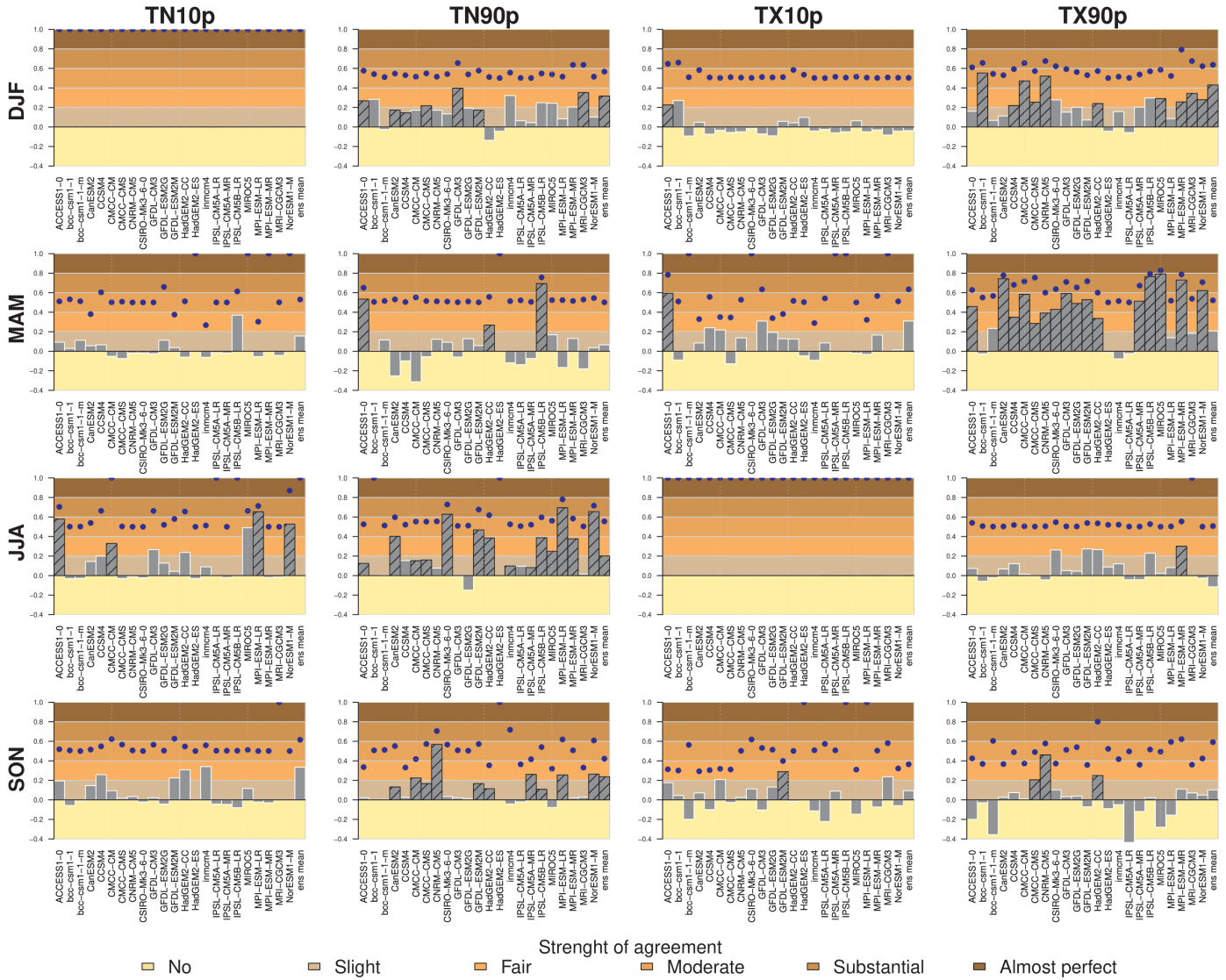

Figure S1: Cohen's Kappa (bars) and Mapcurves (points) indices comparing correlations (HadEX3 data set against GCMs). The strength of agreement of Cohen's Kappa coefficient is shown in colors and dashed indicate significant kappa at 5%.

