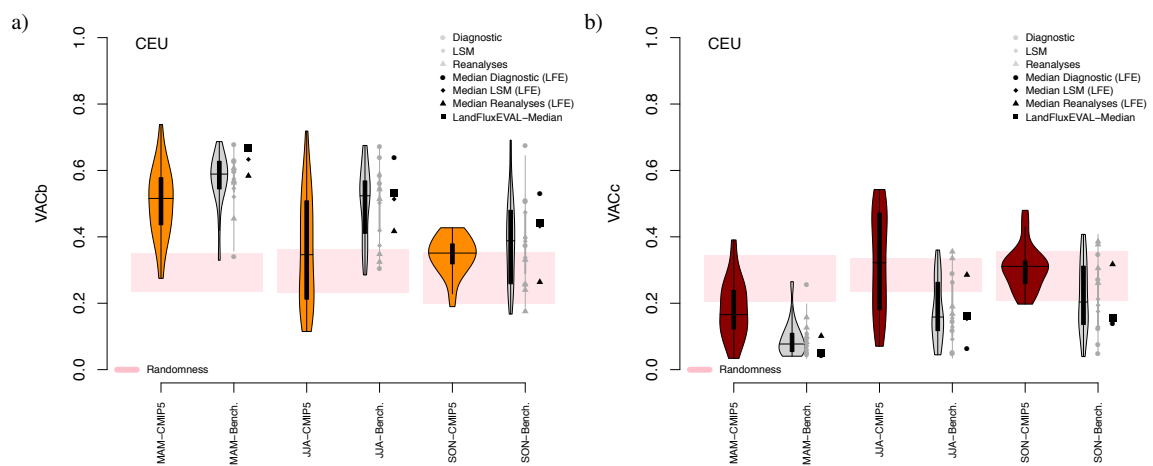
118 We include more discussion on this issue in the revised manuscript (Section 2.1).

119

120 **Regarding temperature from NCEP/DOE reanalysis vs. CFSR reanalysis: Thanks for**  
 121 **pointing out that NCEP/DOE is an old version. We have replaced the NCEP/DOE reanalysis**  
 122 **by the newer CFSR reanalysis and redone all calculations with the three temperature**  
 123 **datasets CRU, ERAI, and CFSR (instead of CRU, ERAI, and NCEP/DOE previously). Results**  
 124 **are almost exactly the same (as to be expected due to well-correlated temperature**  
 125 **datasets, see above).**

126

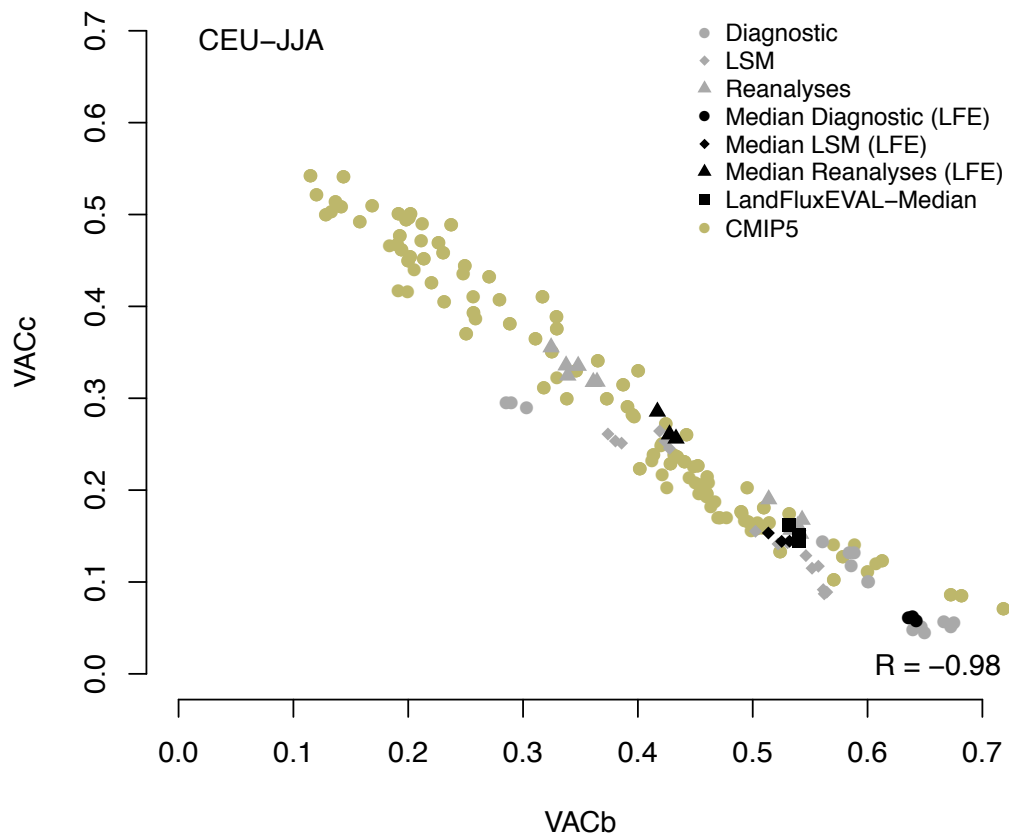
127



128

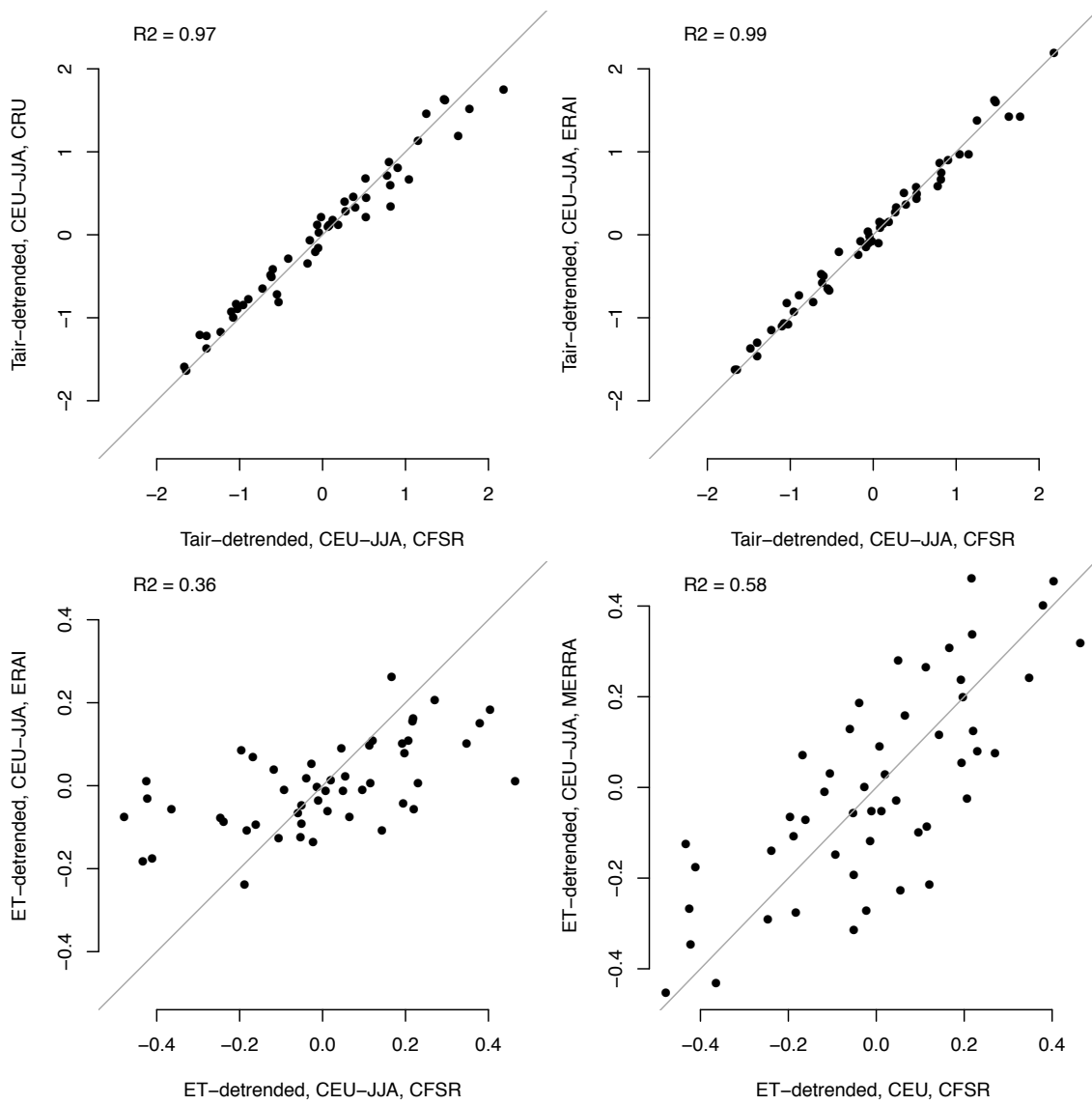
129 **Figure 1: Evaluation of T-ET coupling in global climate models. (a, b) VACb (a) and VACc**  
 130 **coupling in the CMIP5 climate model ensemble and observations-based benchmarking**  
 131 **datasets in Central Europe (CEU, 1989-2005, area-average) with systematic warm season**  
 132 **differences. Randomness indicates the 5th to 95th percentile range obtained by randomly**  
 133 **permutating both time series with respect to the other (N = 100 times) to obtain independent**  
 134 **data (reproduced from Fig. 2a,b in the main manuscript).**

135



136

137 Figure 2: VACb plotted against VACc (Central Europe, JJA) for CMIP5 models and various  
 138 T-ET combinations. All ET datasets have been combined with each CRU-temperature,  
 139 NCEP-temperature, and ERAI-temperature, and therefore each dot is plotted three  
 140 times.



141

142 **Figure 3: Correlation between different Tair datasets (top), and different ET datasets (bottom).**

143

144 Pg 4, lines 19-20: Make sure to reemphasize this when discussing the results.

145 **We reemphasize this point in the Results and Discussion section.**

146 Pg 4, line 28: This is a good use of supplementary material as there is sufficient description to  
147 know that it contains a list of the 37 models used, but it is not necessary for understanding and  
148 interpreting the results of this paper.

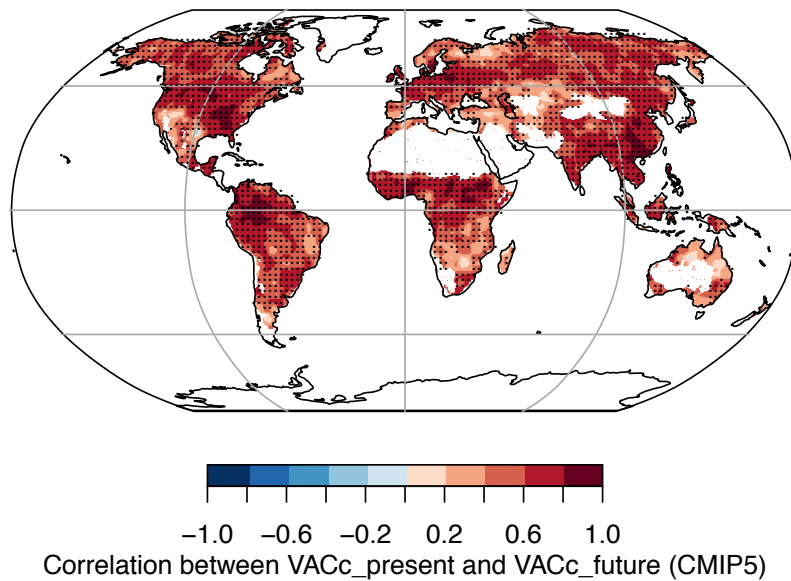
149 Pg 4, lines 28-30: This is an important assumption for this study and the description is a bit  
150 vague. What does “tend” mean? It would be better to provide some sort of quantitative measure  
151 of the variability across ensembles. Is this true for all locations? Given the importance of this  
152 assumption there needs to be further analysis and discussion as to why you think this is a  
153 reasonable assumption.

154 **We agree this is an important point that needs more clarification: VACc at present is a**  
155 **very good predictor for VACc in the future (significant correlation globally across almost**  
156 **all locations, see Figure 4 below). This indicates that variability in land-atmosphere**  
157 **coupling across models as diagnosed by the VACc index is largely a model-inherent**  
158 **feature that is determined by model structure, and model-internal variability on 30-year**  
159 **time scales plays only a comparatively minor role. More importantly, Figures S6 and S7 in**  
160 **the Supplementary Material of the Discussion version confirm this interpretation: While**  
161 **there is some variation in VACc within individual models in the 1989-2005 period (mostly**  
162 **in models with many ensemble members including perturbed physics), the large model**  
163 **spread overall is clearly dominated by across-model variability (and therefore we have**  
164 **chosen only one ensemble member per model -the often-chosen standard run r1i1p1- to**  
165 **investigate variability across the CMIP5 ensemble without being confounded by unequal**  
166 **ensemble sizes for individual models).**

167 **However, it is important to note that there might be substantial variability in land-**  
168 **atmosphere coupling within models or ensemble runs of the same model for individual**  
169 **(e.g. extreme) years (see e.g. Sippel et al., 2016 for a regional model over Europe) or on**  
170 **multi-year but not climatological time scales, or for model ensembles with perturbed**  
171 **physical parametrisations (which is beyond the scope of this manuscript but could be**  
172 **promising future research). We clarify all these points in the revised manuscript.**

173





174

175 **Figure 4: Correlation between present-day and future land-atmosphere coupling reveal that**  
 176 **occurrences of VAC<sub>c</sub> are largely model-inherent features.**

177

178 Pg 6, line 2: I find this equation and description extremely confusing. It took me several times of  
 179 reading the text to understand the metric. I am still have no idea of what is represented in the  
 180 equation particularly in the summation. Where does the 1 come from and why does it have a  
 181 subscript? The equation is more confusing than the text and does not help at all with  
 182 understanding the metric. From what I understand from the metric, the VAC<sub>b</sub> gives the percent  
 183 of the highest 30% of temperatures that correspond with highest 30% of ET, while VAC<sub>c</sub> gives  
 184 the percent of the highest 30% temperatures that correspond with the lowest 30% ET. It is  
 185 difficult to remember which was VAC<sub>b</sub> and VAC<sub>c</sub>. It would be helpful if there was naming  
 186 convention that is more descriptive instead of b and c.

187 **We agree that the equation is confusing and aim to explain it better in the revised**  
 188 **manuscript. Your interpretation of the VAC<sub>b</sub> and VAC<sub>c</sub> metric is correct - and we will**  
 189 **clarify this interpretation in the text (i.e. include a verbal description in the text after the**  
 190 **equations).**

191 **The equation is simply there to state that we count coincidences of T and ET in a given**  
 192 **category (e.g. positive T and negative ET) to get the average coincidence rate r\_VAC<sub>b</sub>. The**  
 193 **"1" in the equation just means we count 1 for each occurrence of VAC<sub>b</sub> - and 0 otherwise**  
 194 **(the indicator function, [https://en.wikipedia.org/wiki/Indicator\\_function](https://en.wikipedia.org/wiki/Indicator_function)).**

195 **Lastly, we also put more emphasis on the meaning of VACb ("energy-limitation") and**  
196 **VACc ("water-limitation"). However, we would not rename VACb and VACc, just to be**  
197 **consistent with earlier studies that introduced the VAC-metric (Zscheischler et al., 2015).**

198

199 Pg 6, line 10: I thought Fig 1a,b was just a conceptual example, how does this connect with the  
200 simple example of monthly time series referenced here? It seems to me that Fig 1a,b are  
201 completely unrelated with the rest of the figure, so why put them together? If they are related,  
202 then there needs to be more description as to how they are connected.

