Supplement of Earth Syst. Dynam., 11, 1033–1049, 2020
https://doi.org/10.5194/esd-11-1033-2020-supplement
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Earth System Dynamics
Open Access
EGU

*Supplement of*

# Calibrating large-ensemble European climate projections using observational data

**Christopher H. O'Reilly et al.**

*Correspondence to:* Christopher H. O'Reilly (christopher.oreilly@physics.ox.ac.uk)

The copyright of individual parts of the supplement might differ from the CC BY 4.0 License.

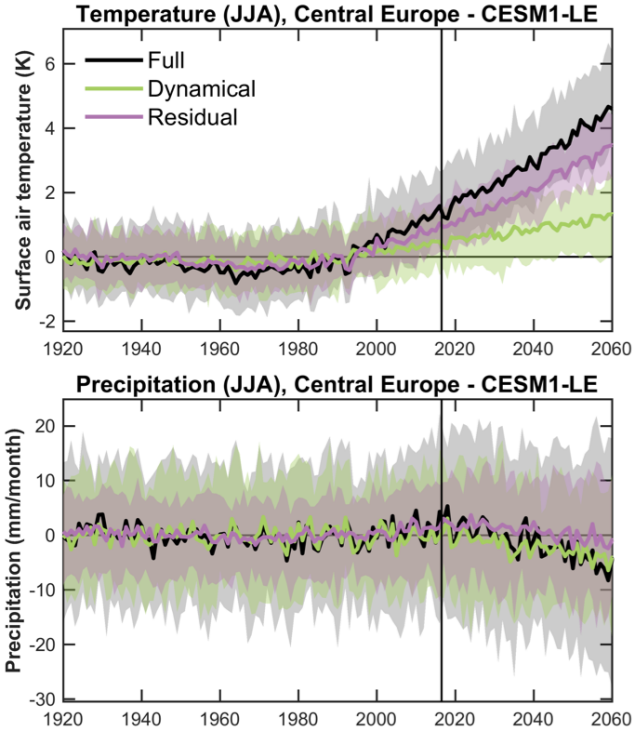



**Figure S1:** Example of the CESM1-LE projection of summertime Central European temperature (top) and precipitation (bottom) decomposed from the full anomalies into dynamical and residual components. The lines show the ensemble medians and the shading shows the 90% range of the ensemble.