Note that the kappa coefficient is null for all models for TN10p in DJF and TX10p in JJA because only the non-significant category has been observed, so the accuracy probability is equal to the probability of agreement by chance.

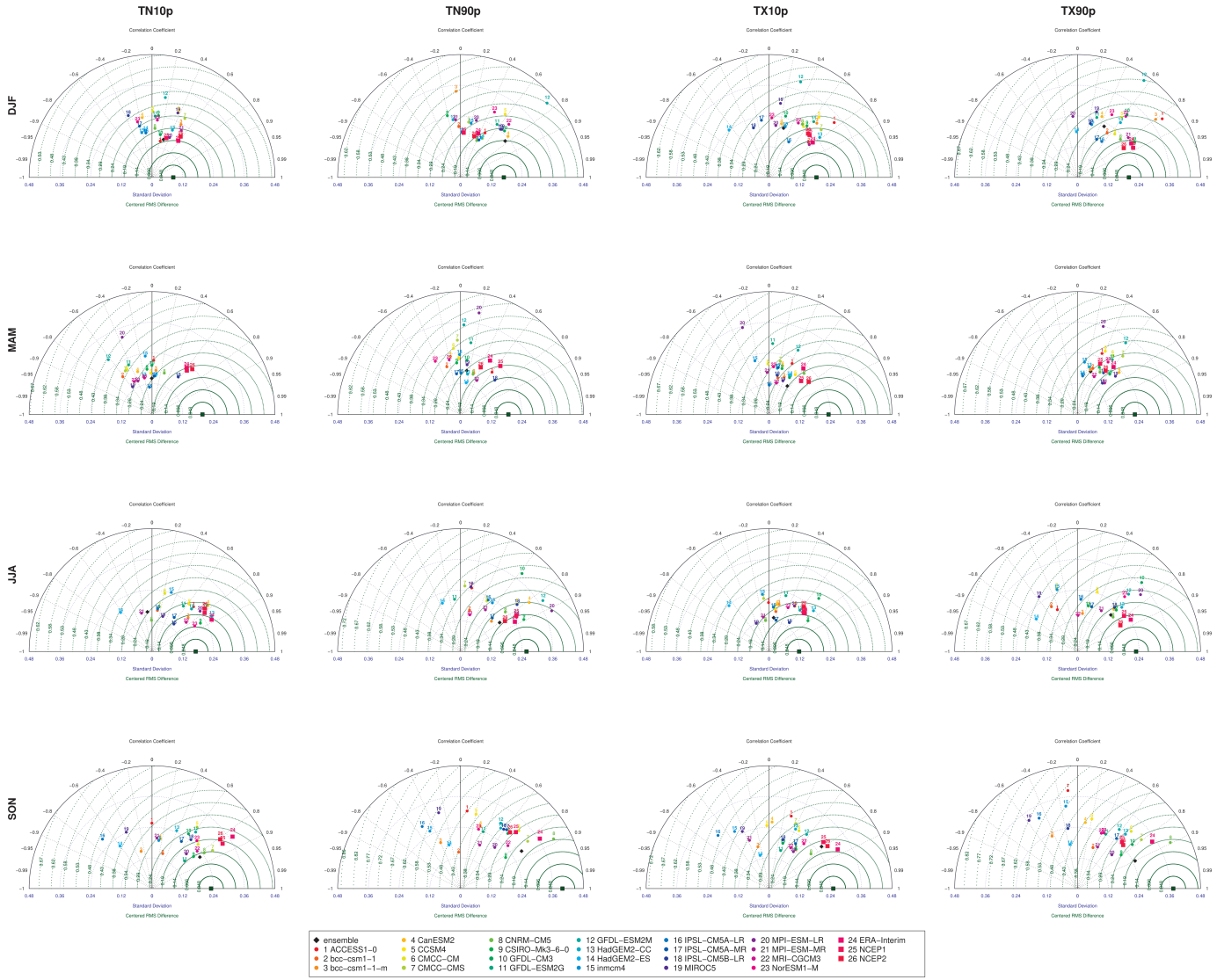

Figure S2: Taylor diagrams comparing the observed correlation between SST3.4 and extreme temperature indices of HADEX3 against reanalyses and modeled correlations.