203 **The subfigures are indirectly related: In Fig. 1a and Fig. 1b we conceptualize "wet" and**  
204 **"dry" regimes, respectively. Fig. 1c/e and Fig. 1d/f then illustrate how time series and the**  
205 **T-ET correlations look like in models that predominantly produce "wet" (NorESM1-M)**  
206 **and "dry" (ACCESS1-3) regimes, respectively. In the revised version, we explain this point**  
207 **better in the caption and in the text. The new text reads (i.e. Section 2.2):**

208 "Fig. 1c-f shows a simple example of monthly time series of T and ET simulated from two CMIP5  
209 models and occurrences of VACb and VACc are highlighted. Note that for the same region (area-  
210 average over Central Europe, CEU) and time of the year (monthly data for June, July, and  
211 August), one model produces predominantly energy-limited regimes (VACb, Fig. 1c,e and  
212 compare to conceptual illustration in Fig. 1a), whereas the other model produces predominantly  
213 water-limited regimes (VACc, Fig. 1d,f and concept in Fig. 1b)."

214 Pg 6, line 14-16: "Might" is not very reassuring and is an inherently weak argument. Correlation  
215 is also universally known and if you use a rank correlation it can also pick-up on the non-  
216 linearities.

217 **"Might" can be safely removed because in the next sentence as we say (and show) that it**  
218 **does pick up non-linearities (see next comment below).**

219 Pg 6, lines 16-19: Not sure what this means. This needs more discussion. Also, this is an  
220 inappropriate use of supplementary information. There is no information about what is plotted  
221 in S1. Furthermore, this seems like an important justification of the methodology and should be  
222 included in the text.

223 **In the revised manuscript we provide an expanded explanation what we mean:**  
224 **Essentially, correlations (Pearson, rank, etc.) emphasize the whole distribution, while our**  
225 **coincidence analysis using VACb and VACc only looks at the warm tail. Figure S1 shows**

226 **that, by correlating VACb and VACc in both models (red), observations (black), and**  
227 **artificial data sampled from a two-dimensional Gaussian distribution with different**  
228 **covariances, both models and observations (red and black dots in bottom left subfigure in**  
229 **S1) the VACb and VACc rates exceed those that would be expected in artificial data (grey**  
230 **dots). This deviation indicates that the warm tail is indeed different to the remainder of**  
231 **the distribution (e.g. no such deviation is detected for the cold tail, VACa and VACd in the**  
232 **bottom right subfigure); and hence an evaluation metric that focuses on the (warm) tail is**  
233 **indeed useful.**

234 Pg 6, lines 18-19: If it yields quantitatively similar results then why bother with the VAC? It is  
235 much easier to understand correlation. What about the significance level for the VAC? Is there  
236 any way to statistically quantify the significance of the VAC as to provide some level of  
237 confidence? If not, that is a major disadvantage over a traditional correlation metric and needs  
238 to be discussed. I am not against using the VAC, but as presented here the justification for using  
239 it is extremely weak. You need to convince the reader that this obscure metric is worth using.

240 **Please consider a few things regarding the method:**

241 **First, as noted above, we believe that it is an advantage to use "non-traditional"**  
242 **evaluation metrics in this case because these focus on the tails rather than the whole**  
243 **distribution (and this would become even more important for sub-monthly or daily time**  
244 **series).**

245 **Second, the method is robust and non-parametric - i.e. it is not sensitive to numerical**  
246 **values or outliers in the data. For this reason it has been used in a number of studies**  
247 **dating back quite a long time (e.g. Quiroga et al., 2002), but also very recent studies**  
248 **(Siegmund et al., 2016).**

249 **Significance can be established quite simply with coincidence metrics, see e.g. Donges et**  
250 **al. (2016) for an overview. In our paper we use a permutation-based scheme to find the**  
251 **range of VACc-rates one would obtain in random data (e.g. Fig. 2). We have made all these**  
252 **points more clear in the revised manuscript.**

253 Pg 6, line 20: This is a reasonable use of a supplementary figure because it provides more depth  
254 to the analysis, but is not directly necessary for understanding the paper. It could be improved  
255 by providing a better description. For example, "the model evaluation as shown in Fig X is  
256 presented for a 90th percentile threshold in Fig S2 and shows . . ." If this figure is completely

257 different from any other figure presented in the text, then you need to include a more  
258 descriptive discussion about it in the text.

259 **Thanks for the suggestion! We have taken it up.**

260 Pg 8, lines 4-5: A figure is not a reference, don't use it like one. There needs to be discussion  
261 about the figure and what it shows. Parts a and b need to be explained more. I am unsure as to  
262 what the different shapes represent and the colors.

263 **We appreciate this point and make sure that every figure is appropriately explained and**  
264 **discussed, both in the text and in the caption. Shapes represent the different ET datasets**  
265 **used (diagnostic, land surface model-based, and reanalyses) - and the violins indicate the**  
266 **distribution over all CMIP5 models (orange or dark red) and observations-based datasets**  
267 **(gray).**

268 Pg 8, line 10: Fig. 3b does not say Pearson correlation, make sure it is consistent with the other  
269 plots.

270 **Thanks, fixed.**

271 Pg 8, line 25: It is ok to reference a figure like this if you have already discussed it but since this  
272 is the first time that fig 4 has been mentioned you need to describe what is being plotted. Also,  
273 this is an inappropriate use of supplementary material. There needs to be explanation in the  
274 text. How does it differ from what is being plotted in fig 4a-b. Seems like the authors are using  
275 supplementary figures instead of actually discussing the important aspects of the analysis.

276 **Thanks for insisting on a proper discussion of all material. We have expanded the description and**  
277 **discussion of this Figure, and also explained that the Supplementary Figures highlight additional**  
278 **details in that they illustrate all individual occurrences of VACc and VACb events.**

279 Pg 8, line 31: I think Fig. 4 is the best and most impactful figure in this paper. Make sure that you  
280 emphasize its importance.

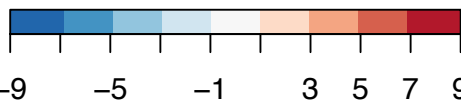
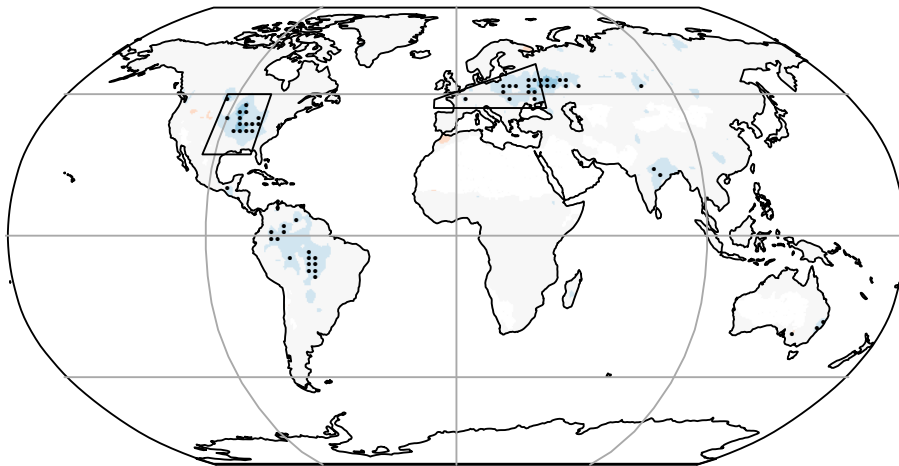
281 **Thanks, we will emphasize the points that are raised by Fig. 4 more in the revised**  
282 **manuscript.**

283 Pg 9, line 2: What does "substantially" mean? Is it significant statistically speaking?

284 **Thanks for this hint. We have indeed tested whether the reductions in TXx and in the 90th**  
285 **percentile TXx (ensemble average across the 90th percentile TXx in each model) are**

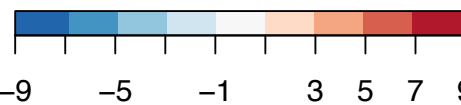
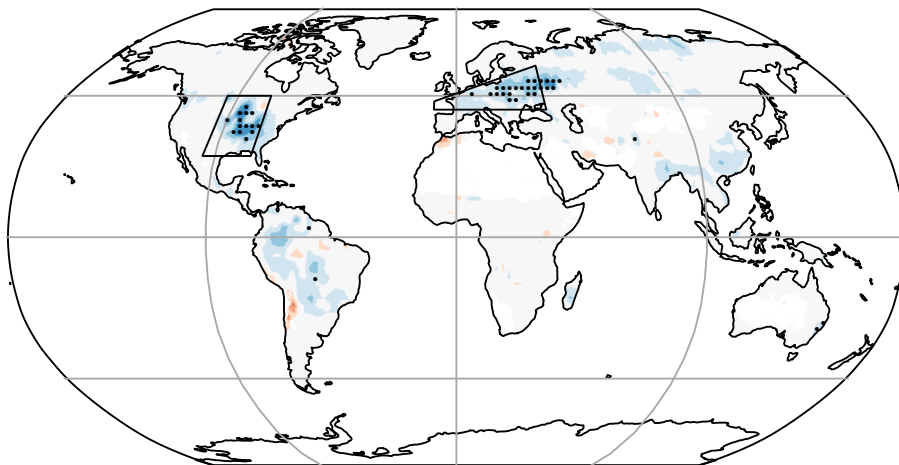
286 **statistically significant, using a non-parametric permutation t-test. Indeed, as would be**  
287 **expected, these reductions are significant in coupling-sensitive transitional regions (CEU,**  
288 **CNA, partly in the Amazon); whereas in regions that are not sensitive to land-atmosphere**  
289 **coupling the constraint does not induce any significant changes:**

290



Absolute reduction in txx-mean [°C] (warm seas., 1981–2010)

291



Absolute reduction in txx-90th perc. [°C] (warm seas., 1981–2010)

292

293 Pg 9, line 25: Again there is no information about what is plotted in the figure and the text makes  
294 it sound relevant for understanding and interpreting the results for this paper and therefore it  
295 should be included in the actual paper and not as supplementary material.

296 **The intention was to indicate that an appropriate representation of land-atmosphere**  
297 **coupling is not only relevant for T and ET - but for related to biases in other variables as**  
298 **well (but which is not the main focus of this paper). We will make this clear.**

299 Pg 10, Line 3: Again, supplementary plots are not a reference. There needs to be an explanation  
300 of the figure.

301 **We have removed the reference to Fig. S13 here.**

302 Pg 10, Line 8: If you reference the same supplementary figure more than once, then that is a  
303 good indication that it should be included in the text. There is more discussion in the text on  
304 S13 then on Fig. 5.

305 **Thanks for these useful comments. We will provide an extended explanation and**  
306 **discussion in the revised manuscript.**

307

## 308 **References**

309 Donges JF, Schleussner CF, Siegmund JF, Donner RV. Event coincidence analysis for quantifying  
310 statistical interrelationships between event time series. The European Physical Journal Special  
311 Topics. 2016;225(3):471-87.

312 Mueller B, Hirschi M, Jimenez C, Ciais P, Dirmeyer PA, Dolman AJ, Fisher JB, Jung M, Ludwig F,  
313 Maignan F, Miralles DG. Benchmark products for land evapotranspiration: LandFlux-EVAL multi-  
314 data set synthesis. Hydrology and Earth System Sciences. 2013; 17:3707-3720.

315 Quiroga RQ, Kreuz T, Grassberger P. Event synchronization: a simple and fast method to  
316 measure synchronicity and time delay patterns. Physical review E. 2002;66(4):041904.

317 Schleussner CF, Donges JF, Donner RV, Schellnhuber HJ. Armed-conflict risks enhanced by  
318 climate-related disasters in ethnically fractionalized countries. Proceedings of the National  
319 Academy of Sciences. 2016;113(33):9216-21.

320 Siegmund JF, Wiedermann M, Donges JF, Donner RV. Impact of temperature and precipitation  
321 extremes on the flowering dates of four German wildlife shrub species. *Biogeosciences*.  
322 2016;13(19):5541.

323 Sippel S, Otto FE, Forkel M, Allen MR, Guillod BP, Heimann M, Reichstein M, Seneviratne SI,  
324 Thonicke K, Mahecha MD. A novel bias correction methodology for climate impact simulations.  
325 *Earth System Dynamics*. 2016;7(1):71.

326 Zscheischler J, Orth R, Seneviratne SI. A submonthly database for detecting changes in  
327 vegetation-atmosphere coupling. *Geophysical Research Letters*. 2015;42(22):9816-24.

328

329 **Reply to Anonymous Referee #2 comment on Sippel et al., 2016, Earth System Dynamics**  
330 **Discussions, doi:10.5194/esd-2016-48**

331

332 The manuscript by Sippel et al. addresses the reduction of ensemble temperature projections by  
333 using best estimates of soil moisture-temperature coupling diagnostics under current climate  
334 conditions. Although the technique itself has been applied in several other studies, the current  
335 application is novel and the results are highly relevant for our understanding of projections of  
336 temperature extremes. The manuscript is generally well-written and results are presented in a  
337 concise manner. The work seems technically sound, and I could not detect any major flaws in the  
338 reasoning and/or analysis, although some minor points were identified that will need to be  
339 addressed. Therefore, I believe the manuscript can be accepted for publication after minor  
340 revisions.

341 **We appreciate the positive evaluation of our study and research.**

342 I have the following remarks/observations:

- 343 • VAC is based on the 30/70th percentile, whereas the authors consider the 90th  
344 percentile of TXx. Please motivate if and why this is justified and consistent (coupling  
345 might be different for highest percentiles).