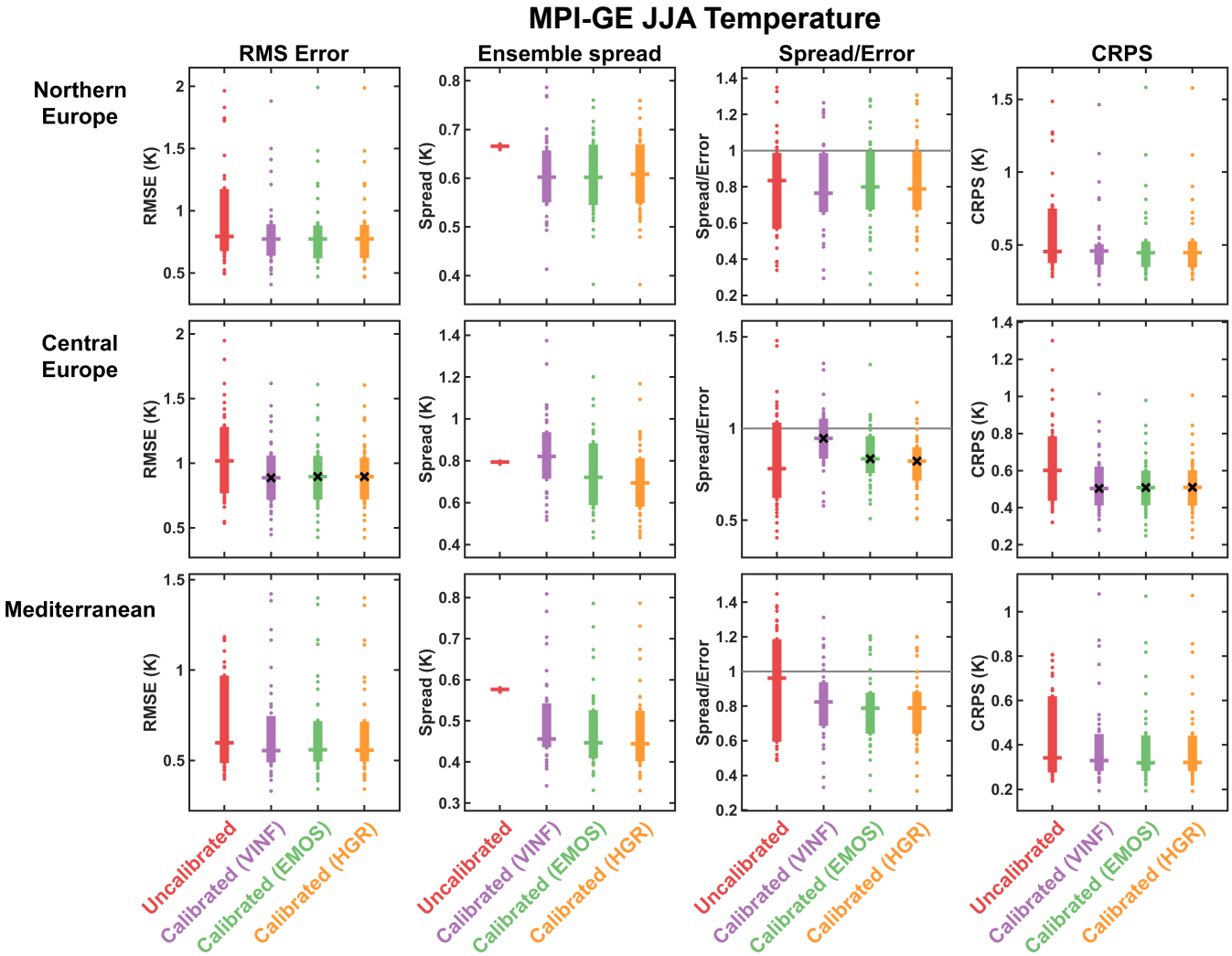



**Figure S2:** Comparison of calibration methods applied to the MPI-GE summer temperature projections calibrated to the CMIP5 models over the observational period (1920-2016) and verified using the 44 years in the out-of-sample period (1917-2060). The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines. For the calibrated RMS Error, spread/error and CRPS values, the black crosses indicate where the calibration represents a significant improvement over the uncalibrated (but bias-corrected) ensemble at the 90% significance level. The significance levels were calculated using the non-parametric Mann-Whitney U-test.

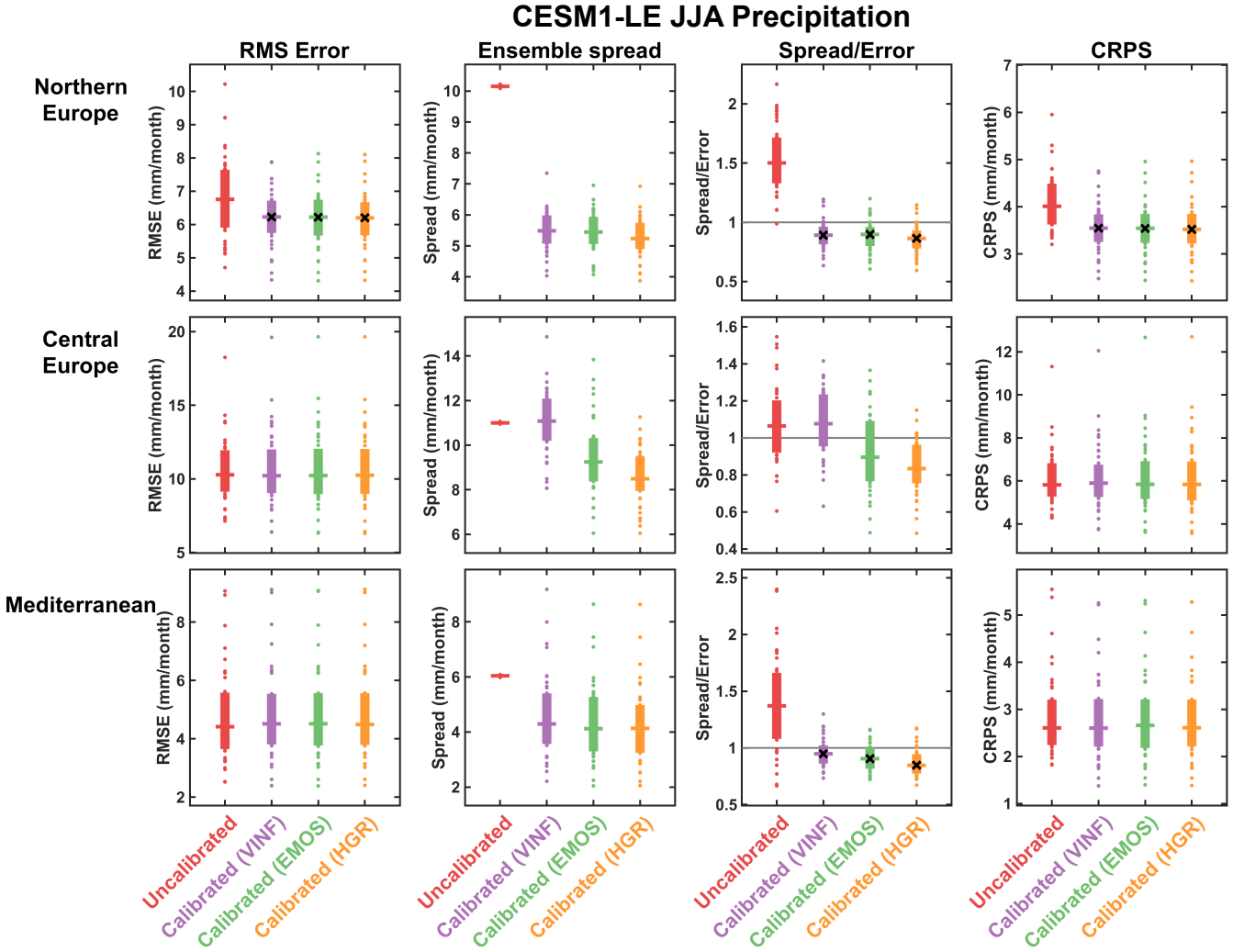



**Figure S3:** Comparison of calibration methods applied to the CESM1-LE summer precipitation projections calibrated to the CMIP5 models over the observational period (1920-2016) and verified using the 44 years in the out-of-sample period (1917-2060). The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines. For the calibrated RMS Error, spread/error and CRPS values, the black crosses indicate where the calibration represents a significant improvement over the uncalibrated (but bias-corrected) ensemble at the 90% significance level. The significance levels were calculated using the non-parametric Mann-Whitney U-test.