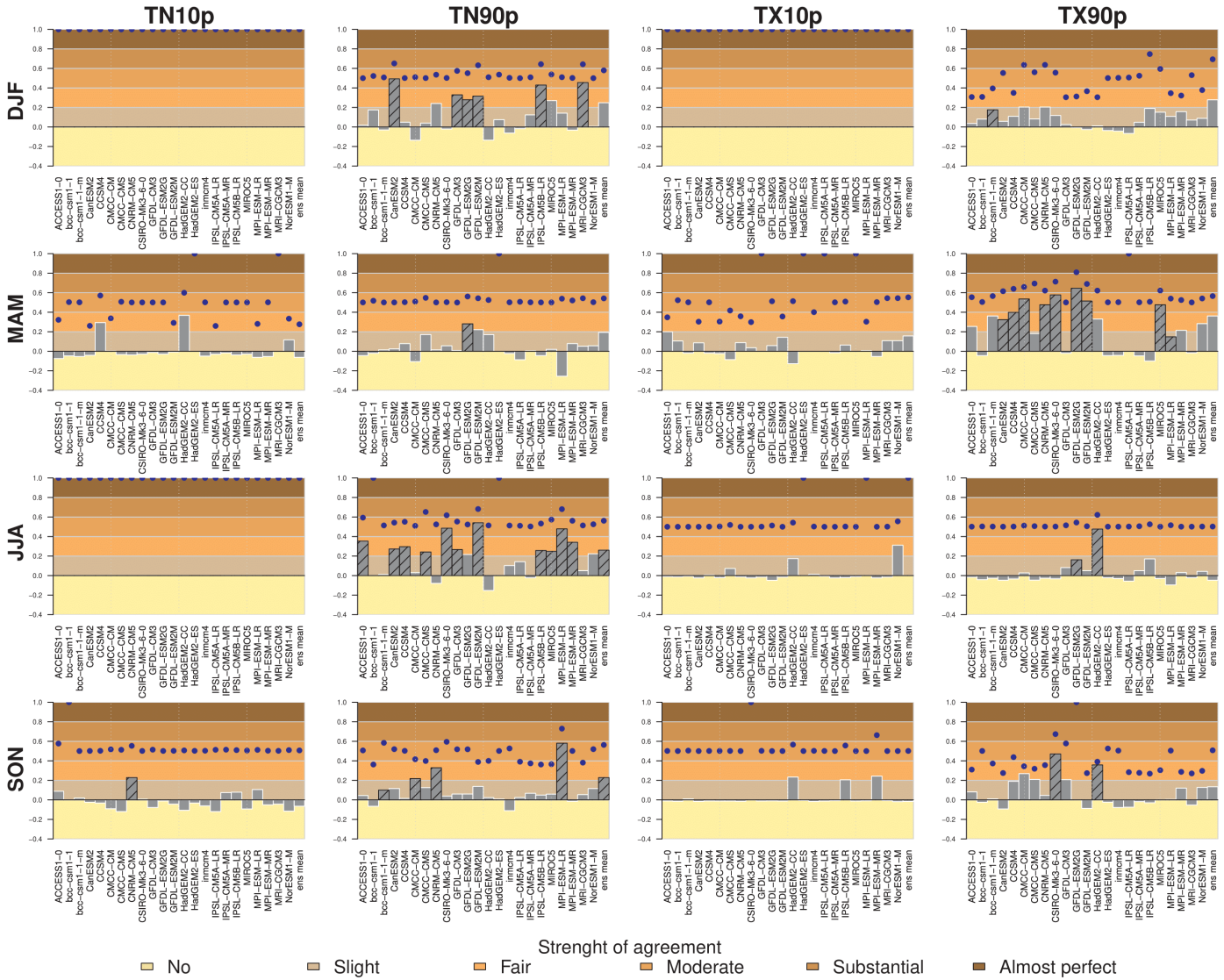

Figure S3: Cohen's Kappa (bars) and Mapcurves (points) indices comparing slopes of the quantile regression (HadEX3 data set against GCMs). The strength of agreement of Cohen's Kappa coefficient is shown in colors and dashed indicate significant kappa at 5%.