346 **Yes, indeed - The reviewer is correct in that coupling might be very different far in the tail**  
347 **of e.g. the temperature distribution (e.g. for the highest percentiles temperature**  
348 **extremes vs. warm, but not extremely warm conditions). This is an important caveat of**  
349 **our study (since we are unable to address very rare events because observations-based**  
350 **datasets are generally short in time and in many cases only available on monthly**  
351 **resolution).**

352 **In the end, both choices are somewhat subjective: The choice for the 30/70th percentile**  
353 **for determining the coupling metric has been discussed (only) briefly in the manuscript:**  
354 **Here, the point is that the threshold choice is basically a trade-off between having enough**  
355 **data while still looking at warm conditions (for both VACb and VACc). An additional**  
356 **analysis using the 10th/90th percentile for computing VACb and VACc yields very similar**  
357 **results (Figure S5), therefore increasing the confidence in our results independent of the**  
358 **specific threshold choice, but unavoidably throws away more data.**

359 **For TXx, we look at both ensemble mean TXx and the 90th percentile TXx across the**  
360 **ensemble (cf. Fig. 4d for TXx ensemble mean and Fig. 4f for TXx 90th percentile). While**



361 **the metric "ensemble mean TXx" is quite natural, the consideration to choose the "90th**  
362 **percentile TXx" arose mainly from considering the "upper end" of projected TXx values**  
363 **(similar metrics based on ensemble spread are also taken as the uncertainty bounds for**  
364 **heat extremes, see e.g. Seneviratne et al., 2016, *Nature*). Again, changes in ensemble mean**  
365 **TXx and 90th percentile TXx are consistent - i.e. the changes induced by the constraint**  
366 **have the same sign, but are more pronounced for the 90th percentile of TXx. Therefore,**  
367 **we believe that these choices are well-justifiable, and we make these considerations more**  
368 **clear in the revised manuscript.**

369 **However, the inherent subjectivity of these choices also means that there is scope for**  
370 **additional research that would look at coupling characteristics under very strong**  
371 **heatwaves e.g. in a small number of models with a large number of ensemble members to**  
372 **test the within-model variability in land-atmosphere coupling and its relation to extreme**  
373 **events. We have included this point in the Discussion in Section 3.1.**

374

375 • While the manuscript has a balanced number of display items, I found the link between  
376 the information displayed and that discussed in the text weak. Many sub-panels are  
377 never mentioned or discussed, and too much is left for the reader to interpret. Please  
378 make sure all relevant information in the figures is referred to, as well as all figures and  
379 sub-panels themselves. In particular a more in depth- discussion of the results in Figures  
380 5 and 6 is needed.

381 **Thanks for these suggestions. We have restructured the discussion section and put more**  
382 **emphasis on the discussion of each single display item (please see also similar comments**  
383 **made by Reviewer #1). In particular Fig. 5 and Fig. 6 are discussed in significantly more**  
384 **detail (Section 3.2) and feature expanded captions. Also, in the revised manuscript we**  
385 **refer to the individual sub-panels of the figures to make the connection between the**  
386 **discussion and the relevant figure sub-panels clear.**

387 • The selection of references doesn't always to justice to work that other groups have been  
388 doing in this area or on this specific topic. In the introduction on weighing models in  
389 large ensembles (Page 3, lines 8–12), some examples are provide but interestingly the  
390 ones most relevant to the current work are not cited (i.e. Fischer et al., 2012 and  
391 Stegehuis et al., 2013). In this way, the suggestion is made that this study is the first to  
392 apply model selection on temperature extremes. Please include references to these  
393 works.

394 **Thanks for pointing this out. Our intention was by no means to claim that there has been**  
395 **no application of other model constraints on soil-moisture temperature coupling (as the**  
396 **Reviewer correctly points out, e.g. H in the Stegehuis et al., 2013, paper; Interannual**  
397 **temperature variability in the Fischer et al., 2012, paper). We have discussed and cited**  
398 **both papers mentioned in the discussion section of our manuscript, but it is true that we**  
399 **should have referred to them also in the motivation. In the revised version we have fixed**  
400 **this.**

401 Also, model selection/weighing has been applied to other aspects/fields such as snow albedo  
402 feedback (Hall and Qu, 2006) and hydrological drought projection (Van Huijgevoort et al., 2014).

403 **Thanks for these references. They are indeed highly relevant to the study and we refer to**  
404 **them both in the motivation section in the revised manuscript.**

405 When discussing the vegetation-atmosphere coupling index (VAC), the authors refer to previous  
406 work from the group (e.g. Seneviratne et al., 2006; Lorenz et al., 2012) from which VAC was  
407 developed, but not to other alternative indices that are based on a similar philosophy (for  
408 instance the metric developed by Miralles et al., 2012, although this paper is cited in a different  
409 context).

410 **Thanks for this suggestion and the reference to the Miralles et al, 2012 paper. We have**  
411 **also fixed this.**

412

#### 413 **References**

414 Hall, A. & X. Qu (2006) Using the current seasonal cycle to constrain snow albedo feedback in  
415 future climate change. Geophys. Res. Lett., 33, L030502, doi:10.1029/2005GL025127.

416 Van Huijgevoort, M.H.G., et al. (2014) Identification of changes in hydrological drought  
417 characteristics from a multi-GCM driven ensemble constrained by observed discharge. J. Hydrol.,  
418 512, 421–434.

419

#### 420 **References**

421 **Fischer EM, Rajczak J, Schär C. Changes in European summer temperature variability revisited.**  
422 **Geophysical Research Letters. 2012;39(19).**

423 Seneviratne SI, Donat MG, Pitman AJ, Knutti R, Wilby RL. Allowable CO2 emissions based on  
424 regional and impact-related climate targets. *Nature*. 2016;529:477-483.

425 Stegehuis AI, Teuling AJ, Ciais P, Vautard R, Jung M. Future European temperature change  
426 uncertainties reduced by using land heat flux observations. *Geophysical Research Letters*.  
427 2013;40(10):2242-5.

428

429

430

431

432

433

# Refining multi-model projections of temperature extremes by evaluation against land-atmosphere coupling diagnostics

Sebastian Sippel<sup>1,2</sup>, Jakob Zscheischler<sup>2</sup>, Miguel D. Mahecha<sup>1</sup>, Rene Orth<sup>2</sup>, Markus Reichstein<sup>1</sup>, Martha Vogel<sup>2</sup>, and Sonia I. Seneviratne<sup>2</sup>

<sup>1</sup>Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, 07745 Jena, Germany.

<sup>2</sup>Institute for Atmospheric and Climate Science, ETH Zürich, 8075 Zürich, Switzerland.

Correspondence to: Sebastian Sippel (ssippel@bgc-jena.mpg.de)

**Abstract.** The Earth's land surface and the atmosphere are strongly interlinked through the exchange of energy and matter (e.g. water and carbon). This coupled behaviour causes various land-atmosphere feedbacks, and an insufficient understanding of these feedbacks contributes to uncertain global climate model projections. For example, a crucial role of the land surface in exacerbating summer heat waves in mid-latitude regions has been identified empirically for high-impact heatwaves, but individual climate models differ widely in their respective representation of land-atmosphere coupling. Here, we combine an ensemble of 54 combinations of observations-based and simulated temperature (T) and evapotranspiration (ET) benchmarking datasets and investigate coincidences of T anomalies with ET anomalies as a proxy for land-atmosphere interactions during periods of anomalously warm temperatures. We demonstrate that a relatively large fraction of state-of-the-art climate models from the Coupled Model Intercomparison Project (CMIP5) archive produces systematically too frequent coincidences of high T anomalies with negative ET anomalies in mid-latitude regions during the warm season and in several tropical regions year-round. Further, we show that these coincidences (high T, low ET) are closely related to the representation of temperature variability and extremes of simulated temperatures across the multi-model ensemble. Thus, our approach offers a physically consistent, diagnostic-based avenue to evaluate these ensembles, and subsequently reduce model biases in simulated and predicted extreme temperatures. Following this idea, Second, we derive a land-coupling constraint based on the spread of 54 combinations of the T-ET benchmarking datasets and consequently retain only a subset of CMIP5 models that produce a land-coupling behaviour that is compatible with these observations-based benchmark estimates. The constrained multi-model projections exhibit lower temperature extremes in simulations exhibit more realistic temperature extremes of reduced magnitude in present climate in regions where models show substantial spread in T-ET coupling, and in addition, i.e. biases in the climate-model ensemble are consistently reduced. Also the multi-model simulations for the coming decades display decreased absolute temperature extremes in the constrained ensemble. On the other hand, the differences between projected and present-day climate extremes are affected to a lesser extent by the applied constraint, i.e. projected changes are reduced locally by around 0.5°C to 1°C - but this remains a local effect in regions that are highly sensitive to land-atmosphere coupling. In summary, our approach offers a physically consistent, diagnostic-based avenue to evaluate multi-model ensembles, and subsequently reduce model biases in simulated and projected extreme temperatures.

## 1 Introduction

The exchange of matter and energy between the land surface and the atmosphere is a crucial feature of the Earth's climate (Bonan, 2015) (Seneviratne et al., 2010b; Bonan, 2015; van den Hurk et al., 2016). On one hand, the atmosphere exerts a key influence on land surface processes such as vegetation growth by supplying light, water and carbon dioxide (Köppen, 1900).  
5 On the other hand, the land surface feeds back to the atmosphere, for example through the partitioning of energy into latent and sensible heat fluxes, or by modifying land surface properties, thus implying a direct link to near-surface climate (Koster et al., 2004; Seneviratne et al., 2010b). Conceptually, coupling between the atmosphere and the land surface is often classified into two qualitatively different regimes, a so-called "energy-limited" and "water-limited" regime (Seneviratne et al., 2010b):  
10 In the wet (energy-limited) regime, the land surface is largely controlled by the atmosphere through radiation (see conceptual Fig. 1a,b), implying a positive association between near-surface temperature (T) and evapotranspiration (ET). In contrast, in a dry, water-limited state, the land controls near-surface climate through a lack of soil moisture, and a corresponding reduction in evapotranspiration and latent cooling (see conceptual Fig. 1a,b) with a negative association between T and ET. Therefore, the state of the land surface and land-atmosphere feedbacks modulate and amplify climatic extreme events such as heat waves in mid-latitude regions (Seneviratne et al., 2006; Fischer et al., 2007; Hirschi et al., 2011; Whan et al., 2015; Hauser et al., 2016).  
15 An understanding of these feedbacks might yield improved seasonal predictability of extremes (Quesada et al., 2012), and could help to constrain and better predict model-simulated present and future climate variability in these regions (Seneviratne et al., 2006; Lorenz et al., 2012; Dirmeyer et al., 2013; Seneviratne et al., 2013; van den Hurk et al., 2016; Davin et al., 2016).

However, at present large uncertainties and methodological inconsistencies prevail in both understanding and quantification of land-atmosphere coupling at various spatial and temporal scales, which relate to

- 20 i. scarcity of accurate observational products of soil moisture or evapotranspiration at large spatiotemporal scales and relatively short observational periods (Mueller and Seneviratne, 2014) (Seneviratne et al., 2010b),
- ii. the metrics and variables used to quantify land-atmosphere coupling differ widely in the variables they address (Seneviratne et al., 2010b), and in emphasizing either the whole distribution (Dirmeyer, 2011; Lorenz et al., 2012; Miralles et al., 2012), or the tails of relevant variables (Zscheischler et al., 2015).

25 As a consequence, uncertainties and methodological inconsistencies contribute to a greatly diverging representation of land-atmosphere coupling in state-of-the art climate models (Koster et al., 2004; Boé and Terray, 2008, see also Fig. 1a,b for a simple conceptual example), and further contribute to uncertainties related to projected increases in summer temperature variability in the 21st century in mid-latitude regions (Seneviratne et al., 2006; Dirmeyer et al., 2013). In this context, it has been noted that accurate simulations of temperature variability and extremes require a realistic representation of land-atmosphere  
30 interactions (Seneviratne et al., 2006; Fischer et al., 2012; Bellprat et al., 2013). In other words, biases in temperature variability and extremes might in part stem from an unrealistic representation of land-atmosphere interactions (Fischer et al., 2012; Lorenz et al., 2012; Davin et al., 2016), likely leading to temperature-dependent biases in multi-model ensembles (Boberg and Christensen, 2012; Bellprat et al., 2013).

A model evaluation focus on interpretable land-atmosphere coupling diagnostics might serve as a complementary strategy to traditional model validation and testing (Seneviratne et al., 2010a; Santanello et al., 2010; Mueller et al., 2011b; Mueller and Seneviratne, 2014). Hence, this approach is intended towards testing and understanding the spread and physical consistency in simulated relationships in state-of-the-art multi-model ensembles (e.g. the Coupled Model Intercomparison Project, CMIP5  
5 Taylor et al., 2012) against available observations-based datasets. For example, in the context of land-atmosphere coupling, earlier studies used bivariate correlation- or regression-based metrics to test and evaluate coupling behaviour (Hirschi et al., 2011; Lorenz et al., 2012). Conceptually, the notion of "diagnostic-based model evaluation" as discussed here is consistent with so-called "pattern-oriented model evaluation" (Grimm and Railsback, 2012; Reichstein et al., 2011) - the latter being applied in the context of evaluating simulated and observed patterns at multiple scales in a data-driven way (e.g. in the context  
10 of ecosystem carbon turnover times, Carvalhais et al., 2014).

In the context of extracting credible and relevant information from large (multi-)model ensembles, weighting or selecting models based on ~~relevant~~, observations-based constraints has become increasingly popular recently (Tebaldi and Knutti, 2007; Knutti, 2010), as a priori model ensembles might be seen as a somewhat arbitrary collection of model runs (or "ensembles of opportunity"). For example, empirical and/or physics-based criteria have been used to constrain snow-albedo feedbacks  
15 (Hall and Qu, 2006), constrain carbon cycle projections (Cox et al., 2013; Wenzel et al., 2014; Mystakidis et al., 2016), or in the context of refining precipitation projections (Orth et al., 2016). Moreover, empirical diagnostics are applied to select models for event attribution analyses (Perkins et al., 2007; King et al., 2016; Otto et al., 2015), ~~in the context of refining precipitation projections (Orth et al., 2016)~~ and analyses of drought projections based on model performance (Van Huijgevoort et al., 2014), or to resample large initial-condition ensembles to alleviate biases without distorting the multivariate structure of climate model  
20 output (Sippel et al., 2016b). ~~However, In the context of land-atmosphere coupling, Fischer et al. (2012) and Stegehuis et al. (2013) have constrained a regional model ensemble over Europe using present-day interannual variability of summer temperature, and observations-based estimates of summer sensible heat fluxes. However, these studies came to somewhat conflicting results with respect to the obtained change in warming projections, which probably was due to the underlying choices of datasets to obtain the constraints (Stegehuis et al., 2013).~~ Hence, care is needed in that these practices might not necessarily translate into improved future climate projections or reduced uncertainties. That is because the selection of relevant metrics is clearly not  
25 trivial ~~and-but~~ subjective, and because good model performance w.r.t. any given metric does not translate directly into (more) reliable projections (Knutti, 2008).

~~Hence~~Therefore, the starting point for the present analysis ~~is~~ -in the sense of being necessary, but not sufficient to assure reliability of future climate projections- ~~is~~ that physically motivated, observations-based diagnostics might offer

- 30
1. a link to identify and interpret relevant processes across multiple models (i.e., model evaluation), and
  2. to reduce biases by focusing the interpretation of multi-model ensembles on models that are "right for the right reasons". Most notably climate impacts, including extremes, typically depend on the multivariate structure of climate variables, where simple univariate statistical bias correction methods are prone to failure (Ehret et al., 2012; Cannon, 2016).

In this study, we first evaluate land-atmosphere coupling in state-of-the-art global climate models from the CMIP5 archive and a large ensemble of observations-based ET datasets (Mueller et al., 2013) that has been compiled to address the aforementioned uncertainties in land-atmosphere coupling. In our analyses a land-atmosphere coupling metric that is based on coincidences of temperature and evapotranspiration anomalies is applied. The idea behind a coincidence metric as opposed to a traditional univariate evaluation of model simulated ET fluxes or temperature is that it is insensitive to biases in the simulated means or variances, and thus focusses only on an abstract property of the data, namely the bivariate dependence structure of T and ET. Secondly, we derive a model constraint based on the physically motivated land-coupling diagnostic and the ensemble of benchmarking datasets in order to explore the implications of a reduced ensemble but with land-atmosphere coupling that is *within the range* of the benchmarking datasets.