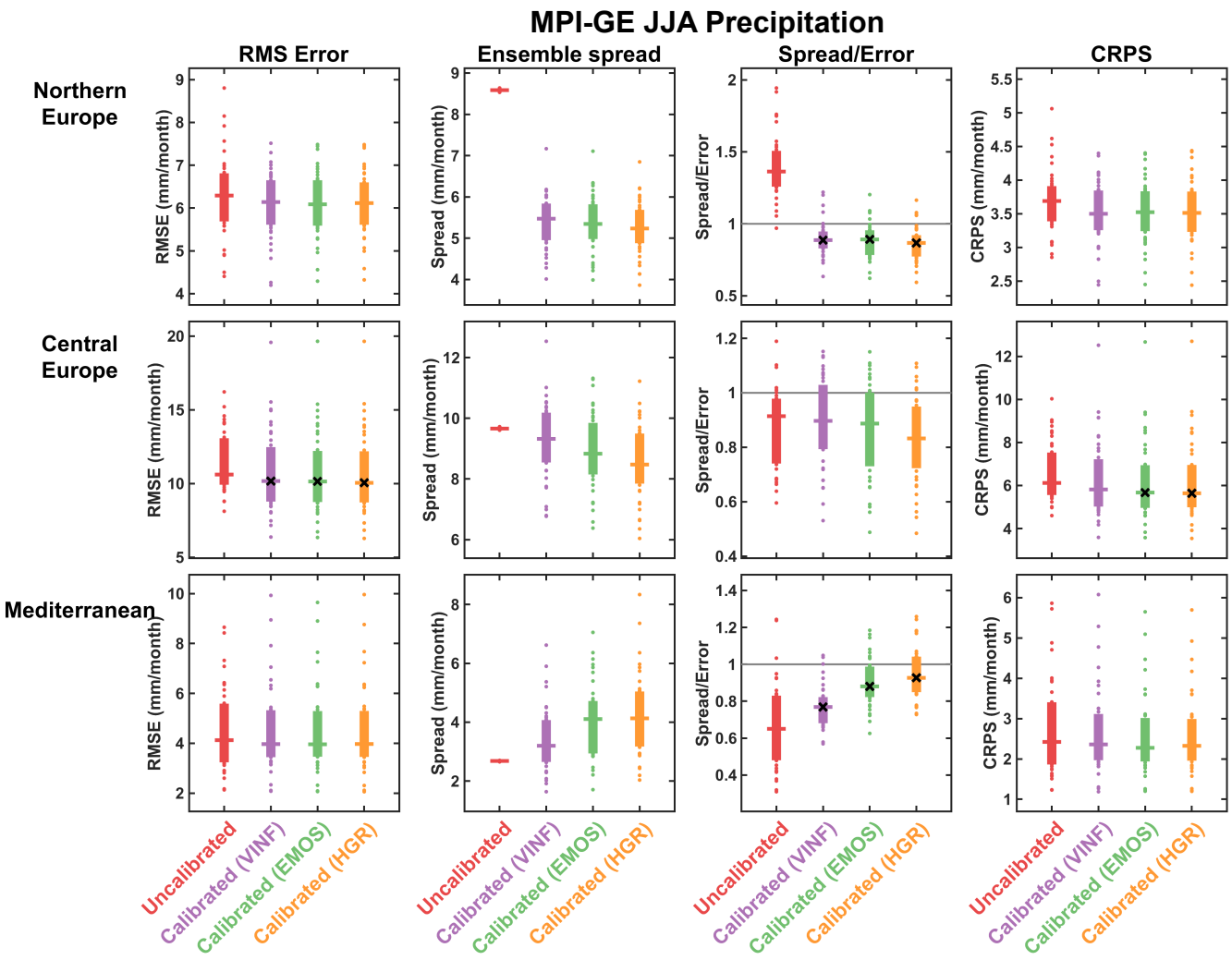



**Figure S4:** Comparison of calibration methods applied to the MPI-GE summer precipitation projections calibrated to the CMIP5 models over the observational period (1920-2016) and verified using the 44 years in the out-of-sample period (1917-2060). The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines. For the calibrated RMS Error, spread/error and CRPS values, the black crosses indicate where the calibration represents a significant improvement over the uncalibrated (but bias-corrected) ensemble at the 90% significance level. The significance levels were calculated using the non-parametric Mann-Whitney U-test.

