Supplement to "Emerging climate signals in the Lena River catchment: a non-parametric statistical approach"

Eric Pohl[1], Christophe Grenier[1], Mathieu Vrac[1], Masa Kageyama[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement (LSCE/IPSL), UMR CEA-CNRS-UVSQ, Gif-sur-Yvette, 91120, France

Correspondence to: Eric Pohl (Eric.Pohl@lsce.ipsl.fr)

Supplementary information, data and figures

The used climate model simulations (CS) obtained through the 'esd' R package and their IDs are given in Table S1. A pre-analysis of meteorological station data with regards to the temporal coverage is displayed in Fig. S1. From the six stations that cover more than 80 years in the time period after 1900, only five stations cover more than 10 years in the reference time period. These are the stations whose associated pixels in the CRUNCEP dataset are finally taken to perform the comparison with the CS. The comparison between the Hellinger Distance (HD) evolution of CRUNCEP and the CMIP5 CS for temperature (*T)* and precipitation (*P)* is shown in Fig. S4 and Fig. S5.

Figure S2 is an extended sensitivity analysis of the introduced method in Section 4 of the main manuscripts. Figure S2 showcases the evolution in HD for strictly positive data with an example of a gamma distribution and different shape parameters. Because the KDE approach is fitting symmetrical kernels, the approach must introduce some uncertainty if the true distribution is strictly positive. Therefore, in Figure S2 we calculated the HDs based on the actual PDFs of gamma distributions with different shape parameters (x-axis). We then compare these HDs with the HDs based on the KDE approach (y-axis). We again showcase the effect of different bandwidths (by using according sample sizes) on the outcome. Figure S2 demonstrates that for actual HDs of lower than 30%, there is a pronounced positive bias of the KDE approach. This bias occurs not only for HDs between distributions with small shape parameters, i.e. distributions with a pronounced peak near 0, but also when distributions have their mean far away from 0. The bias is gradually reduced for a bigger sample size and bandwidth. For larger HD (true and estimated), distances using different sample sizes are in close agreement. The fact that the difference between HD of true and estimated PDFs are also in close agreement for small shape parameters supports possible application for non-normal distributed data, e.g. precipitation data at high temporal resolution.

Figure S3 is showing the lower internal data variability within the CRUNCEP dataset in the northwest compared to the central region. The lower variability is observed where long-term operating meteorological stations are missing (cf. Fig. S1). Figure S4 and S5 show the results from the comparison and ranking process (cf. Fig S6 and Fig. S7) for all 65 CS and the subset of 5 pixels with validated long-term records, respectively.

Figure S8 and Fig. S9 extend the sensitivity analysis in the main text (Fig. 10xxx). They show the changes in absolute time of emergence and standard deviation for different emergence levels for the two variables *T* and *P* for different choices in the meta-parameters window width and split year.

Figure 10 shows the sensitivity of ToE with respect to the choices in meta-parameters for CRUNCEP (cf. Fig 10 for CMIP5 CS).

Table S1: Overview of Climate model simulations from the obtained CMIP5 collection using the 'esd' R-package