## 2 Data & Methods

### 2.1 Datasets for T-ET coupling analysis and model evaluation

#### Global temperature and evapotranspiration datasets

In order to evaluate T-ET coupling in global climate models, an ensemble of 18 gridded ~~evapotranspiration-ET~~ estimates, taken from the LandFlux-EVAL multi-data set synthesis project (Mueller et al., 2013), are combined with three different observations-based and reanalysis-driven temperature datasets, yielding in total 54 T-ET combinations (see Table 1). T-ET coincidence rates are calculated from each of those 54 combinations to evaluate and constrain the multi-model ensemble of global climate models (Section 3). The ensemble of ET reference datasets has been generated by combining a wide range of different ET estimates, ~~based on consisting of~~ five diagnostic (~~observations-based~~based on remote sensing or in-situ observations) products, five land surface models driven by ~~observations-observed climate forcing~~ and four reanalysis products (Mueller et al., 2013). The three temperature datasets are based on one observational product (~~the Climate Research Unit dataset, (Harris et al., 2014)~~) (Climate Research Unit dataset, Harris et al., 2014) and two reanalysis products (~~The ERA-Interim reanalysis (ERA-Interim, Dee et al., 2011), and Climate Forecast System Reanalysis (CFSR, Saha et al., 2010), (Dee et al., 2011), and the National Center of Environmental Prediction (NCEP) reanalysis (Kalnay et al., 1996)~~, see Table 1). ~~As fewer temperature than evapotranspiration datasets are used for the present study, we~~ for details. The large number of T-ET dataset combinations is used in order to take uncertainties in both T- and ET datasets into account. We have tested that the spread between individual ~~temperature-ET~~ datasets is substantially ~~smaller than the differences between individual ET products. Therefore, larger than the spread between individual T datasets (not shown). This indicates that the largest source of uncertainty stems from the choice of ET dataset, and therefore we consider only three different T datasets.~~ Each of the 54 T-ET ~~coincidence datasets-dataset combinations~~ (denoted as "T-ET coupling benchmarks" in the remainder of the paper) ~~represent a relatively is consistently derived from observations, and thus can be expected to represent relevant features in T-ET coupling under different assumptions that underlie diagnostic datasets, reanalyses and land surface models. Therefore, these datasets represent a very large spread of plausible T-ET coupling estimates, but and the spread can be considered as a conservative benchmark for model evaluation (including observational noise, i.e. allowing a wide range of T-ET coupling in models).~~ However, it should be emphasized that the datasets are not independent realizations. Thus, we only use the spread of ~~this observations-based ensemble of the~~ T-ET ~~datasets as a measure of uncertainty~~coupling benchmarks, but we do not interpret the probability distribution of dataset combinations.

For the analysis of historical and future simulations of the monthly maximum value of daily maximum temperatures (TX<sub>x</sub>) in Section 3.2 we use ERA-Interim (Dee et al., 2011) as a reference dataset.

#### 30 Multi-model ensemble simulations

The Climate Model Intercomparison Project (CMIP5) has been designed to allow for multi-model comparison and evaluation studies (Taylor et al., 2012). Although large model spread, biases and uncertainties remain in the ensemble projections (Knutti



and Sedláček, 2013), for example with respect to extremes (Sillmann et al., 2013a), the water (Mueller et al., 2011b; Mueller and Seneviratne, 2014), and land carbon cycle (Anav et al., 2013), the archive of standardized scenario-driven model experiments provides one of the main avenues to study climate variability and change (e.g. (Stocker et al., 2013)), including present and future climate extremes (Sillmann et al., 2013b; Seneviratne et al., 2016). We use one ensemble member from 37 individual models or model variants (Table S1) ~~.-We have tested that to avoid unequal sample sizes in the multi-model ensembles.~~ Furthermore, this choice is made to assess variability in land-atmosphere coupling across models, because individual ensemble members from the same model ~~tend to show a show~~ comparably small spread in ~~VAC-coupling, indicating land-atmosphere coupling and present-day and future land-atmosphere coupling are highly correlated (Fig. S1, metric and definition is provided below).~~ This indicates that the large spread ~~across models likely arises from differences in model structure between models is~~ dominated by variability across models, and thus land-atmosphere coupling is a model-inherent feature on climatological time scales (Fig. S1 and Fig. S2, see further discussion below). On shorter (e.g. annual or seasonal) time scales, models indeed show substantial variability in their land-atmosphere coupling (Sippel et al., 2016b), which could be used as a constraint in large single-model ensembles but is beyond the scope of the present study.

## Data processing and analysis

All datasets were remapped to a common  $2.5^\circ \times 2.5^\circ$  spatial resolution for analysis and before computing T-ET coincidences. For model evaluation (Section 3.1), all computations and analyses are performed on a monthly temporal resolution and are restricted to the time period 1989-2005 due to data availability constraints of the ET reference datasets (Mueller et al., 2013). Thus, the reference period for model evaluation corresponds to the last 17 years of the "historical" scenario in CMIP5 models. T-ET coincidences are computed based on monthly deseasonalized and linearly detrended time series of T and ET, and coincidence rates are calculated separately for each individual season. Only land pixels outside of desert regions following the Köppen-Geiger climate classification are considered (Kottek et al., 2006). The model evaluation is conducted based on all individual pixels, and additionally on area-averages for so-called IPCC-SREX regions (IPCC, 2012).

## 2.2 Diagnostic-based model evaluation using T-ET coupling

### The T-ET link and the Vegetation-Atmosphere Coupling (VAC) Index

An adequate characterization of the coupling between soil moisture and temperature is key to model evaluation using observations-based datasets, ~~and the latter.~~ This coupling is often diagnosed by correlation-based metrics such as for example ~~the Pearson correlation~~ between T and ET,  $\rho_{(T,ET)}$  (Seneviratne et al., 2006; Lorenz et al., 2012), ~~or the difference in the covariability of temperature and sensible heat, where the latter is calculated with and without accounting for soil moisture deficits (Miralles et al., 2012).~~ Here, we aim to exploit the T-ET coupling by using a natural extension of  $\rho_{(T,ET)}$  that focusses on the tails of T-ET dependencies. Deseasonalized and detrended time series of ET ( $x_i^{ET}$ ) and T ( $x_i^T$ , ~~with  $i$  and  $N$  denoting denotes~~ the time step ~~and time series length, respectively~~), are partitioned into five distinct classes of Vegetation-Atmosphere Coupling (VAC) following

(Zscheischler et al., 2015), resulting in a time series of discrete events  $x_i^{VAC}$ :

$$x_i^{VAC} = \begin{cases} a, & \text{if } x_i^T < th_{lower}^T \text{ and } x_i^{ET} < th_{lower}^{ET}, \\ b, & \text{if } x_i^T > th_{upper}^T \text{ and } x_i^{ET} > th_{upper}^{ET}, \\ c, & \text{if } x_i^T > th_{upper}^T \text{ and } x_i^{ET} < th_{lower}^{ET}, \\ d, & \text{if } x_i^T < th_{lower}^T \text{ and } x_i^{ET} > th_{upper}^{ET}, \\ 0 & \text{otherwise.} \end{cases}$$

Event thresholds  $th_{lower}$  and  $th_{upper}$  might be chosen relative to the variability of each time series by fixing the probability  $p$  to exceed or fall below a threshold through the choice of an appropriate quantile:

$$5 \quad Pr[X > th_{upper}] = Pr[X \leq th_{lower}] = p \quad (1)$$

Taking time series length restrictions into account, we choose the 30th and 70th percentile as lower and upper thresholds in all time series (i.e. such that  $Pr[X \leq th_{lower}] = Pr[X > th_{upper}] = 0.3$   $Pr[X < th_{lower}] = Pr[X > th_{upper}] = 0.3$ ). Here, we focus on coincidences of *warm temperature anomalies* ("T-events":  $x_i^T > th_{upper}^T$ ) with anomalies in ET ("ET-events", i.e. either  $x_i^{ET} > th_{upper}^{ET}$  for  $VAC_b$  or  $x_i^{ET} < th_{lower}^{ET}$  for  $VAC_c$ ), i.e. we derive coincidence rates  $r_{VAC_b}$  by counting the number of  $VAC_b$ -events (see Quiroga et al., 2002; Donges et al., 2016, for earlier formulations of event coincidence analysis); (see Quiroga et al. (2002) for earlier formulations of event coincidence analysis, and e.g. Rammig et al. (2014); Siegmund et al. (2016) for applications in an ecological context):

$$r_{VAC_b} = \frac{1}{N_0} \sum_{i=1}^N 1_{[VAC_b][b]}(x_i^{VAC})$$

Here,  $1_A(x)$  is the indicator function, defined as  $1_A(x) = 1$  if  $x \in A$  and  $1_A(x) = 0$  otherwise.  $N$  denotes the length of the time series. Hence, we simply count coincidences of T and ET in a given category (e.g. positive T and positive ET for  $VAC_b$ ) to get the average coincidence rate ( $r_{VAC_b}$ ).  $N_0$  acts as a normalization constant and is chosen in our study such that  $0 \leq r_{VAC_b} \leq 1$ , i.e. we normalize with the total number of "T-events",  $N_0 = \sum_{i=1}^N 1_{[x^T > th_{upper}^T]}(x_i^T)$ . In other words, Hence, if all (or none) of the "T-events" in the time series would coincide with "ET-events" (defined for  $VAC_b$ ), then the average coincidence rates would be given by  $r_{VAC_b} = 1$  (or  $r_{VAC_b} = 0$ ). For independent time series, i.e. no coupling,  $r_{VAC_b}$  would approximate the occurrence rate of "ET-events" in the time series (defined for  $VAC_b$ ) that is governed by the chosen threshold, i.e.  $r_{VAC_b} = \frac{1}{N} \sum_{i=1}^N 1_{[x_i^{ET} > th_{upper}^{ET}]}(x_i^{ET})$  (hence,  $r_{VAC_b} \approx 0.3$  in our case). Coincidence rates  $r_{VAC_c}$  follow equivalently by replacing  $VAC_b$  with  $VAC_c$  and in the definition of "ET-events" in the previous description. We compute  $r_{VAC_b}$  and  $r_{VAC_c}$  for all seasons but with an emphasis on the warmest periods-season of the year. In this study, significance of coincidence rates is established by randomly permuting one time series with respect to the other 100 times. Hence, VAC-rates from models or observations-based benchmarks that fall outside the 5th to 95th percentile range of the VAC-rates obtained from randomly permuted time series are significantly different from independent data at the 0.1 level.

In other words,  $r_{VAC_b}$  gives the fraction of the highest 30% of temperatures that coincide with the highest 30% of ET (i.e., occurrence rate of 'energy-limited regimes'), while  $r_{VAC_c}$  denotes the fraction of the highest 30% temperatures that correspond

with the lowest 30% ET (i.e., occurrence rate of ‘water-limited regimes’). Fig. 1c-d shows a simple example of monthly time series of T and ET simulated from two CMIP5 models and occurrences of  $VAC_b$  and  $VAC_c$  are highlighted, and Fig. 1e-f shows the correlation of T and ET. Note that for the same location (area-averaged region (area-average over Central Europe, CEU) ; and occurrences of  $VAC_b$  and time of the year (monthly data for June, July, and August), one model produces predominantly energy-limited regimes ( $VAC_b$ , Fig. 1c,e and compare to conceptual illustration in Fig. 1a), whereas the other model produces predominantly water-limited regimes ( $VAC_c$  are highlighted. Please note that event coincidence analyses are frequently applied in the context of ecosystem science (e.g. Rammig et al., 2014; Siegmund et al., 2016) -, Fig. 1d,f and concept in Fig. 1b).

We abbreviate the average occurrence rates  $r_{VAC_b}$  and  $r_{VAC_c}$  as  $VAC_b$  and  $VAC_c$  for convenience in the remainder of the paper. In comparison to more traditional coupling metrics, such as e.g.  $\rho_{(T,ET)}$ ,  $VAC_{a-d}$  might be expected to yield similar results on very long time scales, whereas on shorter time scales the  $VAC_{a-d}$  index might pick  $VAC$  index picks up non-linearities in the tails (e.g. during warm temperature anomalies). We note that on  $\Delta t$  the monthly time scale (as used in the present this study),  $VAC_b$  and  $VAC_c$  detect distinct non-linearities are detected in models and observations in summer T-ET coupling e.g. in Central Europe, where a larger number of CEU: Fig. S3 shows that, by correlating  $VAC_b$  with  $VAC_c$  events occurs than derived from individual models, observations-based benchmarks, and from a two-dimensional Gaussian distribution,  $VAC_b$  and  $VAC_c$  rates in models and observations-based benchmarks exceed those that would be inferred from a correlation-based metric  $\rho_{(T,ET)}$  (expected in random data. This deviation indicates that the warm tail is indeed different to the remainder of the distribution (we observe no such deviation for the cold tail, Fig. S1). However,  $\rho_{(T,ET)}$  yields qualitatively similar results S3), and hence an evaluation metric that focuses on the tail such as the  $VAC$  index is indeed useful for our present purpose. In addition to the main text, the model evaluation is presented for a 90th percentile threshold, and for  $\rho_{(T,ET)}$  to demonstrate robustness to the chosen methodological approach (cf. Fig. S4), and for the  $VAC$  index using a 90th percentile threshold (Fig. S2)-S4). Both alternatives show qualitatively similar results (see Results and Discussion section).

### A constraint on T-ET coupling in multi-model ensembles

In general, a constraint links an observations-based diagnostic with a key model output variable across multiple models (Cox et al., 2013), and thus can be used to reduce model uncertainties and spread. Here, we derive a T-ET coupling constraint as the uncertainty range from the 54 combinations of T-ET benchmarking datasets. A Gaussian kernel with reliable data-based bandwidth selection (Sheather and Jones, 1991) is fitted over all 54 1989-2005 coincidence rates ( $r_{VAC_c}$ ) for each meteorological season and pixel (and each SREX region average). Throughout this paper, the 5th to 95th percentile range of the fitted Gaussian kernels is taken as the plausible range of observations, and the reduced (constrained) ensemble of CMIP5 simulations is obtained by retaining only those CMIP5 models that simulate T-ET coincidences that fall within the this range of observational uncertainty.

### 3 Results and Discussion

In this section, we first evaluate land-coupling in CMIP5 models explicitly against an observations-based ensemble of T-ET combinations and explore the link to temperature variability and extremes (Section 3.1). All model evaluation results are presented globally and exemplarily for Central Europe (CEU) as a region where global models and observations differ widely. Subsequently, we constrain the ensemble of CMIP5 models using each model's land-coupling as diagnosed through the  $VAC_c$  index and discuss implications for biases in simulated present-day temperature extremes and warming projections (Section 3.2).

#### 3.1 Evaluation of land-atmosphere coupling in CMIP5 models and the link to temperature variability and extremes

##### Evaluation of T-ET coupling in CMIP5 models.

Models and observations-based datasets show a relatively large spread in their representation of T-ET coupling, as expressed exemplarily in ~~Central Europe through both  $r_{VAC_b}$  and  $r_{VAC_c}$~~  CEU through both  $VAC_b$  and  $VAC_c$  across various seasons (Fig. 2, ~~top a,b~~) or diagnosed through more traditional coupling metrics such as  $\rho_{(T,ET)}$  (~~Appendix B Fig. S4~~). Individual models indicate pronounced qualitative differences in the warm season, where some models point to energy-limited, whereas others ~~point to indicate~~ predominantly water-limited conditions (Fig. 2a,b, ~~top~~, and Fig. 1, for an illustrative example). Observations-based T-ET datasets agree qualitatively, i.e., indicating energy-limited to neutral conditions in the ~~Central European CEU~~ example, thus implying an overestimation of water-limited regimes in ~~Central Europe CEU~~ in roughly 50% of CMIP5 models (Fig. 2).