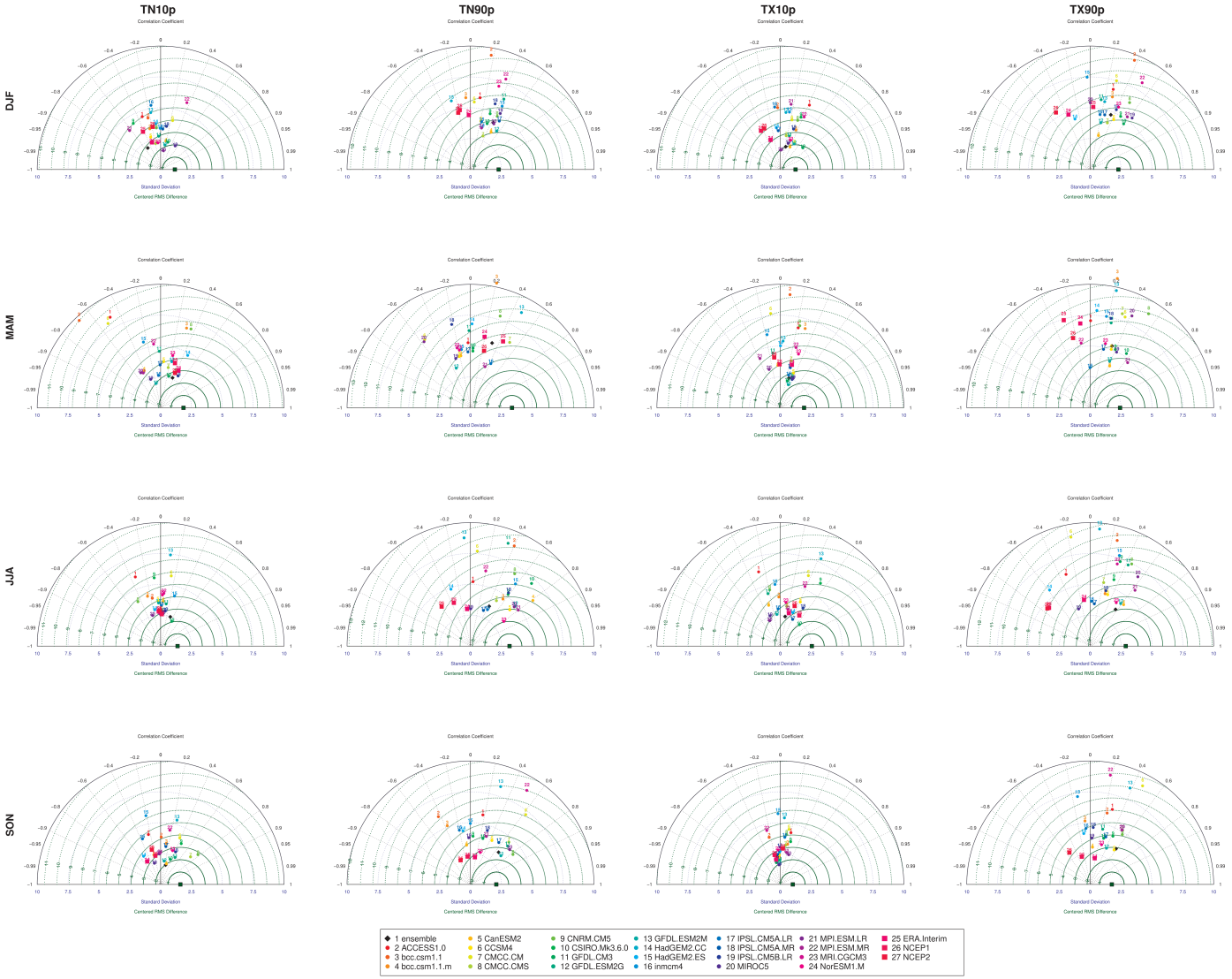

Figure S4: Taylor diagrams comparing the observed slope of the quantile regression between SST3.4 and extreme temperature indices of HADEX3 against reanalyses and modeled slopes.