000-ACCESS1-0-r1i1p1	022-CNRM-CM5-r4i1p1	044-FIO-ESM-r1i1p1
001-ACCESS1-3-r1i1p1	023-CNRM-CM5-r6i1p1	045-FIO-ESM-r2i1p1
002-bcc-csm1-1-r1i1p1	024-CNRM-CM5-r10i1p1	046-FIO-ESM-r3i1p1
003-BNU-ESM-r1i1p1	025-CSIRO-Mk3-6-0-r1i1p1	047-GFDL-CM3-r1i1p1
004-CanESM2-r1i1p1	026-CSIRO-Mk3-6-0-r2i1p1	048-GFDL-ESM2G-r1i1p1
005-CanESM2-r2i1p1	027-CSIRO-Mk3-6-0-r3i1p1	049-GFDL-ESM2M-r1i1p1
006-CanESM2-r3i1p1	028-CSIRO-Mk3-6-0-r4i1p1	050-GISS-E2-H-r1i1p1
007-CanESM2-r4i1p1	029-CSIRO-Mk3-6-0-r5i1p1	051-GISS-E2-H-r1i1p2
008-CanESM2-r5i1p1	030-CSIRO-Mk3-6-0-r6i1p1	052-GISS-E2-H-r1i1p3
009-CCSM4-r1i1p1	031-CSIRO-Mk3-6-0-r7i1p1	053-GISS-E2-R-r1i1p1
010-CCSM4-r2i1p1	032-CSIRO-Mk3-6-0-r8i1p1	054-GISS-E2-R-r1i1p2
011-CCSM4-r3i1p1	033-CSIRO-Mk3-6-0-r9i1p1	055-GISS-E2-R-r1i1p3
012-CCSM4-r4i1p1	034-CSIRO-Mk3-6-0-r10i1p1	056-HadGEM2-AO-r1i1p1
013-CCSM4-r5i1p1	035-EC-EARTH-r1i1p1	057-HadGEM2-CC-r1i1p1
014-CCSM4-r6i1p1	036-EC-EARTH-r2i1p1	058-HadGEM2-ES-r1i1p1
015-CESM1-BGC-r1i1p1	037-EC-EARTH-rcp85	059-HadGEM2-ES-r2i1p1
016-CESM1-CAM5-r1i1p1	038-EC-EARTH-r8i1p1	060-HadGEM2-ES-r3i1p1
017-CESM1-CAM5-r2i1p1	039-EC-EARTH-r9i1p1	061-HadGEM2-ES-r4i1p1
018-CMCC-CM-r1i1p1	040-EC-EARTH-r11i1p1	062-inmcm4-r1i1p1
019-CMCC-CMS-r1i1p1	041-EC-EARTH-r12i1p1	063-IPSL-CM5A-LR-r1i1p1
020-CNRM-CM5-r1i1p1	042-EC-EARTH-r13i1p1	064-IPSL-CM5A-LR-r2i1p1
021-CNRM-CM5-r2i1p1	043-FGOALS-g2-r1i1p1	

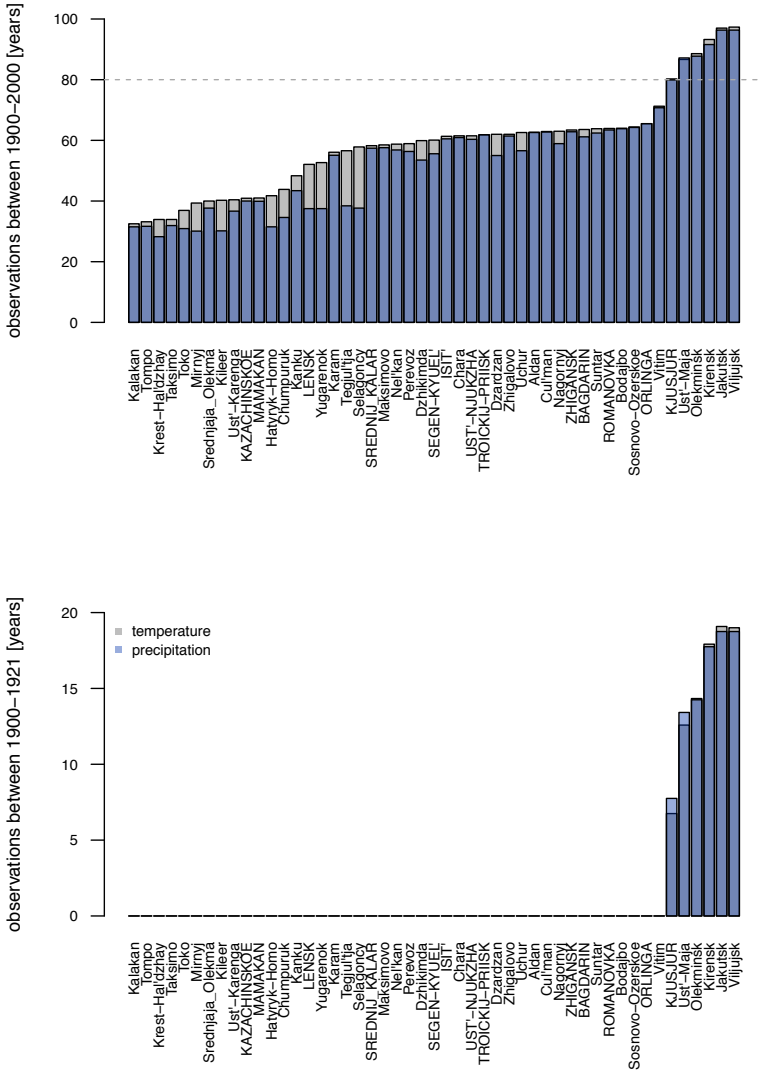

Figure S1: Number of observations for temperature and precipitation for each of the 49 meteorological stations in the Lena River catchment from the RIHMI-WDC dataset for the period 1900-2000 (top) and within the reference period (bottom).