This pattern holds across most regions of the globe, as many CMIP5 models consistently overestimate occurrences of  $VAC_c$  regimes (and correspondingly underestimate  $VAC_b$  occurrences) in the warm season of the year (Fig. 2, ~~c,d~~, see Fig. ~~S2-S5~~ for a definition of the warm season in each pixel). In mid-latitude and several tropical regions (e.g. Central North America, Central Europe, the Amazon, India, parts of Africa), more than 25% and up to 50% of CMIP5 models lie outside the observational range ~~(Fig. 2d)~~. These discrepancies hold also if metrics that emphasize the whole distribution ( $\rho_{(T,ET)}$ ) or more extreme parts of the tail (VAC based on a 90th percentile threshold) are used for model evaluation (~~Figs. S3-S5~~ Fig. S4, results for individual seasons are presented for  $VAC_c$  and  $VAC_b$  in Fig. S6 and Fig. S7, respectively). Moreover, the spread between the individual models' representation of land-atmosphere coupling strongly exceeds the spread in observational datasets, although different diagnostic, reanalyses and land surface model datasets are included in the observations-based ensemble (Fig. 2e for CMIP5 model spread and Fig. 2f for spread in observations-based benchmark datasets).

Furthermore, the models' land-atmosphere coupling, as diagnosed here through the VAC-index, is a highly model-inherent feature, as different model variants or ensemble members from the same model generally ~~tend to lie~~ relatively close to each other (Figs. ~~S6-S7~~ S1-S2). However, model-specific signatures of model output are not unusual, as diagnosed before ~~e.g.~~ for spatial patterns of temperature and precipitation (Knutti et al., 2013) or the statistical information content in carbon fluxes (Sipfel et al., 2016a). Furthermore, present-day land-atmosphere coupling is strongly related to future land-atmosphere coupling in the individual models (Fig. ~~S6~~ S1). A ~~detailed~~ detailed overview of  $VAC_c$  coupling in individual models and ensemble mem-

bers relative to the benchmark datasets for Central Europe and Central North America is presented in Fig. S6-S7S1-S2. Despite regionally pronounced qualitative discrepancies, it should be noted that on a global scale, the distribution of water-limited and energy-limited patterns in models and observations agrees qualitatively (Fig. S8). Likewise, the findings of climatologically too pronounced water-limited regimes in individual models w.r.t. observations does not exclude the possibility of future changes in the coupling strength in transitional regions (Seneviratne et al., 2006) or of strong water limitations during extreme events in the real world (Miralles et al., 2012; Whan et al., 2015)~~or possible future changes of the coupling strength. Further, we~~. To this end, an evaluation of the year-to-year variability of the coupling behaviour in larger ensembles of individual models, including very rare events, could constitute a topic for further study, as this study was restricted to relatively moderate events in a 16 year period (70th percentile threshold for the computation of the VAC-index) and one ensemble member per model. Besides, we also note that observations-based benchmark datasets ~~also~~ show systematic (albeit smaller) differences in the representation of land-atmosphere coupling: Diagnostic datasets indicate more frequent energy-limited regimes (see e.g. Fig. 2), and thus differ consistently to generally drier land surface models and reanalysis products, consistent with earlier findings (Santanello et al., 2015).

#### **T-ET coincidences and the link to temperature variability and extremes.**

15 The representation of T-ET coupling as diagnosed through the VAC index largely determines the variability of temperatures at monthly and inter-annual time scales across the CMIP5 multi-model ensemble in ~~Central Europe~~ CEU (Fig. 3a) and in most regions of the globe except in some subarctic climates (Fig. 3b). Therefore, this relationship is indicative for the strong influence of land-atmosphere coupling on surface climate. This is consistent with previous findings in Europe in models with and without land-atmosphere interactions (Seneviratne et al., 2006; Fischer and Schär, 2009; Fischer et al., 2012). An important result is  
20 that models that produce ~~VAC<sub>e</sub>~~ VAC<sub>c</sub> indices within the range of benchmark datasets also produce a realistic near surface temperature variability, whereas models that fall too frequently in water-limited regimes also overestimate summer temperature variability (Fig. 3a). Moreover, in mid-latitude and tropical regions, the state of the land surface is strongly associated with the mean and variability of temperature extremes at the daily time scale in the warmest season (TXx, Fig. 3c,d). The link  
25 between between the representation of land-atmosphere coupling and simulated temperature extremes and variability in global climate models is consistent with earlier studies, which has been demonstrated for Europe in individual models (Seneviratne et al., 2006; Lorenz et al., 2012; Davin et al., 2016) and in ensembles of regional models (Fischer et al., 2012; Bellprat et al., 2013). Therefore, the relationship between T-ET coincidence rates and temperature extremes might offer an avenue to derive an explicit land-atmosphere coupling constraint "~~"~~(the likely root cause for biases) to alleviate biases in temperature variability and extremes in the multi-model CMIP5 ensemble.

### 3.2 Analysis of constrained multi-model ensemble and implications for future climate projections

#### A constraint on land-atmosphere coupling in the CMIP5 ensemble.

The association between T and ET in the constrained ensemble resembles the observational datasets in land-atmosphere coupling (observations-based benchmarking datasets in T-ET coupling very well (shown as a bivariate density estimate in Fig. 4a-b, Appendix S9-S11 for details), and a corresponding improvement in the representation of temperature extremes at the daily time scale would be expected due for CEU and CNA, respectively), whereas the unconstrained CMIP5 ensemble produces too many occurrences of  $VAC_c$  conditions in both CEU and CNA. Due to the intimate link between land-atmosphere coupling and temperature variability and extremes (see previous Section), we expect that the improvement in the representation of land-atmosphere coupling in the constrained ensembles yields a corresponding improvement in the representation of temperature extremes at the daily time scale in coupling-sensitive regions.

Coupling-sensitive regions are prone to warm season biases in climate models (Christensen and Boberg, 2012; Bellprat et al., 2013). In the present analysis, high biases in temperature extremes are indeed prevalent in the original (unconstrained) CMIP5 ensemble in these regions (Fig. 4c,e). For example, the ensemble mean warm season TXx is overestimated by up to 5°C, and higher biases are detected in the 90th percentile of TXx in Central North America, Central Europe CNA, CEU or the Amazon (all biases in daily variables relative to ERA-Interim, see Fig. 4c,e). In a CMIP5 ensemble constrained by the land-atmosphere coupling metric  $VAC_c$ , the representation of temperature extremes is substantially improved in regions prone to coupling-induced biases (Fig. 4d,f), i.e. both mean TXx and the 90th percentile of TXx are significantly reduced. The ensemble mean of present-day temperature extremes in other regions remains unchanged. Moreover, projected future temperature extremes are reduced in the constrained ensemble (Fig. 5), similarly to present-day reductions in regions prone to present-day biases in land-atmosphere coupling. This is illustrated in Fig. 5a for TXx (monthly area-averages in summer) in CEU, where the hot end of the original model ensemble is in fact never realised in observed temperatures. The application of the constraint thus not only affects mean TXx, but also reduces the spread of the model ensemble (Fig. 5a,b). The reduction in ensemble mean and ensemble spread is retained for the entire 21st century (Fig. 5a,b). Hence, this result reinforces that coupling-related biases are model-inherent features, i.e. models that simulate too many  $VAC_c$ -occurrences today (and associated high biases in extreme temperatures) are very likely to do so in the future. However, one should keep in mind that the reduction in ensemble mean and spread is confined to coupling-sensitive regions in CEU, CNA, and to some degree in the Amazon region (Fig. 5c,d).

Our results imply that an accurate representation of land surface processes is crucially relevant for a correct simulation of temperature extremes, and more generally for simulated near-surface climate variability. Land-atmosphere coupling is thus an important source of bias in state-of-the-art global climate model simulations. By using an observations-based land-atmosphere coupling diagnostic to constrain the multi-model CMIP5 ensemble, we have shown that biases in extremes in the large ensemble can be alleviated to a certain degree. As bias correction methodologies that take the physical causes for biases into account are still widely lacking (Ehret et al., 2012; Bellprat et al., 2013) and multivariate bias correction methods are currently in development (Cannon, 2016), the identification of models with a *physically plausible* representation of near-surface climate and land-atmosphere interactions at the regional scale might be crucial to extract accurate and relevant information about

climate extremes in the context of climatic changes in the 21st century (Mitchell et al., 2016b; Schleussner et al., 2016; Seneviratne et al., 2016). For example, model selection for event attribution studies or a quantification of changes in univariate climate extremes is often based on a statistical performance criterion (Perkins et al., 2007; King et al., 2016; Otto et al., 2015). Our results indicate that these procedures could be further refined through incorporating observations-based diagnostics or constraints in order to analyse model simulations that are indeed "right for the right reasons" (at least given physics-guided and observations-based relationships). Moreover, the impacts of climate and its extremes e.g. on human health or ecosystems (Mitchell et al., 2016a; Frank et al., 2015) are often inherently related to multiple climate variables (Ehret et al., 2012; Leonard et al., 2014). Therefore, simple constraints as motivated for instance in the present study might complement more conventional bias correction procedures (e.g. Hempel et al., 2013) to derive physically consistent estimates of climate impacts. This approach appears promising, because biases within climate models (i.e. in different variables) and across climate model ensembles are often correlated (e.g. Knutti, 2010; Mueller and Seneviratne, 2014; Sippel et al., 2016b). ~~Fig. S12 indicates that~~ Hence, beyond soil moisture control on simulated temperature extremes as the present study's focus, related biases in other variables such as warm season precipitation or ET might be similarly relevant in this context. For example,  $VAC_c$  occurrences across the CMIP5 ensemble are negatively associated with precipitation and evapotranspiration-ET in the warm season in mid-latitude regions (Fig. S9) - both crucial variables in the water cycle that show pronounced summer low biases in CMIP5 models (Mueller and Seneviratne, 2014). Therefore, a constrained model ensemble with improved land-atmosphere coupling, a likely root cause of biases (Lorenz et al., 2012), might not only improve temperature extremes and variability, but additionally might reduce biases in associated variables such as ~~evapotranspiration or rainfall~~ ET or precipitation.

### **Is there a link between present-day land-atmosphere coupling and warming projections?**

We investigate whether the representation of land-atmosphere coupling in climate models affects the magnitude of 21st century warming (e.g. Fischer et al., 2012; Stegehuis et al., 2013). We first note that regions sensitive to land-atmosphere coupling in the CMIP5 model ensemble also show relatively strong warming in daily-scale temperature extremes (TXx), for example Central America or South and Central Europe (Fig. 6, ~~top a,b~~). More importantly, however, models that produce frequent  $VAC_c$  occurrences (water-limited regimes) tend to be associated with larger rates of warming in TXx, although it should be emphasized that this relationship is not simple or linear (~~middle panel in~~ Fig. 6, ~~Fig. S13, e.g. c,d, see also~~ Fischer et al. (2012)). Conversely, this pattern reverses in boreal regions, where strongly energy-limited models (i.e. very few  $VAC_c$  occurrences) tend to produce larger warming. However, in boreal regions this apparent relationship likely stems from a spurious correlation with the individual models' background warming (i.e., warming in annual averages), as the correlation in fact disappears if the background warming is subtracted from summer warming (Fig. ~~S13~~ S10). In contrast, in mid-latitude regions warm season warming that exceeds annual average warming remains confined to the warm season (~~Fig. S13~~). A multi-model projection constrained by a plausible representation of land-atmosphere coupling reduces ~~future differences in~~ TXx estimates in a future climate relative to the present in coupling-sensitive regions such as Central Europe and Central North America by ~~up to~~ locally by around 0.5°C to 1°C ~~but this remains a regional effect (Fig. 6e,f)~~. These results are consistent with earlier

studies that used an ensemble of regional models over Europe that used the standard deviation of temperatures as a constraint (Fischer et al., 2012).



## 4 Conclusions

In the present study, we have evaluated land-atmosphere coupling in state-of-the-art climate models with an ensemble of observations using a diagnostic based on coincidences of large temperature and evapotranspiration anomalies-T and ET anomalies (the so called VAC index). While observations and models broadly agree on spatial patterns of land-atmosphere coupling, our results reveal that models differ widely in coupling-sensitive regions in the mid-latitudes and the tropics. Several models exhibit systematically too frequent coincidences of high temperature anomalies with negative ET anomalies (water-limited regimes) in mid-latitude regions in the warm season, and in several tropical regions year-round. Across the multi-model ensemble, we found a strong association of land-atmosphere coupling with simulated temperature variability and extremes. The spread between models largely explains differences in simulated monthly temperature variability and daily extremes. We applied a land-atmosphere coupling constraint to the multi-model ensemble, which considerably improves the representation of land-atmosphere coupling in the ensemble, and reduces biases in temperature variability and extremes in present-day simulations in a physically consistent manner ~~;~~ and (Fig. 4). Furthermore, the constraint leads to reduced variability and lower extreme temperatures in future projections. However, the overall projected changes in temperature extremes are not so strongly affected (reduction around 0.5 – 1.0°C locally in regions that are sensitive to land-atmosphere coupling), because the models with overestimated land-atmosphere coupling display similar anomalies from the multi-ensemble mean in present and future. In conclusion, we selected models with a *physically plausible* representation of land surface processes (and near-surface climate) using observations-based constraints that are guided by physical considerations. This approach complements more traditional bias correction approaches and offers new avenues to obtain improved estimates of future climate impacts.

*Acknowledgements.* We ~~appreciate the creators~~ are grateful to the creators, maintainers, and providers of all data sets, ~~including~~ CRU data were obtained from the University of East Anglia Climate Research Unit (CRU) ~~for providing temperature data~~ ( British Atmospheric Data Centre, 2008, available from <http://badc.nerc.ac.uk/data/cru/>), ~~ERA-Interim data provided courtesy ECMWF,~~ and NCEP Reanalysis data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA (). Furthermore, we acknowledge the Global Modeling and Assimilation Office and the ~~GES DISC~~ Goddard Earth Sciences (GES) Data and Information Service Center (DISC) for the dissemination of MERRA ~~and~~ MERRA-LAND, ~~and GLDAS-2 products (GL-NOAH-PF), the latter of which were acquired as part of the mission of NASA's Earth Science Division.~~ The CFSR data are from the Research Data Archive which is maintained by the Computational and Information Systems Laboratory at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. The original data are available from the RDA (<http://dss.ucar.edu>) in data set number ds093.0. We would further like to acknowledge the Japanese 25yr ReAnalysis and JMA Climate Data Assimilation System (JCDAS) for the dissemination of JRA-25 data. The LandFlux-EVAL dataset used in this article was coordinated under the Global Energy and Water Exchanges (GEWEX) LandFlux initiative. We thank the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We also thank Urs Beyerle and Jan Sedlacek for preparation and maintenance of CMIP5 data. This research has received funding from the European Commission Horizon 2020 BACI Project (Towards a Biosphere Atmosphere Change Index, grant No. 640176), and the European Research Council (ERC) under grant agreement No. 617518 (DROUGHT-HEAT). This work contributes to the World Climate Research Programme (WCRP) Grand Challenge on Extremes. S.S. is grateful to the German National Academic Foundation for PhD funding.