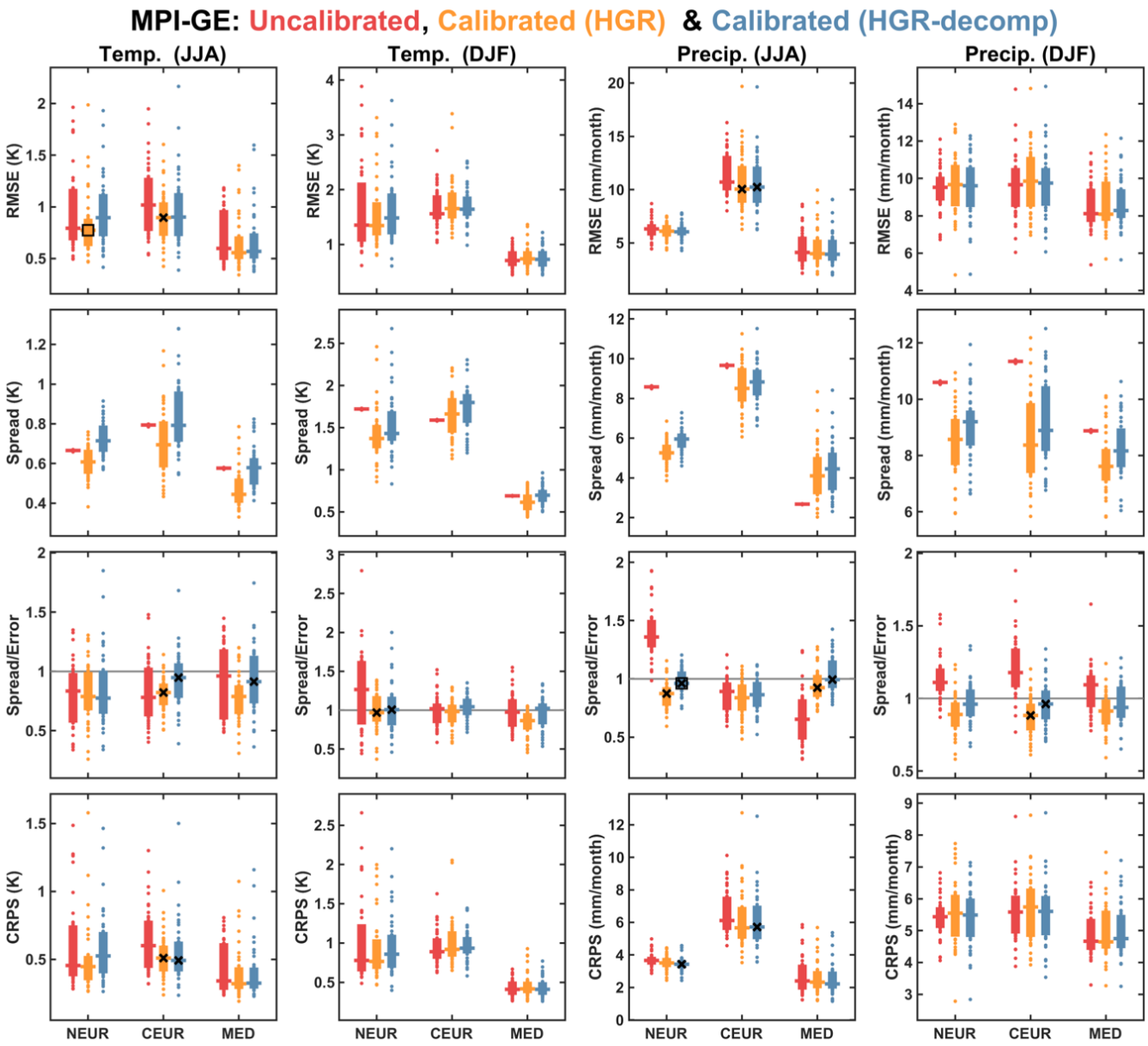



**Figure S5:** Overview of verification of the HGR and HGR-decomp calibration methods compared with the uncalibrated MPI-GE data in the European regions. Results are shown for all of the verification measures, for both summer and winter seasons and for temperature and precipitation. The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines. For the calibrated RMS Error, spread/error and CRPS values, the black crosses indicate where the calibration represents a significant improvement over the uncalibrated (but bias-corrected) ensemble at the 90% significance level. Black circles indicate where the calibration is significantly worse than the uncalibrated ensemble (at the 90% level). The black boxes show where one calibration method is found to be significantly better than the other calibration method for the same variable, season and region at the 90% significance level. The significance levels were calculated using the non-parametric Mann-Whitney U-test.