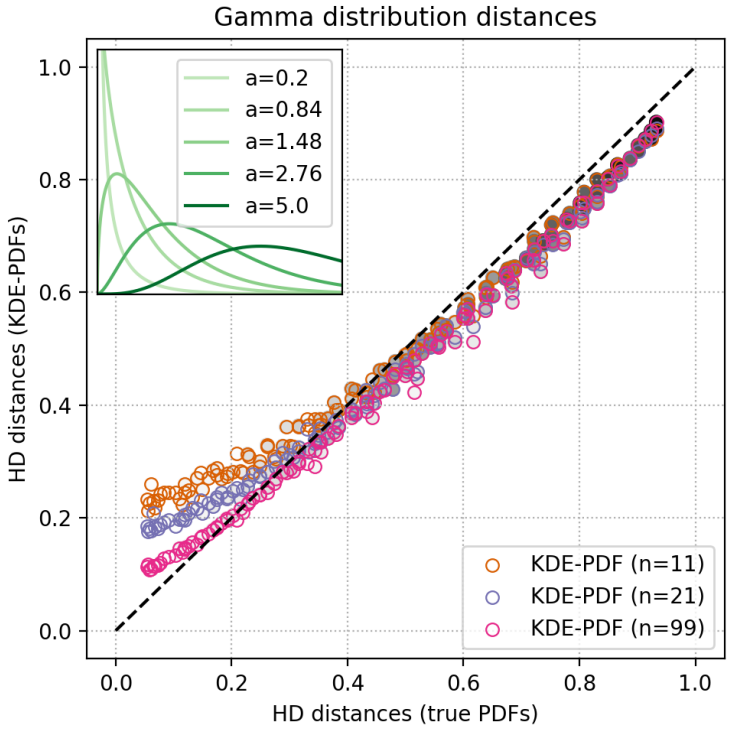

Figure S2: Comparison of Hellinger distances for strictly positive data (by means of a gamma distribution) calculated from the actual PDF vs. the KDE-PDF approach as used in the manuscript. The different sample sizes for the KDE-PDFs correspond roughly to the different bandwidths tested in Figure 2. Inset shows five of the 16 gamma distributions with different shape parameters "a" between 0.2 and 5.0 used for this test. The HD between each possible gamma distribution with a different shape parameter was calculated. This was done for both, the "true" (the actual PDFs) and the estimated (KDE) PDFs. This was repeated 100 times for each distance. The averages of these distances are plotted. The grey-coding of points (face color) represents the difference in shape parameter a, with light colors for small and dark colors for large differences.