## References

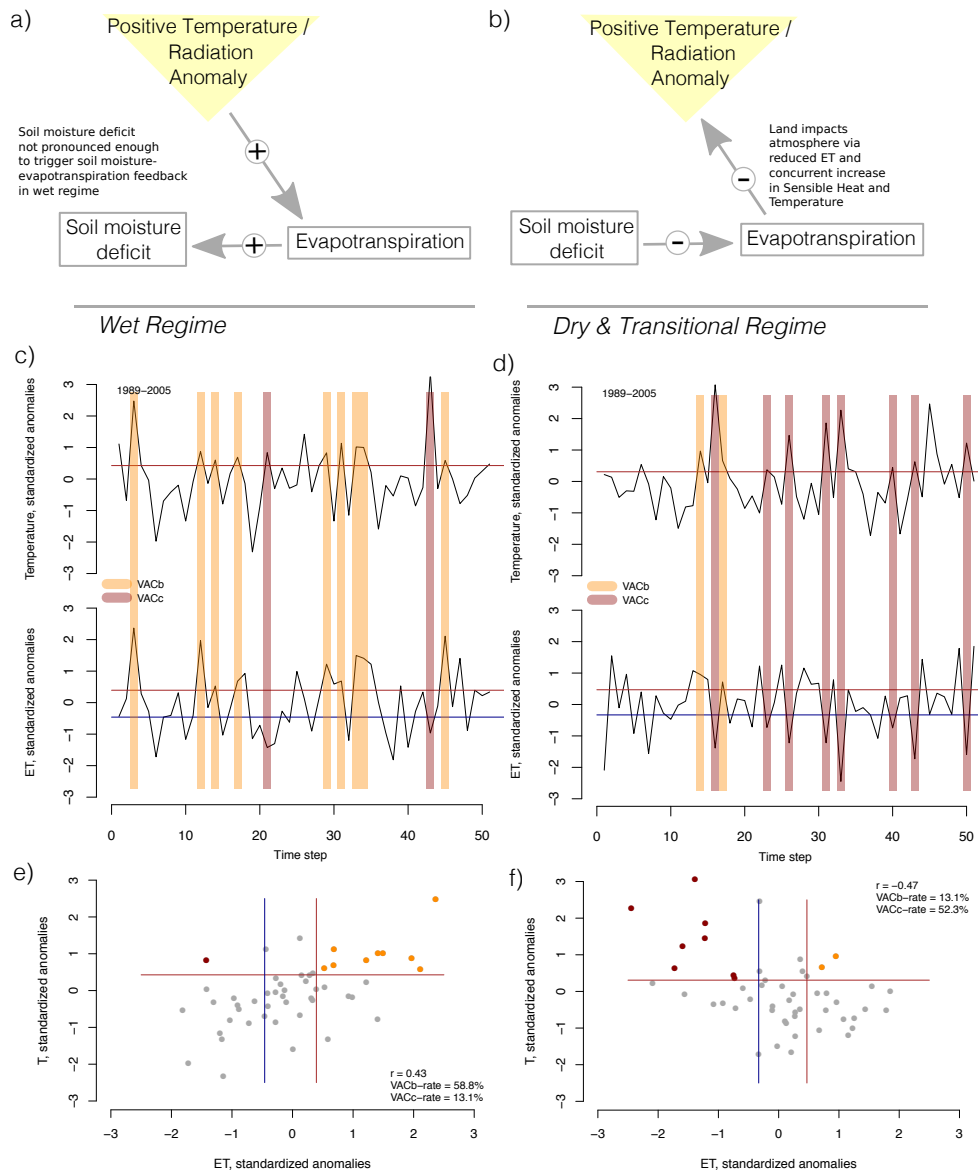
- Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R., and Zhu, Z.: Evaluating the land and ocean components of the global carbon cycle in the CMIP5 Earth System Models, *J Climate*, 26, 6801–6843, 2013.
- Bellprat, O., Kotlarski, S., Lüthi, D., and Schär, C.: Physical constraints for temperature biases in climate models, *Geophys Res Lett*, 40, 4042–4047, 2013.
- 5 Boberg, F. and Christensen, J. H.: Overestimation of Mediterranean summer temperature projections due to model deficiencies, *Nat Clim Change*, 2, 433–436, 2012.
- Boé, J. and Terray, L.: Uncertainties in summer evapotranspiration changes over Europe and implications for regional climate change, *Geophys Res Lett*, 35, 2008.
- 10 Bonan, G.: *Ecological climatology: concepts and applications*, Cambridge University Press, 2015.
- Bosilovich, M. G.: NASA's modern era retrospective-analysis for research and applications: Integrating Earth observations, *IEEE Earthzine*, 1, 82–87, 2008.
- Cannon, A. J.: *Multivariate Bias Correction of Climate Model Outputs: Matching Marginal Distributions and Inter-variable Dependence Structure*, *J Climate*, 2016.
- 15 Carvalho, N., Forkel, M., Khomik, M., Bellarby, J., Jung, M., Migliavacca, M., Mu, M., Saatchi, S., Santoro, M., Thurner, M., et al.: Global covariation of carbon turnover times with climate in terrestrial ecosystems, *Nature*, 514, 213–217, 2014.
- Christensen, J. H. and Boberg, F.: Temperature dependent climate projection deficiencies in CMIP5 models, *Geophys Res Lett*, 39, 2012.
- Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D., and Luke, C. M.: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, *Nature*, 494, 341–344, 2013.
- 20 Davin, E. L., Maisonnave, E., and Seneviratne, S. I.: Is land surface processes representation a possible weak link in current Regional Climate Models?, *Environ Res Lett*, 11, 074027, <http://stacks.iop.org/1748-9326/11/i=7/a=074027>, 2016.
- Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Quart J Roy Meteor Soc*, 137, 553–597, 2011.
- Dirmeyer, P. A.: The terrestrial segment of soil moisture–climate coupling, *Geophys Res Lett*, 38, 2011.
- 25 Dirmeyer, P. A., Jin, Y., Singh, B., and Yan, X.: Trends in land–atmosphere interactions from CMIP5 simulations, *J Hydrometeorol*, 14, 829–849, 2013.
- Donges, J., Schleussner, C.-F., Siegmund, J., and Donner, R.: Event coincidence analysis for quantifying statistical interrelationships between event time series, *The European Physical Journal Special Topics*, 225, 471–487, 2016.
- Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions "Should we apply bias correction to global and regional climate model data?", *Hydrol Earth Syst Sci*, 16, 3391–3404, 2012.
- 30 Fischer, E., Seneviratne, S., Lüthi, D., and Schär, C.: Contribution of land-atmosphere coupling to recent European summer heat waves, *Geophys Res Lett*, 34, 2007.
- Fischer, E., Rajczak, J., and Schär, C.: Changes in European summer temperature variability revisited, *Geophys Res Lett*, 39, 2012.
- Fischer, E. M. and Schär, C.: Future changes in daily summer temperature variability: driving processes and role for temperature extremes, *Climate Dyn*, 33, 917–935, 2009.
- 35 Frank, D., Reichstein, M., Bahn, M., Thonicke, K., Frank, D., Mahecha, M. D., Smith, P., Velde, M., Vicca, S., Babst, F., et al.: Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts, *Glob Chang Biol*, 21, 2861–2880, 2015.

- Grimm, V. and Railsback, S. F.: Pattern-oriented modelling: a "multi-scope" for predictive systems ecology, *Phil Trans R Soc B*, 367, 298–310, 2012.
- Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, *Geophys Res Lett*, 33, 2006.
- 5 Harris, I., Jones, P., Osborn, T., and Lister, D.: Updated high-resolution grids of monthly climatic observations—the CRU TS3. 10 Dataset, *Int J Climatol*, 34, 623–642, 2014.
- Hauser, M., Orth, R., and Seneviratne, S. I.: Role of Soil Moisture vs. Recent Climate Change for the 2010 Heat Wave in Western Russia, *Geophys Res Lett*, 2016.
- Hempel, S., Frieler, K., Warszawski, L., Schewe, J., and Piontek, F.: A trend-preserving bias correction—the ISI-MIP approach, *Earth Syst*  
 10 *Dynam*, 4, 219–236, 2013.
- Hirschi, M., Seneviratne, S. I., Alexandrov, V., Boberg, F., Boroneant, C., Christensen, O. B., Formayer, H., Orłowsky, B., and Stepanek, P.: Observational evidence for soil-moisture impact on hot extremes in southeastern Europe, *Nat Geosci*, 4, 17–21, 2011.
- IPCC: Summary for Policymakers, in: *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*, edited by Field, C., Barros, V., Stocker, T., Dahe, Q., Dokken, D., Ebi, K.,  
 15 Mastrandrea, M., Mach, K., Plattner, G., Allen, S., Tignor, M., and Midgley, P., Cambridge University Press, 2012.
- Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., et al.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, *J Geophys Res: Biogeosci*, 116, 2011.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al.: The  
 20 NCEP/NCAR 40-year reanalysis project, *Bull Amer Meteor Soc*, 77, 437–471, 1996.
- King, A. D., Black, M. T., Min, S.-K., Fischer, E. M., Mitchell, D. M., Harrington, L. J., and Perkins-Kirkpatrick, S. E.: Emergence of heat extremes attributable to anthropogenic influences, *Geophys Res Lett*, 43, 3438–3443, 2016.
- Knutti, R.: Should we believe model predictions of future climate change?, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366, 4647–4664, 2008.
- 25 Knutti, R.: The end of model democracy?, *Clim Chang*, 102, 395–404, 2010.
- Knutti, R. and Sedláček, J.: Robustness and uncertainties in the new CMIP5 climate model projections, *Nat Clim Change*, 3, 369–373, 2013.
- Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, *Geophys Res Lett*, 40, 1194–1199, 2013.
- Köppen, W.: Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt, *Geogr Z*, 6, 593–611, 1900.
- 30 Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C., Kanae, S., Kowalczyk, E., Lawrence, D., et al.: Regions of strong coupling between soil moisture and precipitation, *Science*, 305, 1138–1140, 2004.
- Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World map of the Köppen-Geiger climate classification updated, *Meteorol Z*, 15, 259–263, 2006.
- Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C.: A dynamic  
 35 global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem Cy*, 19, 2005.
- Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., and Stafford-Smith, M.: A compound event framework for understanding extreme impacts, *Wiley Interdiscip Rev Clim Change*, 5, 113–128, 2014.

- Lorenz, R., Davin, E., and Seneviratne, S.: Modeling land-climate coupling in Europe: Impact of land surface representation on climate variability and extremes, *J Geophys Res: Atmos*, 117, 2012.
- Miralles, D., De Jeu, R., Gash, J., Holmes, T., and Dolman, A.: Magnitude and variability of land evaporation and its components at the global scale, *Hydrol Earth Syst Sci*, 2011a.
- 5 Miralles, D., Holmes, T., De Jeu, R., Gash, J., Meesters, A., and Dolman, A.: Global land-surface evaporation estimated from satellite-based observations, *Hydrol Earth Syst Sci*, 15, 453–469, 2011b.
- Miralles, D., den Berg, M. v., Teuling, A., and Jeu, R. d.: Soil moisture-temperature coupling: A multiscale observational analysis, *Geophys Res Lett*, 39, 2012.
- Mitchell, D., Heaviside, C., Vardoulakis, S., Huntingford, C., Masato, G., Guillod, B. P., Frumhoff, P., Bowery, A., Wallom, D., and Allen,  
10 M.: Attributing human mortality during extreme heat waves to anthropogenic climate change, *Environ Res Lett*, 11, 074006, 2016a.
- Mitchell, D., James, R., Forster, P. M., Betts, R. A., Shiogama, H., and Allen, M.: Realizing the impacts of a 1.5°C warmer world, *Nat Clim Change*, 2016b.
- Mueller, B. and Seneviratne, S.: Systematic land climate and evapotranspiration biases in CMIP5 simulations, *Geophys Res Lett*, 41, 128–134, 2014.
- 15 Mueller, B., Hirschi, M., and Seneviratne, S. I.: New diagnostic estimates of variations in terrestrial water storage based on ERA-Interim data, *Hydrol Processes*, 25, 996–1008, 2011a.
- Mueller, B., Seneviratne, S., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P., Fisher, J., Guo, Z., et al.: Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations, *Geophys Res Lett*, 38, 2011b.
- Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P., Dolman, A., Fisher, J., Jung, M., Ludwig, F., Maignan, F., et al.: Benchmark  
20 products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, *Hydrol Earth Syst Sci*, 2013.
- Mystakidis, S., Davin, E. L., Gruber, N., and Seneviratne, S. I.: Constraining future terrestrial carbon cycle projections using observation-based water and carbon flux estimates, *Global Change Biol*, 2016.
- Onogi, K., Tsutsui, J., Koide, H., Sakamoto, M., Kobayashi, S., Hatsushika, H., Matsumoto, T., Yamazaki, N., Kamahori, H., Takahashi, K., et al.: The JRA-25 reanalysis, *J Meteorol Soc Jpn*, 85, 369–432, 2007.
- 25 Orth, R., Zscheischler, J., and Seneviratne, S. I.: Record dry summer in 2015 challenges precipitation projections in Central Europe, *Sci Rep*, 6, 2016.
- Otto, F. E., Hausteil, K., Uhe, P., Coelho, C. A., Aravequia, J. A., Almeida, W., King, A., Coughlan de Perez, E., Wada, Y., Jan van Oldenborgh, G., et al.: Factors other than climate change, main drivers of 2014/15 water shortage in southeast Brazil, *Bull Amer Meteor Soc*, 96, S35–S40, 2015.
- 30 Perkins, S., Pitman, A., Holbrook, N., and McAneney, J.: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions, *J Climate*, 20, 4356–4376, 2007.
- Quesada, B., Vautard, R., Yiou, P., Hirschi, M., and Seneviratne, S. I.: Asymmetric European summer heat predictability from wet and dry southern winters and springs, *Nat Clim Change*, 2, 736–741, 2012.
- Quiroga, R. Q., Kreuz, T., and Grassberger, P.: Event synchronization: a simple and fast method to measure synchronicity and time delay  
35 patterns, *Physical review E*, 66, 041904, <http://journals.aps.org/pre/pdf/10.1103/PhysRevE.66.041904>, 2002.
- Rammig, A., Wiedermann, M., Donges, J., Babst, F., von Bloh, W., Frank, D., Thonicke, K., and Mahecha, M.: Coincidences of climate extremes and anomalous vegetation responses: comparing tree ring patterns to simulated productivity, *Biogeosciences*, 12, 373–385, 2014.