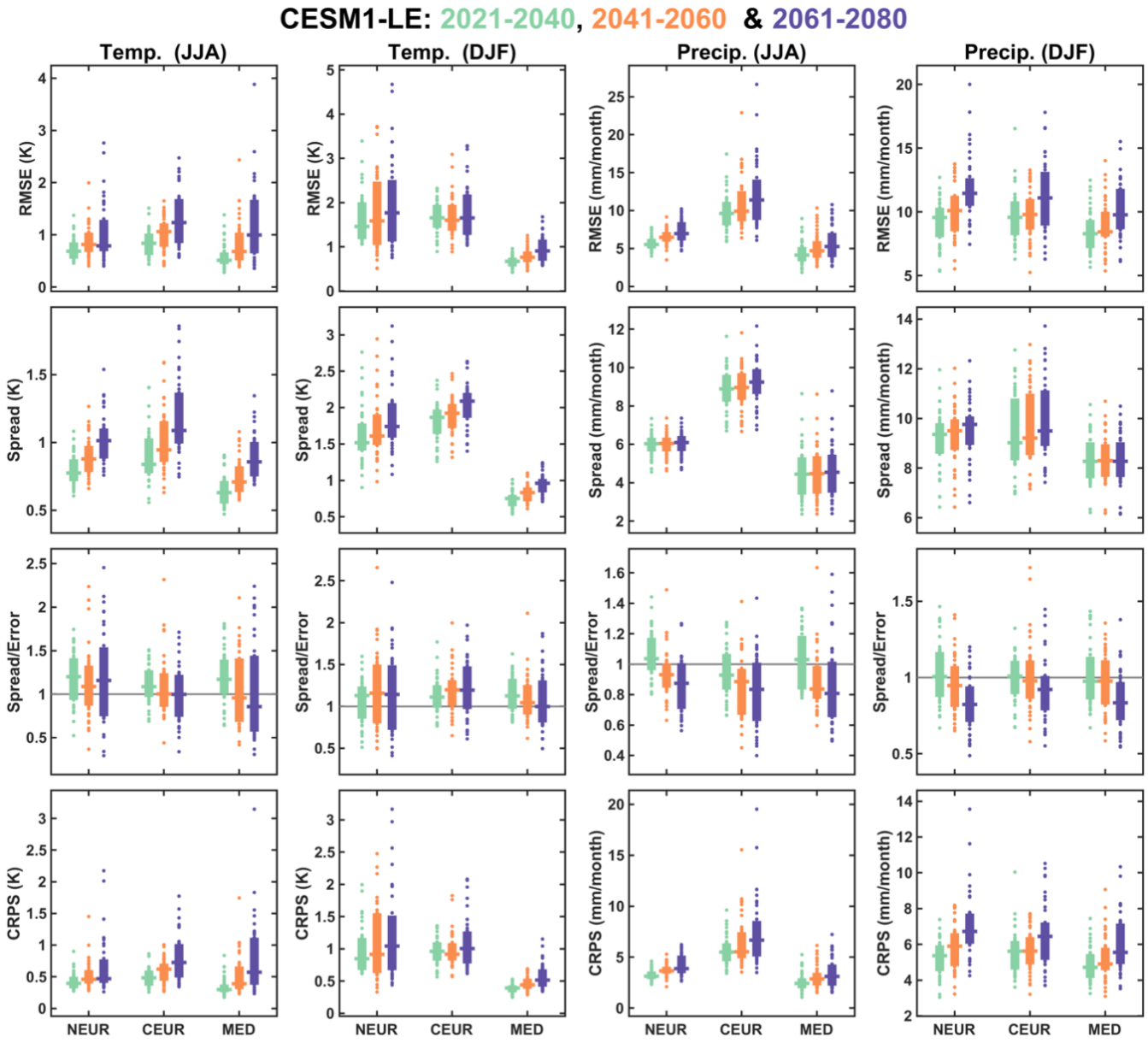



**Figure S6:** Verification of HGR-decomp calibration method applied to CESM1-LE data in the European regions for different verification periods: 2021-2040, 2041-2060 & 2061-2080. Shown for all of the verification measures, for both summer and winter seasons and for temperature and precipitation. The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines.