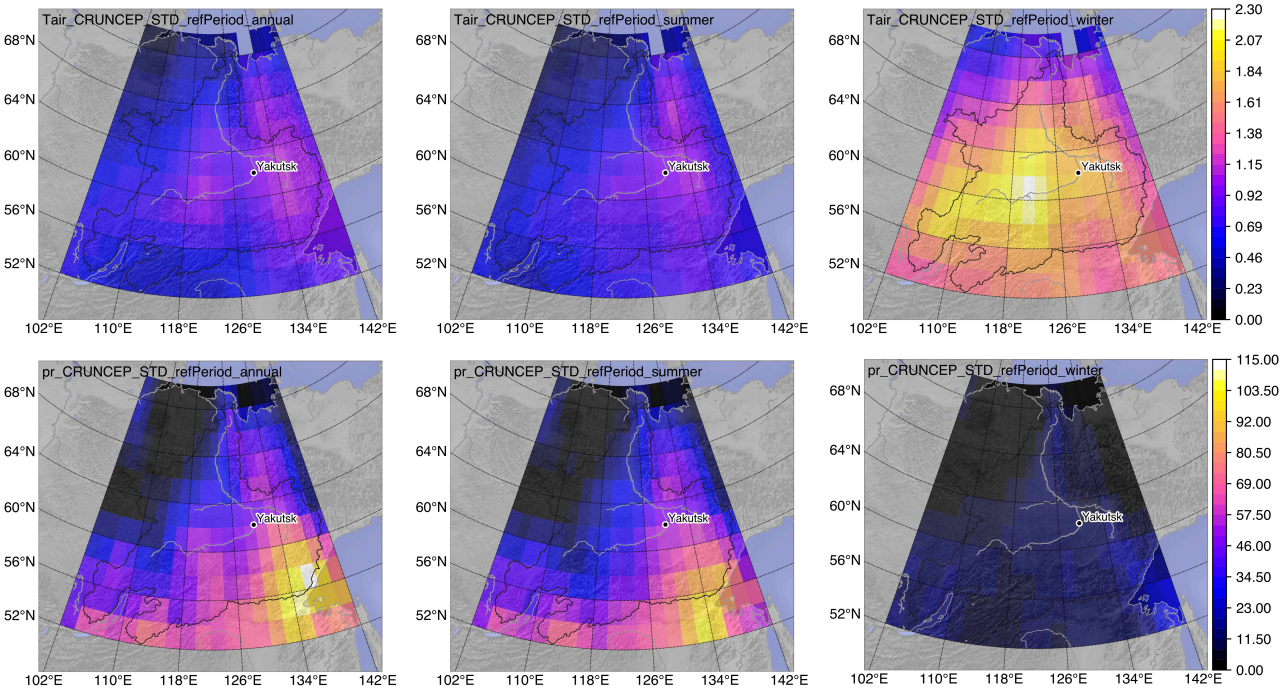

Figure S3: Variability in the CRUNCEP data within the reference period (1901-1921) by means of the standard deviation. Standard deviation for temperature (top) in ºC and for precipitation (bottom) in mm for annual (left), summer (middle), and winter (right) values. Northeastern region where no long-term meteorological stations exist show artificial, repetitive data in the CRUNCEP and associated very low standard deviations.

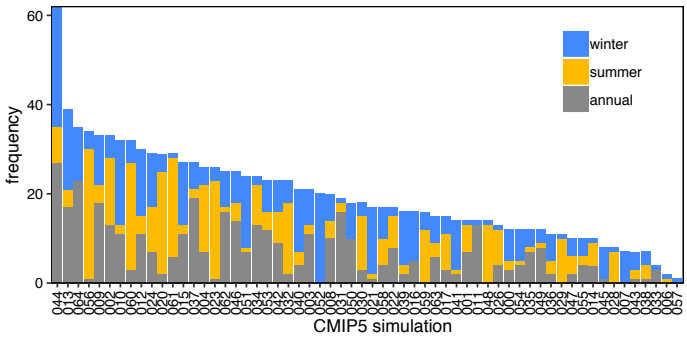

Figure S4: Ranked best model simulations based on Nash-Sutcliffe efficiency between HD of CRUNCEP and individual CMIP5 model simulations for all pixels encompassing meteorological stations (Fig.1).

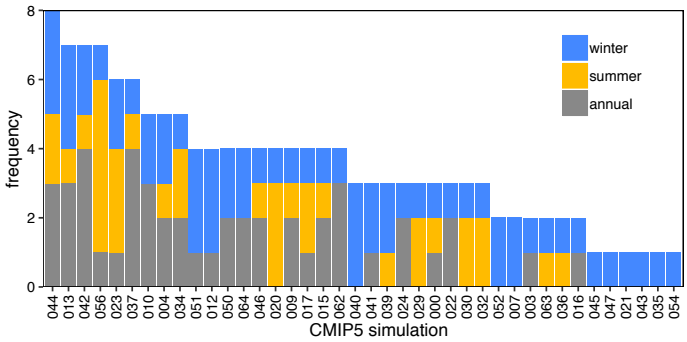

Figure S5: The same as Fig.S3 but for the 5 pixels encompassing the long-term meteorological stations (Fig. 1), excluding Kjusjur. For Kjusjur less than 10 years of data are available in the reference period (Fig. S1) and artificial data recycling was apparent in the CRUNCEP dataset, which causes a sooner and artificially caused emergence.