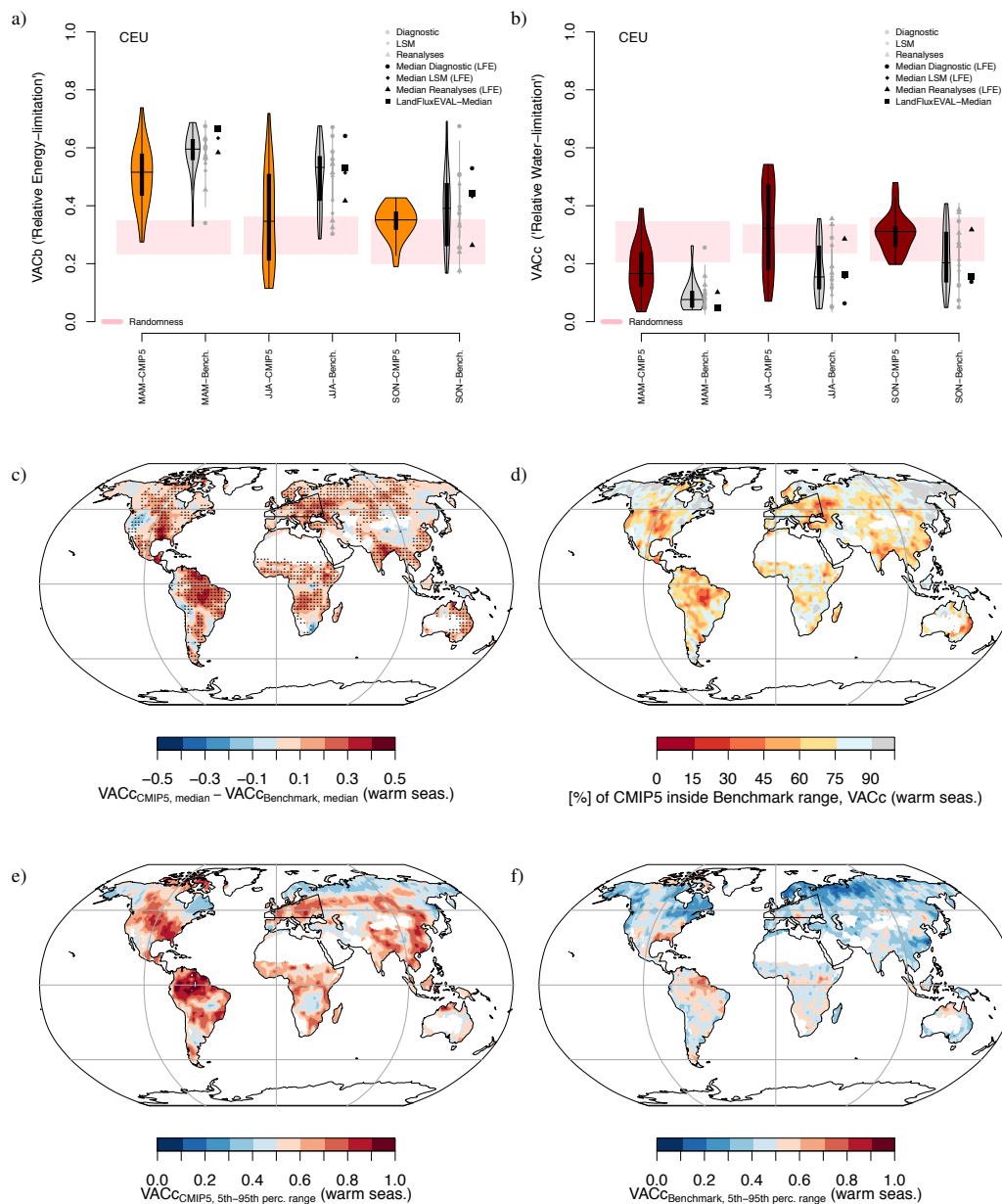
- Reichle, R. H., Koster, R. D., De Lannoy, G. J., Forman, B. A., Liu, Q., Mahanama, S. P., and Touré, A.: Assessment and enhancement of MERRA land surface hydrology estimates, *J Climate*, 24, 6322–6338, 2011.
- Reichstein, M., Mahecha, M. D., Ciais, P., Seneviratne, S. I., Blyth, E. M., Carvalhais, N., and Luo, Y.: Elk-testing climate-carbon cycle models: a case for pattern-oriented system analysis, *iLEAPS Newsletter*, 11, 14–21, 2011.
- 5 Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., et al.: The global land data assimilation system, *Bull Amer Meteor Soc*, 85, 381–394, 2004.
- Rui, H. and Beaudoin, H.: README Document for Global Land Data Assimilation System Version 2 (GLDAS-2) Products, Goddard Earth Sciences (GES) Data and Information Services Center, [https://hydro1.gesdisc.eosdis.nasa.gov/data/GLDAS/README\\_GLDAS2.pdf](https://hydro1.gesdisc.eosdis.nasa.gov/data/GLDAS/README_GLDAS2.pdf), 2016.
- 10 Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., et al.: The NCEP climate forecast system reanalysis, *Bull Amer Meteor Soc*, 91, 1015, 2010.
- Santanello, J. A., Peters-Lidard, C. D., Kumar, S. V., Alonge, C., and Tao, W.-K.: Diagnosing the local land-atmosphere coupling in models and observations, *iLEAPS Newsletter*, 9, 22–24, 2010.
- Santanello, J. A., Roundy, J., and Dirmeyer, P. A.: Quantifying the land-atmosphere coupling behavior in modern reanalysis products over  
15 the US Southern Great Plains, *J Climate*, 28, 5813–5829, 2015.
- Schleussner, C.-F., Rogelj, J., Schaeffer, M., Lissner, T., Licker, R., Fischer, E. M., Knutti, R., Levermann, A., Frieler, K., and Hare, W.: Science and policy characteristics of the Paris Agreement temperature goal, *Nat Clim Change*, 2016.
- Seneviratne, S., Corti, T., Davin, E., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A.: Climate change and soil moisture-climate interactions: Using new diagnostics to identify hot spots of land-atmosphere coupling, *iLEAPS Newsletter*, 9, 18–21, 2010a.
- 20 Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C.: Land-atmosphere coupling and climate change in Europe, *Nature*, 443, 205–209, 2006.
- Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture-climate interactions in a changing climate: A review, *Earth Sci Rev*, 99, 125–161, 2010b.
- Seneviratne, S. I., Wilhelm, M., Stanelle, T., Hurk, B., Hagemann, S., Berg, A., Cheruy, F., Higgins, M. E., Meier, A., Brovkin, V., et al.:  
25 Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment, *Geophys Res Lett*, 40, 5212–5217, 2013.
- Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., and Wilby, R. L.: Allowable CO<sub>2</sub> emissions based on regional and impact-related climate targets, *Nature*, 2016.
- Sheather, S. J. and Jones, M. C.: A reliable data-based bandwidth selection method for kernel density estimation, *J R Stat Soc Series B Stat  
30 Methodol*, 53, 683–690, 1991.
- Sheffield, J. and Wood, E. F.: Characteristics of global and regional drought, 1950–2000: Analysis of soil moisture data from off-line simulation of the terrestrial hydrologic cycle, *J Geophys Res: Atmos*, 112, 2007.
- Sheffield, J., Goteti, G., and Wood, E. F.: Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling, *J Climate*, 19, 3088–3111, 2006.
- 35 Sheffield, J., Wood, E. F., and Munoz-Arriola, F.: Long-term regional estimates of evapotranspiration for Mexico based on downscaled ISCCP data, *J Hydrometeorol*, 11, 253–275, 2010.
- Siegmund, J. F., Wiedermann, M., Donges, J. F., and Donner, R. V.: Impact of temperature and precipitation extremes on the flowering dates of four German wildlife shrub species, *Biogeosciences*, 13, 5541–5555, 2016.

- Sillmann, J., Kharin, V., Zhang, X., Zwiers, F., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, *J Geophys Res: Atmos*, 118, 1716–1733, 2013a.
- Sillmann, J., Kharin, V., Zwiers, F., Zhang, X., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections, *J Geophys Res: Atmos*, 118, 2473–2493, 2013b.
- 5 Sippel, S., Lange, H., Mahecha, M., Hauhs, M., Bodesheim, P., Kaminski, T., Gans, F., and Rosso, O.: Diagnosing the Dynamics of Observed and Simulated Ecosystem Gross Primary Productivity with Time Causal Information Theory Quantifiers, *PLoS One*, doi:doi:10.1371/journal.pone.0164960, accepted, 2016a.
- Sippel, S., Otto, F., Forkel, M., Allen, M., Guillod, B., Heimann, M., Reichstein, M., Seneviratne, S., Thonicke, K., and Mahecha, M. D.: A novel bias correction methodology for climate impact simulations, *Earth Syst Dyn*, 7, 71, 2016b.
- 10 Stegehuis, A. I., Teuling, A. J., Ciais, P., Vautard, R., and Jung, M.: Future European temperature change uncertainties reduced by using land heat flux observations, *Geophys Res Lett*, 40, 2242–2245, 2013.
- Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., et al.: *Climate Change 2013: The Physical Science Basis. Working Group 1 (WG1) Contribution to the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report (AR5)*. Cambridge, United Kingdom and New York, NY, Tech. rep., 2013.
- 15 Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, *Bull Amer Meteor Soc*, 93, 485, 2012.
- Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365, 2053–2075, 2007.
- van den Hurk, B., Kim, H., Krinner, G., Seneviratne, S. I., Derksen, C., Oki, T., Douville, H., Colin, J., Ducharne, A., Cheruy, F., Viovy, N., Puma, M. J., Wada, Y., Li, W., Jia, B., Alessandri, A., Lawrence, D. M., Weedon, G. P., Ellis, R., Hagemann, S., Mao, J., Flanner, M. G.,
- 20 Zampieri, M., Matera, S., Law, R. M., and Sheffield, J.: LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model Intercomparison Project – aims, setup and expected outcome, *Geosci Model Dev*, 9, 2809–2832, doi:10.5194/gmd-9-2809-2016, 2016.
- Van Huijgevoort, M., Van Lanen, H., Teuling, A., and Uijlenhoet, R.: Identification of changes in hydrological drought characteristics from a multi-GCM driven ensemble constrained by observed discharge, *J Hydrol*, 512, 421–434, 2014.
- 25 Wenzel, S., Cox, P. M., Eyring, V., and Friedlingstein, P.: Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models, *Journal of Geophysical Research: Biogeosciences*, 119, 794–807, 2014.
- Whan, K., Zscheischler, J., Orth, R., Shongwe, M., Rahimi, M., Asare, E. O., and Seneviratne, S. I.: Impact of soil moisture on extreme maximum temperatures in Europe, *Weather Clim Extremes*, 9, 57–67, 2015.
- Zhang, Y., Leuning, R., Hutley, L. B., Beringer, J., McHugh, I., and Walker, J. P.: Using long-term water balances to parameterize surface
- 30 conductances and calculate evaporation at 0.05 spatial resolution, *Water Resour Res*, 46, 2010.
- Zscheischler, J., Orth, R., and Seneviratne, S. I.: A submonthly database for detecting changes in vegetation-atmosphere coupling, *Geophys Res Lett*, 42, 9816–9824, 2015.

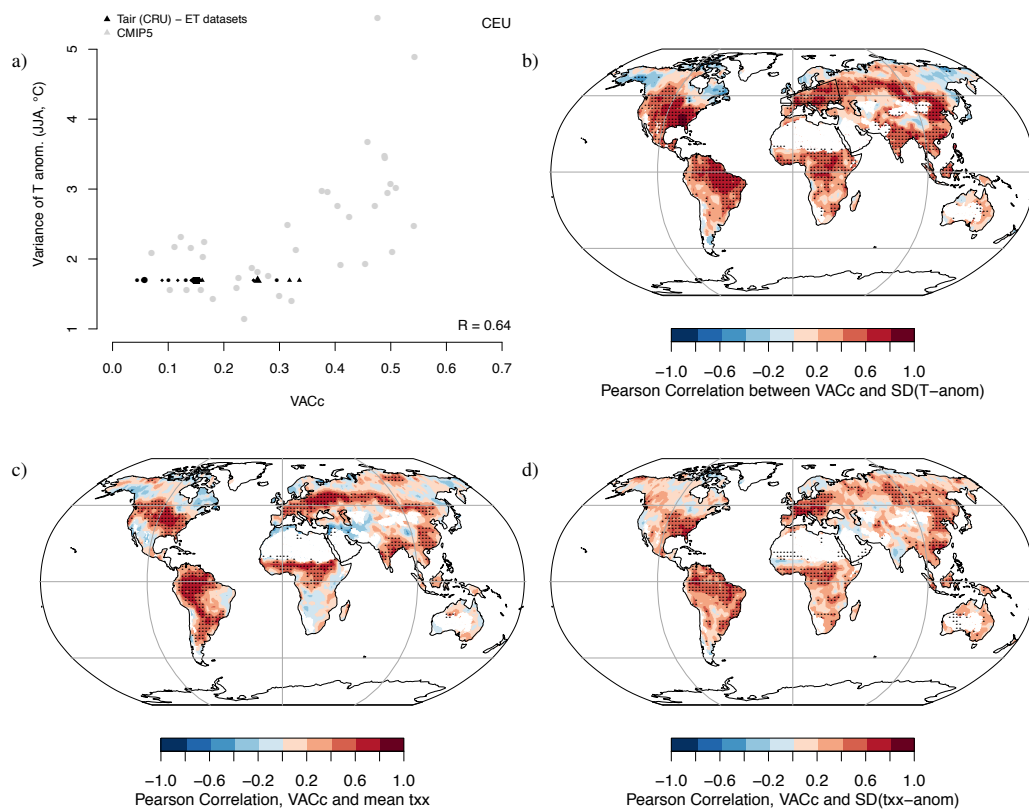


**Figure 1.** Conceptual figure illustrating of qualitatively contrasting warm season temperature-evapotranspiration (T-ET) coupling in global climate models. (a, b) Conceptual illustration of T-ET coupling in (a) wet, and (b) dry & transitional regions-regimes. In wet regimes T and ET are positively associated (atmosphere impacts land) wet-regions, where while in dry & transitional regimes T and ET are negatively associated due to soil moisture-temperature interactions play contrasting roles moisture feedbacks (i.e., land impacts atmosphere via reduced ET and concurrent increases in sensible heat and T). (c-f) Contrasting-Different CMIP5 models show contrasting T-ET coupling behaviour in a coupling-sensitive mid-latitude region in summer (Central Europe, spatial average, JJA, 1989-2005) in two different CMIP5 models: (left, c, e) NorESM1-M produces predominantly wet regime: NorESM1-M; right regimes, i.e. a positive T-ET coupling, while (d, f) ACCESS1-3 produces predominantly dry regime: ACCESS1-3 regimes (negative T-ET coupling), illustrated as time series (c-d) and in the T-ET plane (e-f). Red lines in (c-f) indicate  $th_{upper}^T$  for T and ET, blue lines indicate  $th_{lower}^T$  (70th and 30th percentile in each individual time series, respectively).