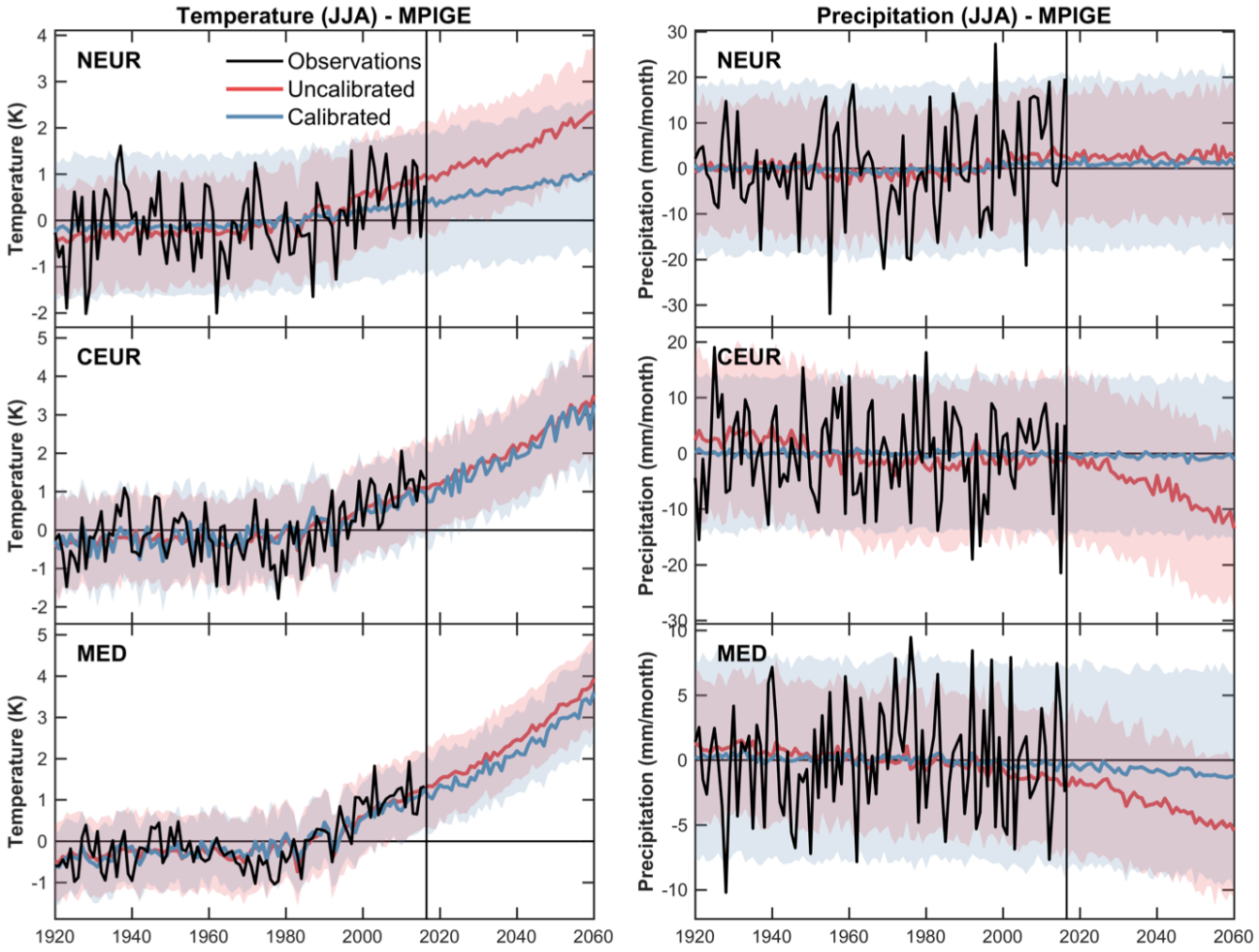



**Figure S7:** Uncalibrated and calibrated (HGR-decomp) MPI-GE projections, where here the calibrated projections have been calibrated against the observations over the period 1920-2016. The lines show the ensemble medians for the uncalibrated and calibrated ensembles. The shading shows the 90% range of the LENS ensemble. Based on the verification out-of-sample tests using the CMIP5 models the calibrated ensemble is expected to be more reliable than the uncalibrated ensemble, particularly for temperatures.

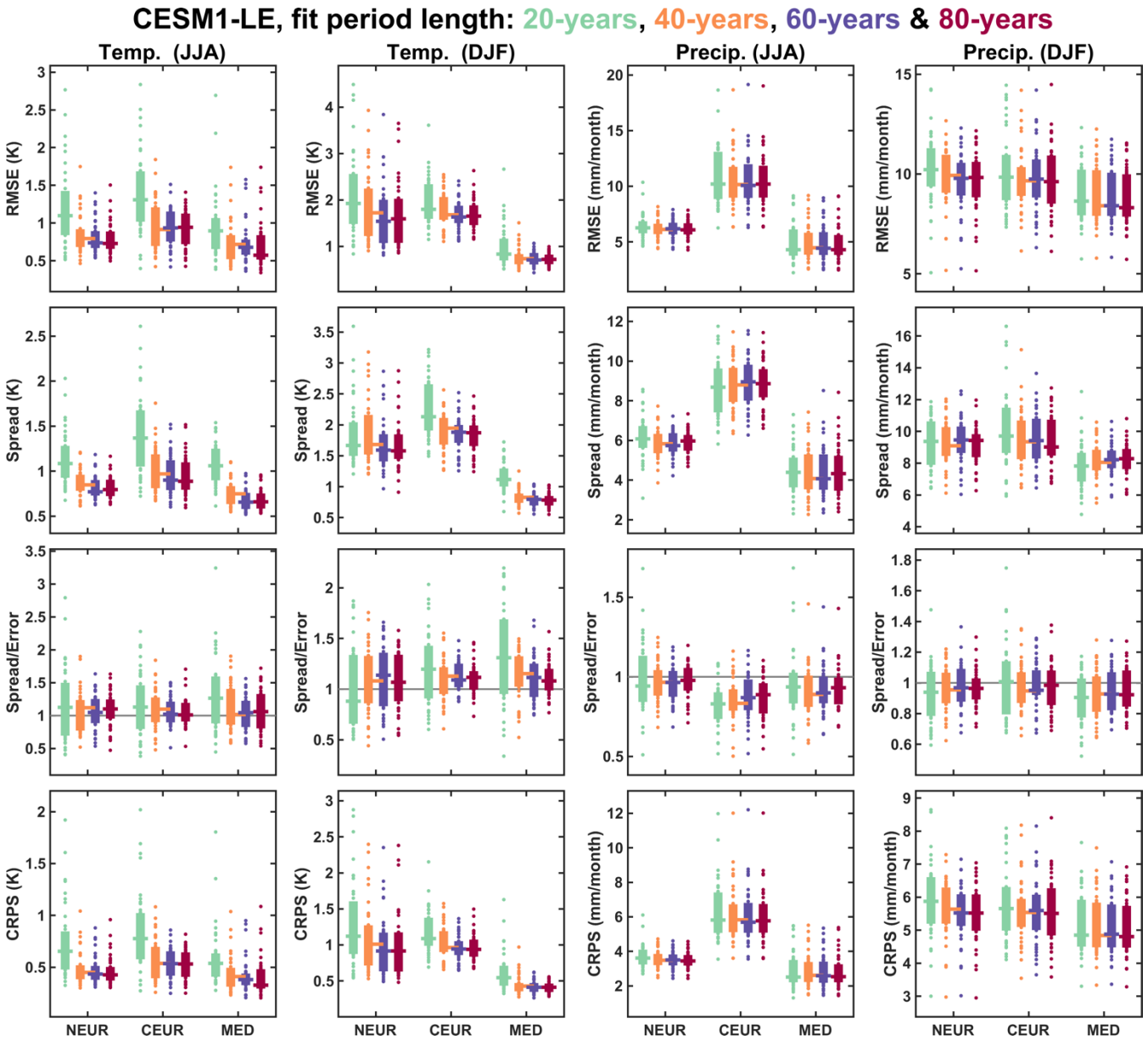



**Figure S8:** Verification of HGR-decomp calibration method applied to CESM1-LE data in the European regions for fit periods of 20, 40, 60 and 80 years. Results are shown for all of the verification measures, for both summer and winter seasons and for temperature and precipitation. The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines.