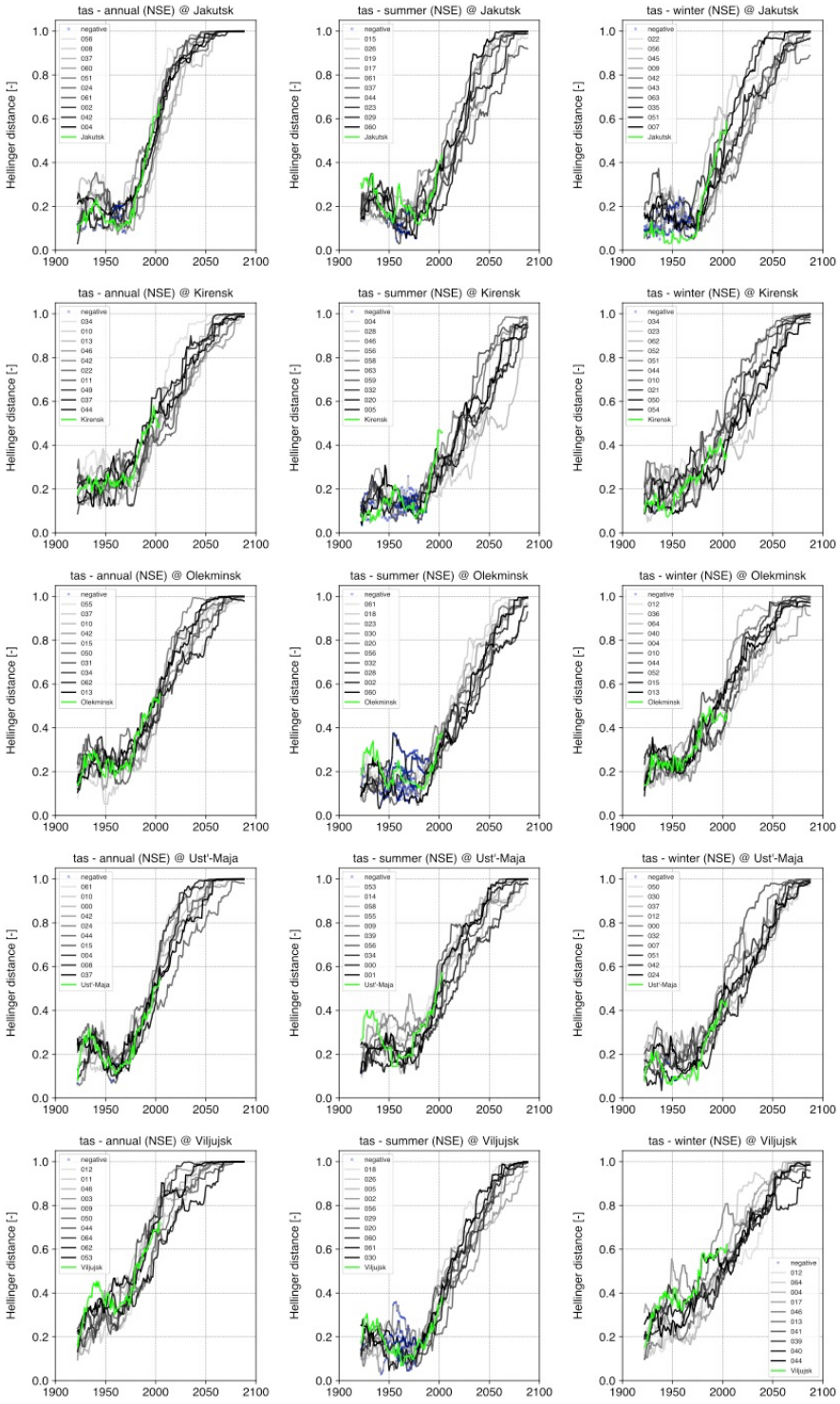

Figure S6: Comparison of Hellinger Distances between CRUNCEP (green) and the 10 best performing climate simulations (in terms of NSE) for temperature at locations where long-term in situ data is available.