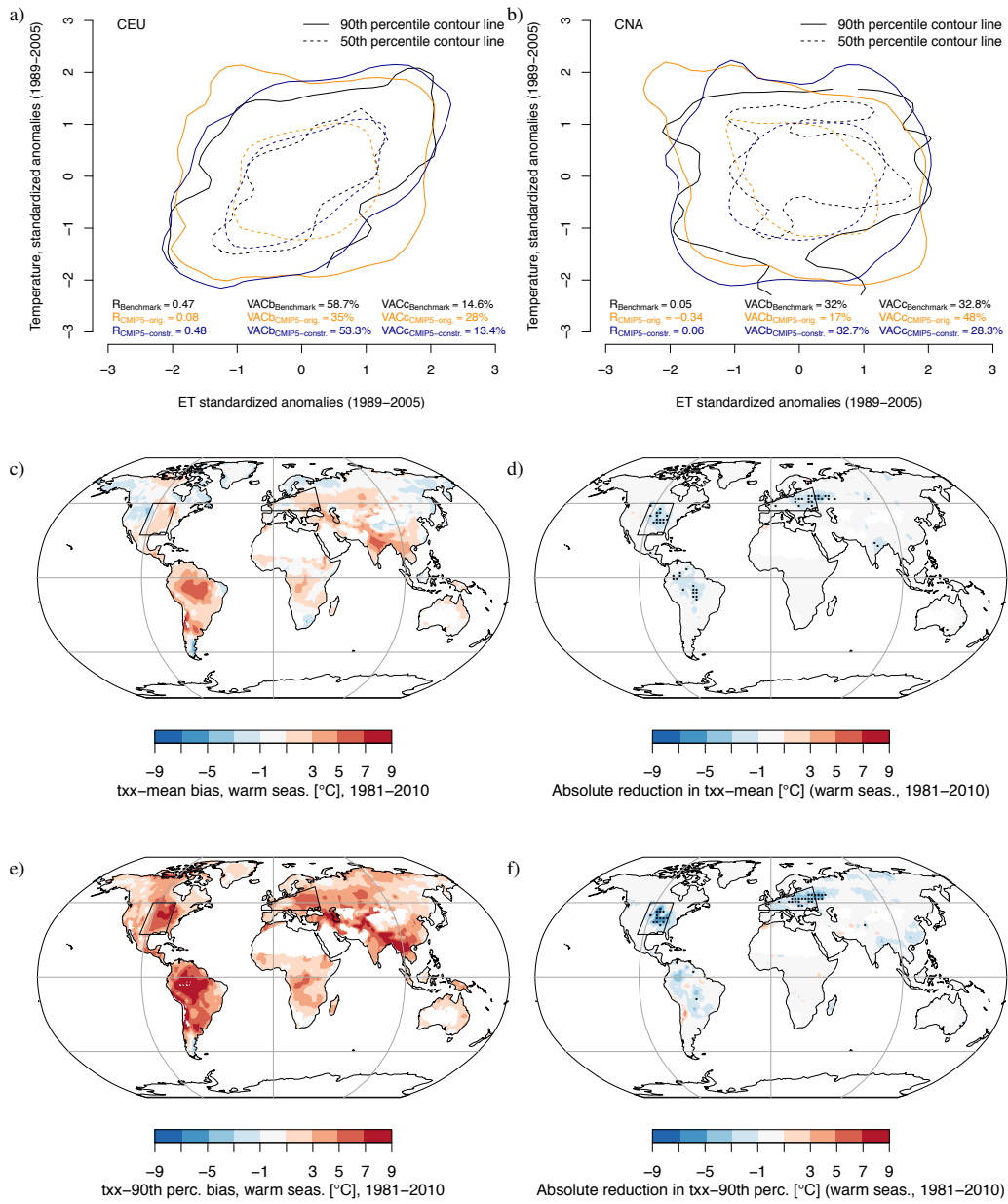




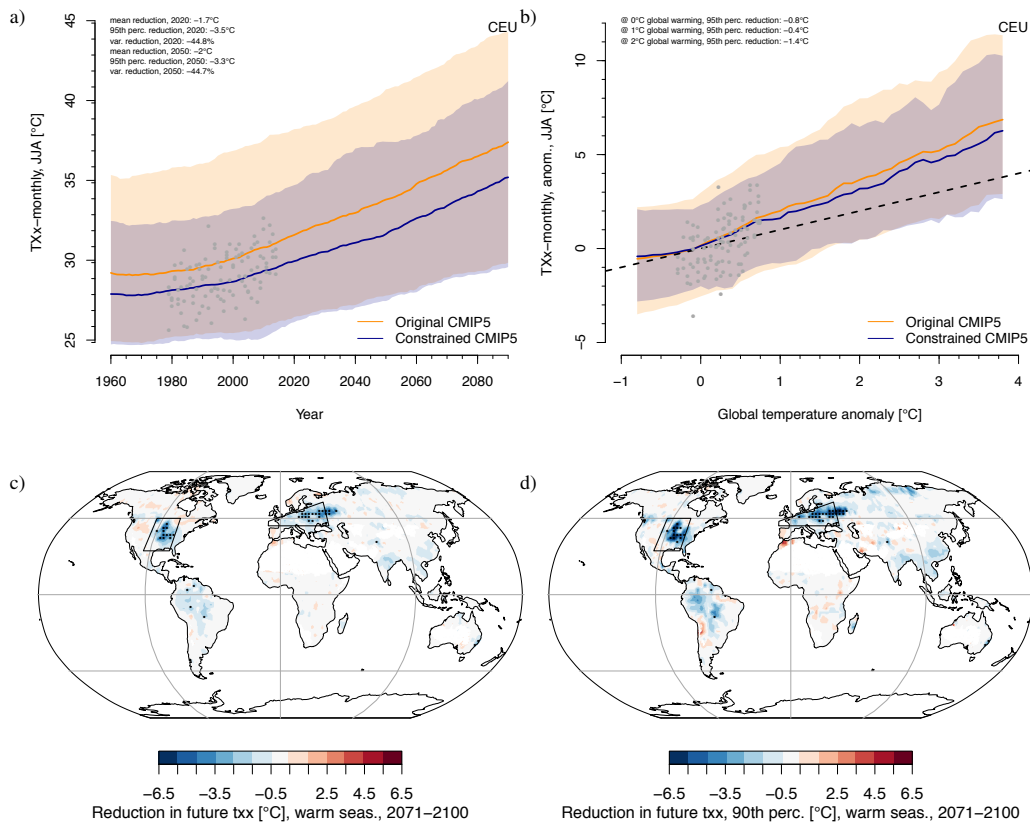
**Figure 2.** Evaluation of T-ET coupling in global climate models. (a, b) VACb (+) and VACc coupling in the CMIP5 climate model ensemble and observations-based benchmarking datasets in Central Europe (CEU, 1989-2005, area-average) with systematic warm season differences (circles, diamonds, and triangles indicate diagnostic, land surface models, and reanalyses reference datasets, respectively). Randomness indicates the 5th to 95th percentile range obtained by randomly permutating both time series with respect to the other ( $N = 100$  times) to obtain independent data. (c) Difference in the VACc median of the CMIP5 ensemble and benchmarking datasets. (d) Fraction of CMIP5 models that are inside the 5th-95th percentile spread of the benchmarking datasets. (e, f) Range of VACc-occurrences (5th to 95th percentile range) in CMIP5 models (e) and in the ensemble of observations (f). **22**



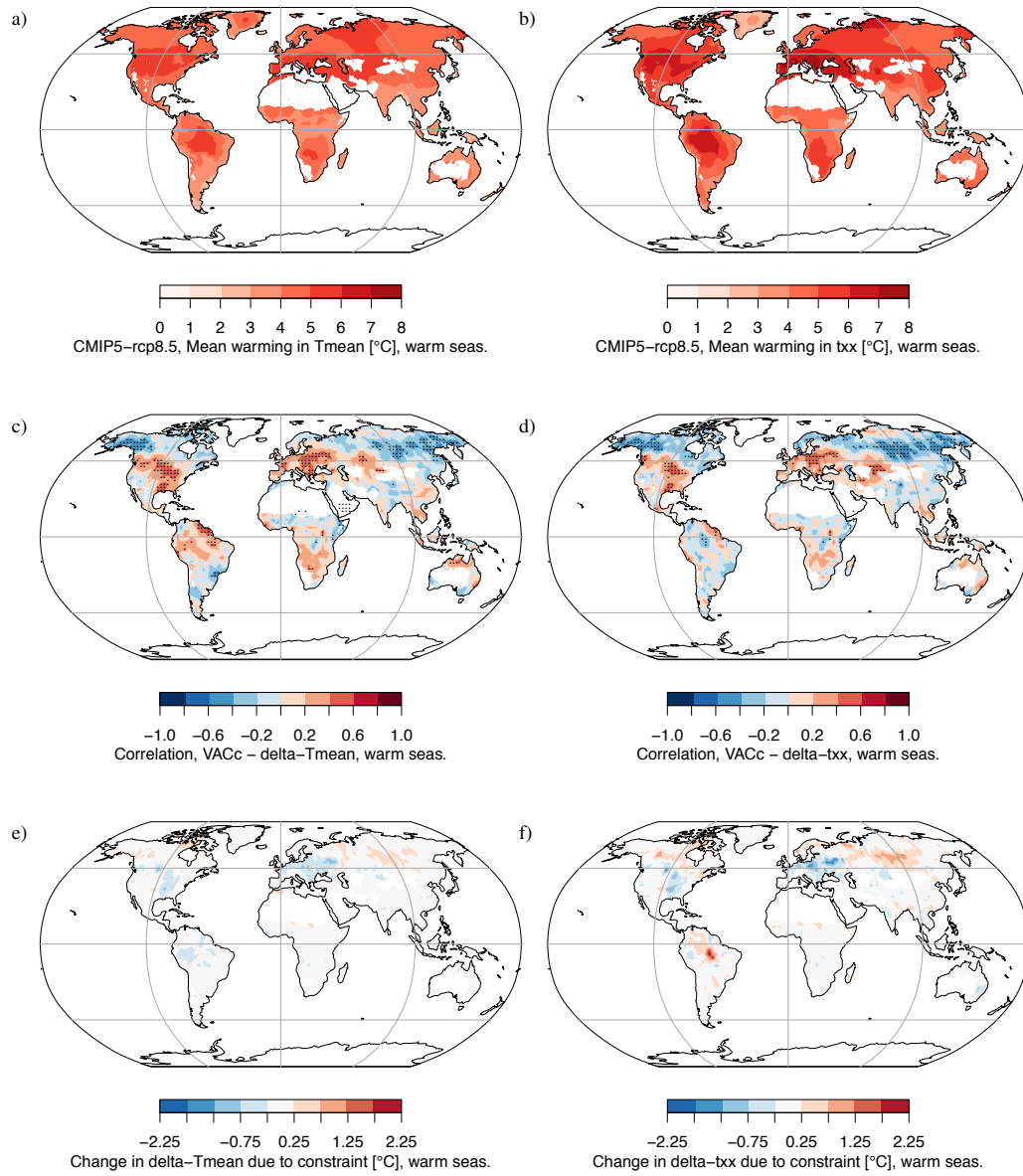
**Figure 3.** (a, b) Relationship between model-specific T-ET coupling (expressed through VACc) and model simulated variability of monthly temperature anomalies (JJA) in Central Europe (a), and globally (b). (c, d) Relationship between VACc-coupling and mean (c) and standard deviation (d) of simulated monthly maximum value of daily maximum temperature (TXx) in summer (JJA).



**Figure 4.** (a-b) Contour lines of bivariate kernel density estimates of T-ET relationship in the benchmarking datasets, the original and constraint CMIP5 ensemble for (a) Central Europe, and (b) Central North America (1989–2005, area-average). (c, e) Biases in warm season (c) TXx mean, and (e) 90th percentile of TXx in the original CMIP5 ensemble, and (d, f) reduction in biases in (d) TXx mean, and (f) 90th percentile TXx through the application of the land-coupling constraint. Regions with a significant reduction in (d) TXx mean, and (f) the cross-model average in the 90th percentile of TXx according to a permutation significance test are stippled.



**Figure 5.** Application of land coupling constraint to CMIP5 ensemble. (a, b) Ensemble prediction of original and constrained multi-model ensemble for (a) future absolute TXx and (b) range of TXx anomalies relative to global mean temperature anomalies in each model, following Seneviratne et al (2016). Envelopes indicate 5th to 95th percentile. (c, d) Global maps of present-day and future-projected changes in the simulated (c) mean TXx, and (d) 90th percentile of TXx in the VACc-constrained CMIP5 ensemble.



**Figure 6.** (a, b) Projected warming in warm season (a) mean temperature, and (b) TXx across the CMIP5 ensemble (RCP8.5 scenario, 2071-2100 relative to 1981-2010). (c, d) Correlation between VACc in the warm season and the projected warming in (c) mean temperature, and (d) TXx. Stippling indicates significant correlations. (e, f) Relative change in (e) mean warming and (f) TXx warming due to the application of the land-atmosphere coupling constraint, warming defined as 2071-2100 relative to 1981-2100.

**Table 1.** Datasets used for model evaluation

Name of dataset	Variable	Type / Group	Provider & Reference
LandFlux-EVAL <sup>a</sup>	ET	Ensemble Median	Mueller et al. (2013)
LandFlux-EVAL <sup>a</sup>	ET	Median of Reanalyses	Mueller et al. (2013)
LandFlux-EVAL <sup>a</sup>	ET	Median of LSMs	Mueller et al. (2013)
LandFlux-EVAL <sup>a</sup>	ET	Median of Diagnostic datasets	Mueller et al. (2013)
PRUNI <sup>a,b</sup>	ET	Diagnostic	<del>Sheffield et al. (2006)</del> <a href="#">Sheffield et al. (2006, 2011)</a>
MPIBGC <sup>a,b</sup>	ET	Diagnostic	Jung et al. (2011)
CSIRO <sup>a,b</sup>	ET	Diagnostic	Zhang et al. (2010)
GLEAM <sup>a,b</sup> , V. 1A	ET	Diagnostic	Miralles et al. (2011a, b)
AWB <sup>a,b</sup>	ET	Diagnostic	Mueller et al. (2011a)
EI-ORCHIDEE <sup>a,b</sup>	ET	LSM	Krinner et al. (2005)
CRU-ORCHIDEE <sup>a,b</sup>	ET	LSM	Krinner et al. (2005)
VIC <sup>a,b</sup>	ET	LSM	<del>Sheffield and Wood (2007)</del> <a href="#">Sheffield et al. (2006)</a> ; <a href="#">Sheffield et al. (2006)</a>
<del>NOAH-PFGL-NOAH-PF</del> <sup>a,b</sup>	ET	LSM	<del>Rodell et al. (2004)</del> ; <a href="#">Rui and Beaudoin (2011)</a>
MERRA-LAND <sup>a,b</sup>	ET	LSM	Reichle et al. (2011)
ERA-Interim <sup>a,b</sup>	ET	Reanalysis	Dee et al. (2011)
CFSR <sup>a,b</sup>	ET	Reanalysis	Saha et al. (2010)
JRA-25 <sup>a,b</sup>	ET	Reanalysis	Onogi et al. (2007)
MERRA <sup>a,b</sup>	ET	Reanalysis	Bosilovich (2008)
CRU-TS3.2 <sup>a</sup>	T	Observations	Harris et al. (2014)
ERA-Interim reanalysis <sup>a</sup>	T	Reanalysis	Dee et al. (2011)
<del>NCEP/DOE reanalysis2</del> <a href="#">CFSR reanalysis</a> <sup>a</sup>	T	Reanalysis	<del>Kalnay et al. (1996)</del> <a href="#">Saha et al. (2010)</a>

<sup>a</sup> All T-ET combinations of marked datasets have been used to derive the ET-T constraint.

<sup>b</sup> Original individual datasets that contributed to the LandFlux-EVAL synthesis project (Mueller et al., 2013).