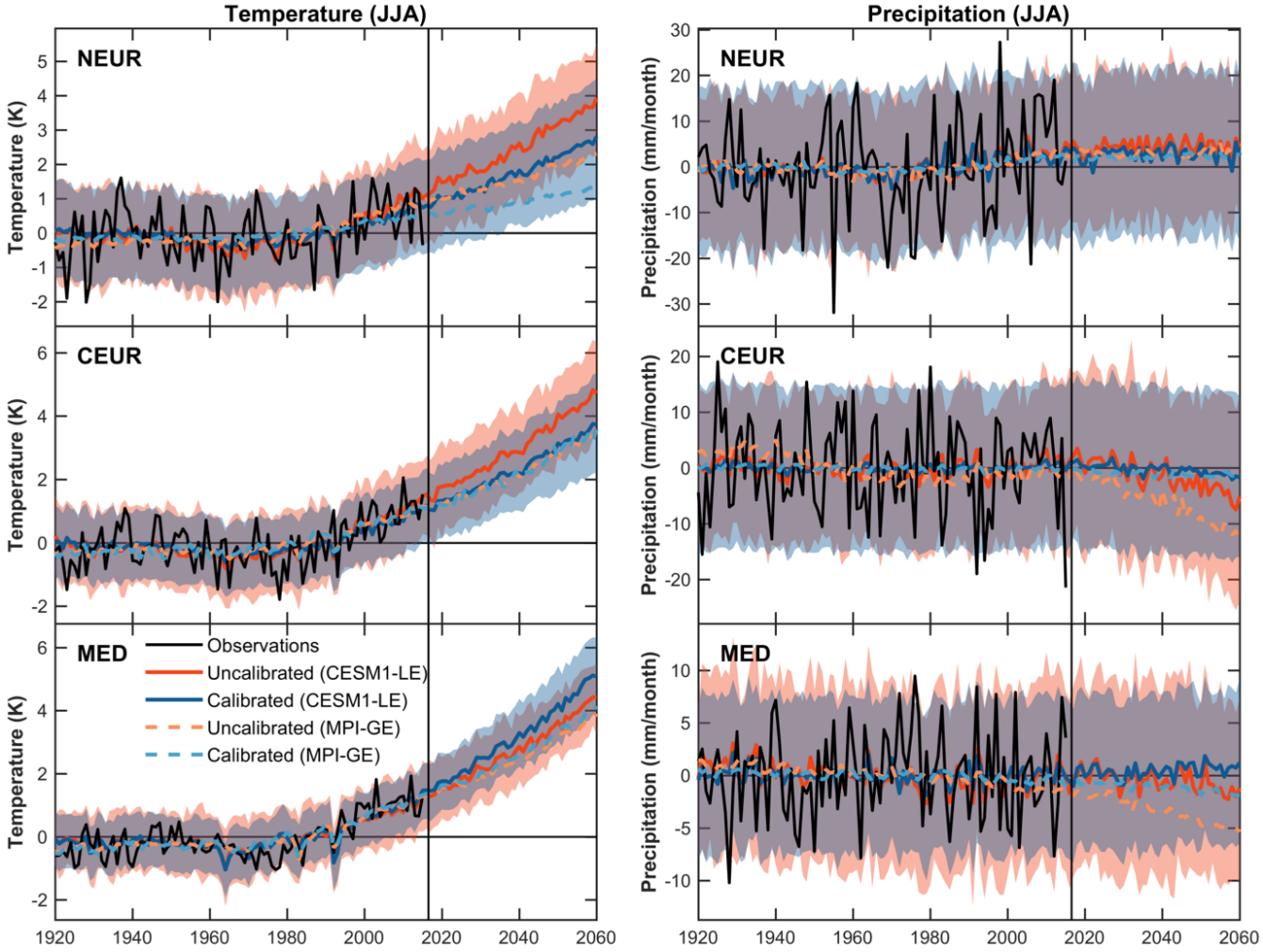
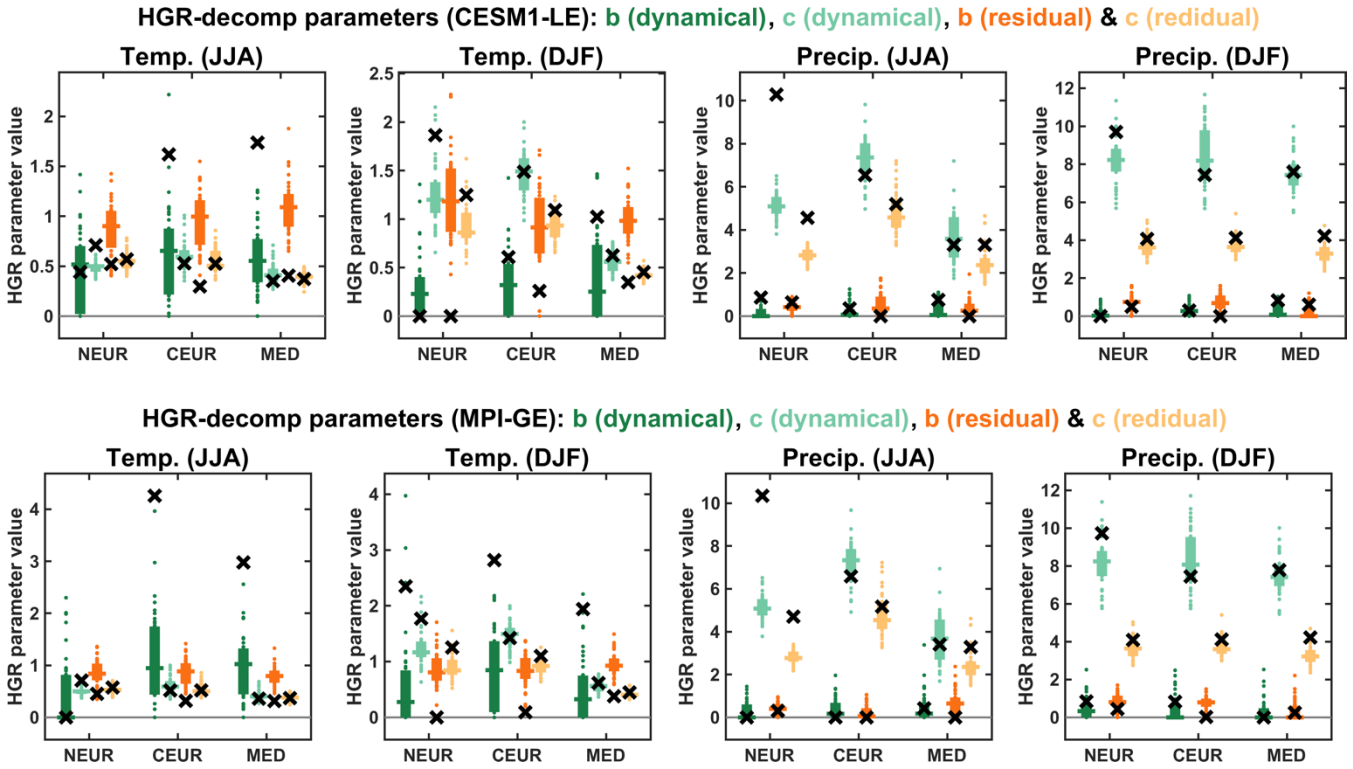