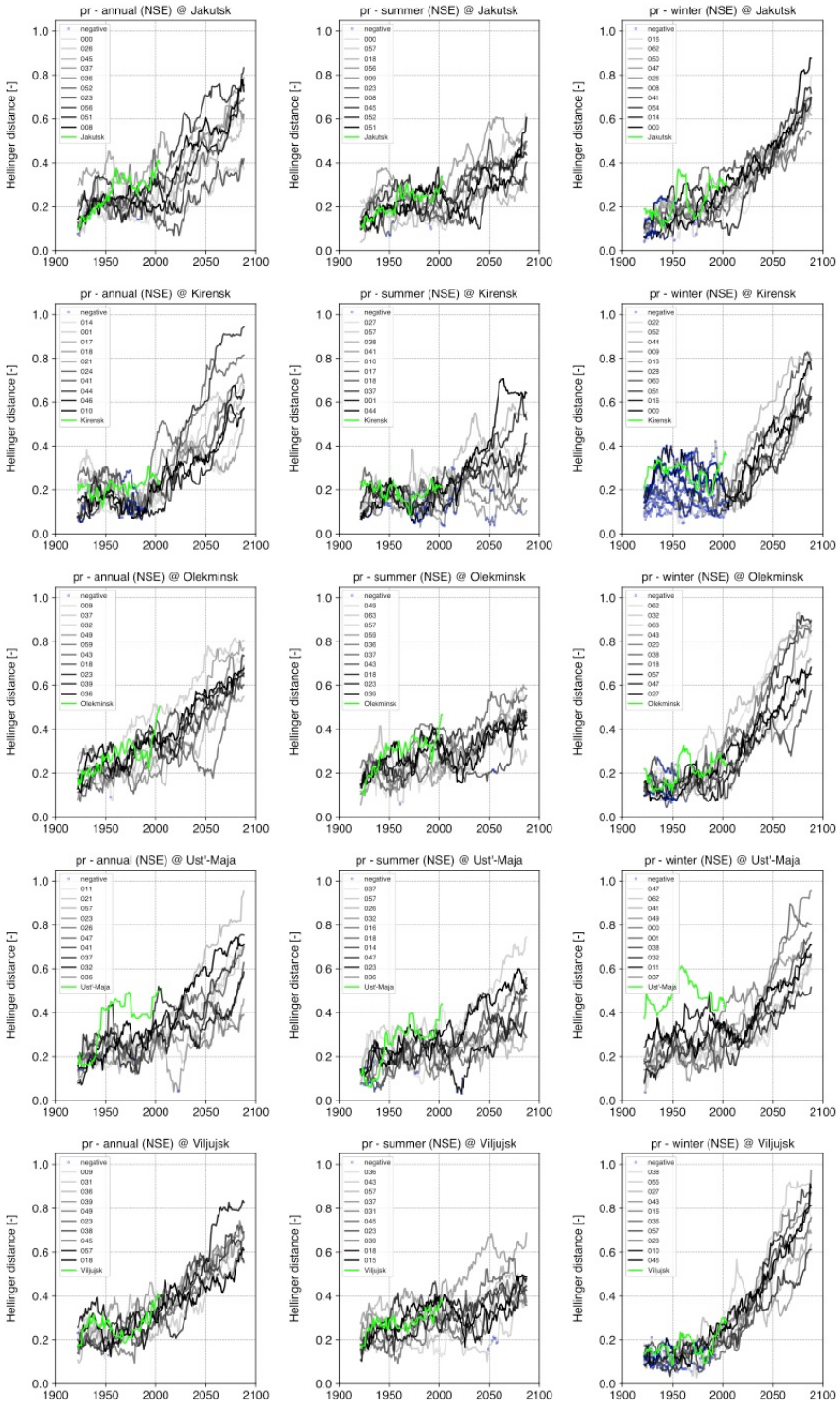

Figure S7: Same as Fig. S3 but for precipitation.

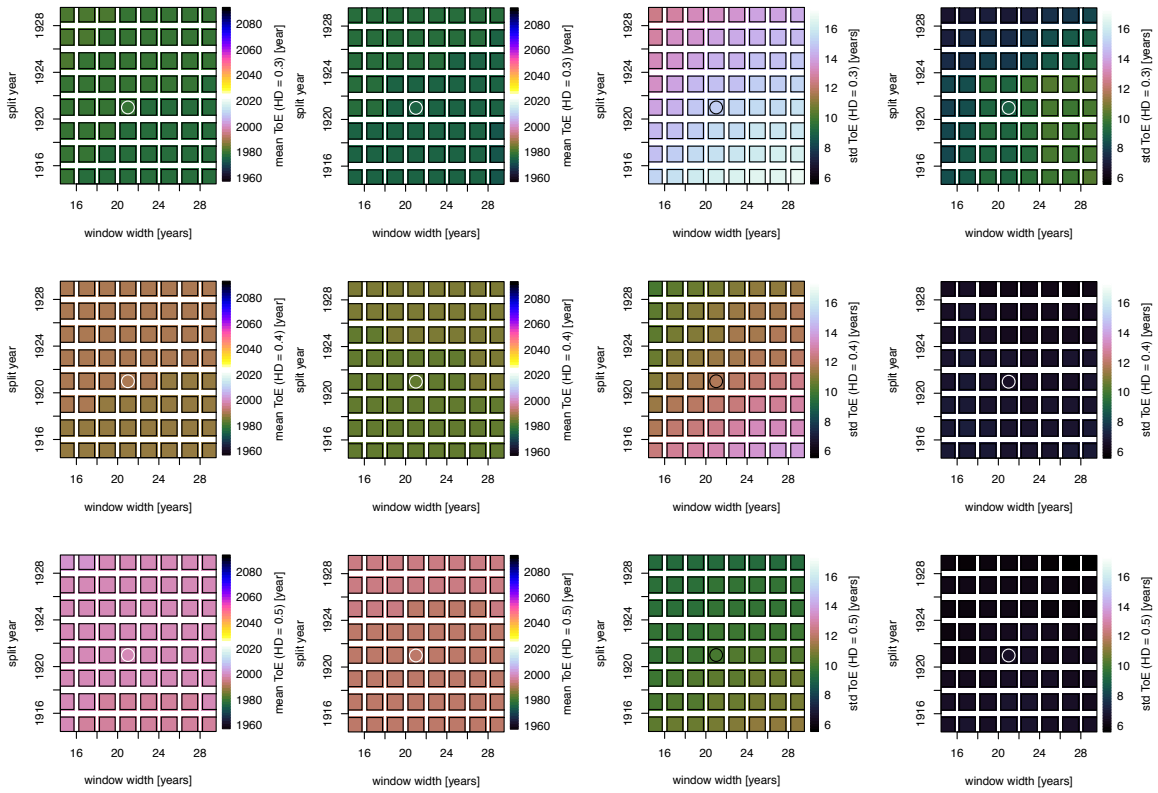

Figure S8: Sensitivity for derived ToE of temperature for different combinations of the meta-parameters window width and split year (i.e. the end of the reference period). Left two columns – derived mean ToE over all 65 climate simulations (first column), and the 10 best simulations (second column). Right two columns – standard deviation of ToE for all 65 climate simulations (first column), and the 10 best simulations (second column). Rows – different emergence levels (30%, 40%, and 50% from top to bottom). Circle marks the used combination for the presented results in the main manuscript.