**Figure S9:** As in Figure 7 of the main paper but using observations decomposed using the 20CR dataset rather than HadSLP2. Uncalibrated and calibrated (HGR-decomp) CESM1-LE projections, where here the calibrated projections have been calibrated against the observations over the period 1920-2016. The lines show the ensemble medians for the uncalibrated and calibrated ensembles for both the CESM1-LE (solid) and MPI-GE (dashed) datasets. The shading shows the 5-95% range of the CESM1-LE ensemble. Based on the verification out-of-sample tests using the CMIP5 models the calibrated ensemble is expected to be more reliable than the uncalibrated ensemble, particularly for temperatures.




95 **Figure S10:** Distribution of the median *b* & *c* parameters (i.e. equation 7 of the main text) from the HGR-decomp methodology applied to the 39 CMIP5 models. Each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median of the CMIP5 ensemble distribution is indicated by the horizontal lines. The black crosses show the equivalent parameter for the HGR-decomp was applied to the observations over the same period.

100

105

110

| | *DYNAMICAL – HadSLP2* | *DYNAMICAL – 20CR* |
|---|---|---|
| $T_{JJA}$ **– NEUR** | 49.0 % | 51.9 % |
| $T_{JJA}$ **– CEUR** | 54.4 % | 51.9 % |
| $T_{JJA}$ **– MED** | 63.0 % | 48.8 % |
| $P_{JJA}$ **– NEUR** | 79.2 % | 73.7 % |
| $P_{JJA}$ **– CEUR** | 59.5 % | 58.0 % |
| $P_{JJA}$ **– MED** | 43.4 % | 54.2 % |
| $T_{DJF}$ **– NEUR** | 64.8 % | 74.0 % |
| $T_{DJF}$ **– CEUR** | 66.4 % | 70.3 % |
| $T_{DJF}$ **– MED** | 69.1 % | 60.5 % |
| $P_{DJF}$ **– NEUR** | 84.0 % | 81.8 % |
| $P_{DJF}$ **– CEUR** | 78.5 % | 77.7 % |
| $P_{DJF}$ **– MED** | 77.4 % | 76.7 % |

**Table S1:** The amount of the total variance explained (i.e. $r^2$, expressed as a percentage) by the DYNAMICAL component of the decomposition, using two different observational SLP datasets, HadSLP2 and 20CR, calculated over the overlapping period (1920-2015).

| | $T_{JJA}$ *(DYN)* | $T_{JJA}$ *(RES)* | $P_{JJA}$ *(DYN)* | $P_{JJA}$ *(RES)* | $T_{DJF}$ *(DYN)* | $T_{DJF}$ *(RES)* | $P_{DJF}$ *(DYN)* | $P_{DJF}$ *(RES)* |
|---|---|---|---|---|---|---|---|---|
| **NEUR** | 0.80 | 0.71 | 0.87 | 0.51 | 0.89 | 0.75 | 0.93 | 0.66 |
| **CEUR** | 0.73 | 0.67 | 0.75 | 0.59 | 0.88 | 0.74 | 0.91 | 0.69 |
| **MED** | 0.63 | 0.56 | 0.74 | 0.67 | 0.83 | 0.70 | 0.88 | 0.58 |

**Table S2:** Correlation between the HadSLP2 and 20CR decomposed observational indices, using data over the overlapping period (1920-2015).