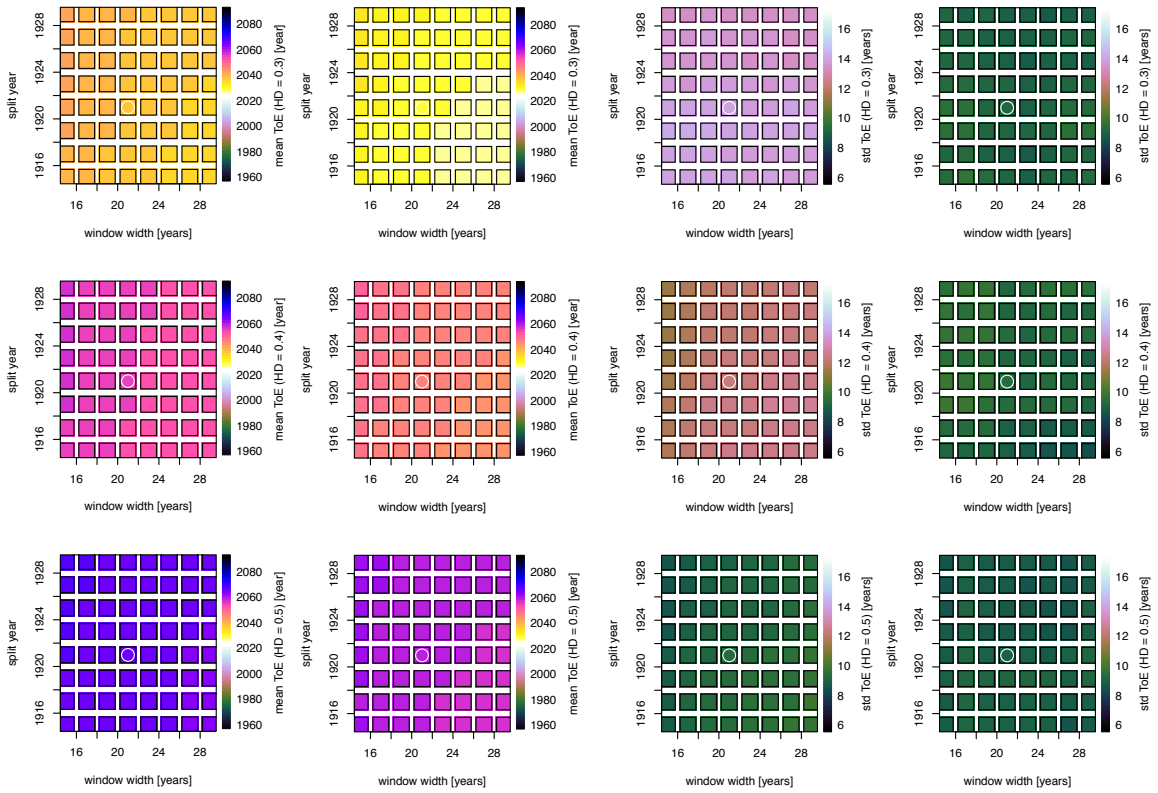

Figure S9: Sensitivity of derived ToE of precipitation for different combinations of the meta-parameters window width and split year (i.e. the end of the reference period). Left two columns – derived mean ToE over all 65 climate simulations (first column), and the 10 best simulations (second column). Right two columns – standard deviation of ToE for all 65 climate simulations (first column), and the 10 best simulations (second column). Rows – different emergence levels (30%, 40%, and 50% from top to bottom). Circle marks the used combination for the presented results in the main manuscript.

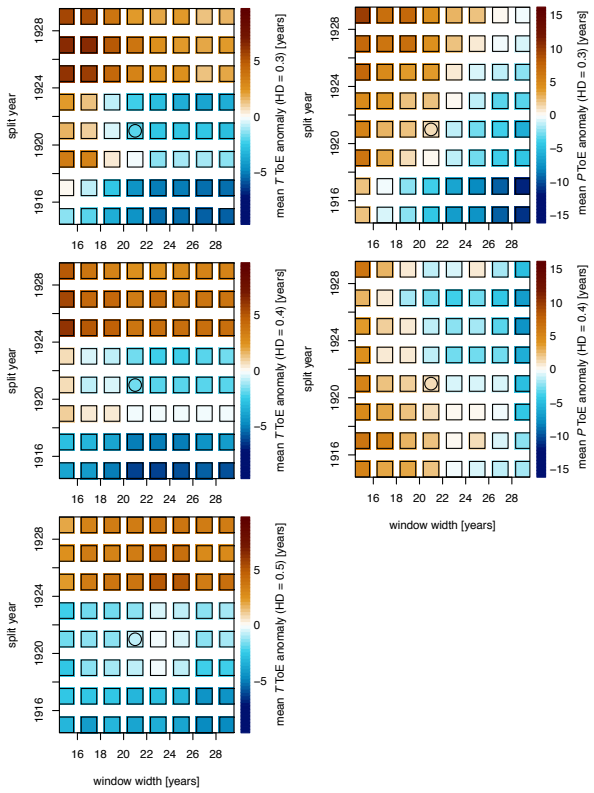

Figure S10: Sensitivity of derived ToE of temperature (left) and precipitation (right) for different combinations of the meta-parameters window width and split year (i.e. the end of the reference period) for the CRUNCEP dataset. Results for precipitation represent the entire study area, including the large fraction in the northwestern part with issues in the early data records (see main text and Fig. S2). Rows – different emergence levels (30%, 40%, and 50% from top to bottom). Circle marks the used combination for the presented results in the main manuscript. Note different scaling for *T* and *P*, and that not all pixels show emergence at 40% and 50% (especially for precipitation – cf. Fig. 4). The resulting values represent thus only a certain fraction of the entire study area.