Text S1. Temporal performance metrics

The correlation between SST3.4 and extreme temperature indices measure the strength of the joint variability of the variables. Therefore, we also decided to evaluate the performance of the GCMs in simulating the variability of each variable individually. We tested the homogeneity of variances of the observed and modeled SST3.4 and extreme indices by using the non-parametric Fligner-Killeen test with a significance level of 5%, which is more robust against departures from normality of the variables (Fligner & Killeen 1976). The null hypothesis states that the variances in each of the samples are the same. More details are in Conover et al. (1981).

We studied if the GCMs were capable of representing the temporal variability of the SST3.4 and the extreme temperature indices. Table S1 shows the observed and modeled variances. The null hypothesis of homogeneity of the variances could not be rejected for several models. In general, the ensemble-mean has a reduced variability that significantly differs from the observed one since the average of the various models tends to flatten the curve. This underestimation of the variances of SST3.4 by the ensemble-mean might be responsible for the few grid points with significant values observed in Figures 6-9. On the other hand, all the models that correctly simulate the association between the SST3.4 and TN90p during winter and spring (CanESM2, CNRM-CM5, CSIRO-Mk3-6-0, HadGEM2-CC, IPSL-CM5B-LR, and NorESM1-M) showed a variance of SST3.4 similar to the observed.

The TN90p variances are generally overestimated by the GCMs and are underestimated by the ensemble (Figure S5-S6). Nevertheless, we observed some models with good performances in simulating the correlations between SST3.4 and TN90p presented similar variance to the observed in almost all the region, e. g. CanESM2 and CSIRO-Mk3-6-0; while other models with good performances in simulating the correlations between SST3.4 and TN90p fail in representing the observed variance in several grid points, e. g. HadGEM2-CC and IPSL-CM5B-LR.

Table S1: Variances of the observed (HADISST) and modeled SST3.4 [°C^2]. Variances significantly different from observations at 5% are in bold and with an asterisk.

	MAM	JJA	SON	DJF
HADISST (obs)	0.45	0.43	0.89	1.17
ACCESS1-0	0.28	0.28	0.68	0.67
bcc-csm1-1	0.21	0.65	0.96	0.57
bcc-csm1-1-m	**0.96***	**1.63***	2.82	**2.48***
CanESM2	0.78	0.57	1.16	1.15
CCSM4	**1.06***	**1.05***	1.67	1.91
CMCC-CM	0.28	0.21	0.34	**0.37***
CMCC-CMS	**0.95***	**1.13***	1.31	1.50
CNRM-CM5	0.24	0.32	0.89	0.88
CSIRO-Mk3-6-0	0.75	0.39	0.49	0.59
GFDL-CM3	0.46	0.78	1.28	1.04
GFDL-ESM2G	0.35	0.32	0.70	0.76
GFDL-ESM2M	**2.15***	**1.74***	**2.48***	**3.24***
HadGEM2-CC	0.33	0.23	0.49	0.61
HadGEM2-ES	0.37	0.73	1.14	1.22
inmcm4	0.28	0.22	**0.25***	**0.32***
IPSL-CM5A-LR	0.56	0.74	0.51	0.61
IPSL-CM5A-MR	0.48	0.75	0.73	0.62
IPSL-CM5B-LR	0.60	0.59	0.79	0.75
MIROC5	0.92	1.29	1.64	1.27
MPI-ESM-LR	0.90	0.80	0.86	1.14
MPI-ESM-MR	0.51	0.40	0.40	0.49
MRI-CGCM3	0.36	0.19	**0.27***	**0.26***
NorESM1-M	0.37	0.68	0.83	0.75
Ensemble mean	**0.03***	**0.03***	**0.04***	**0.04***

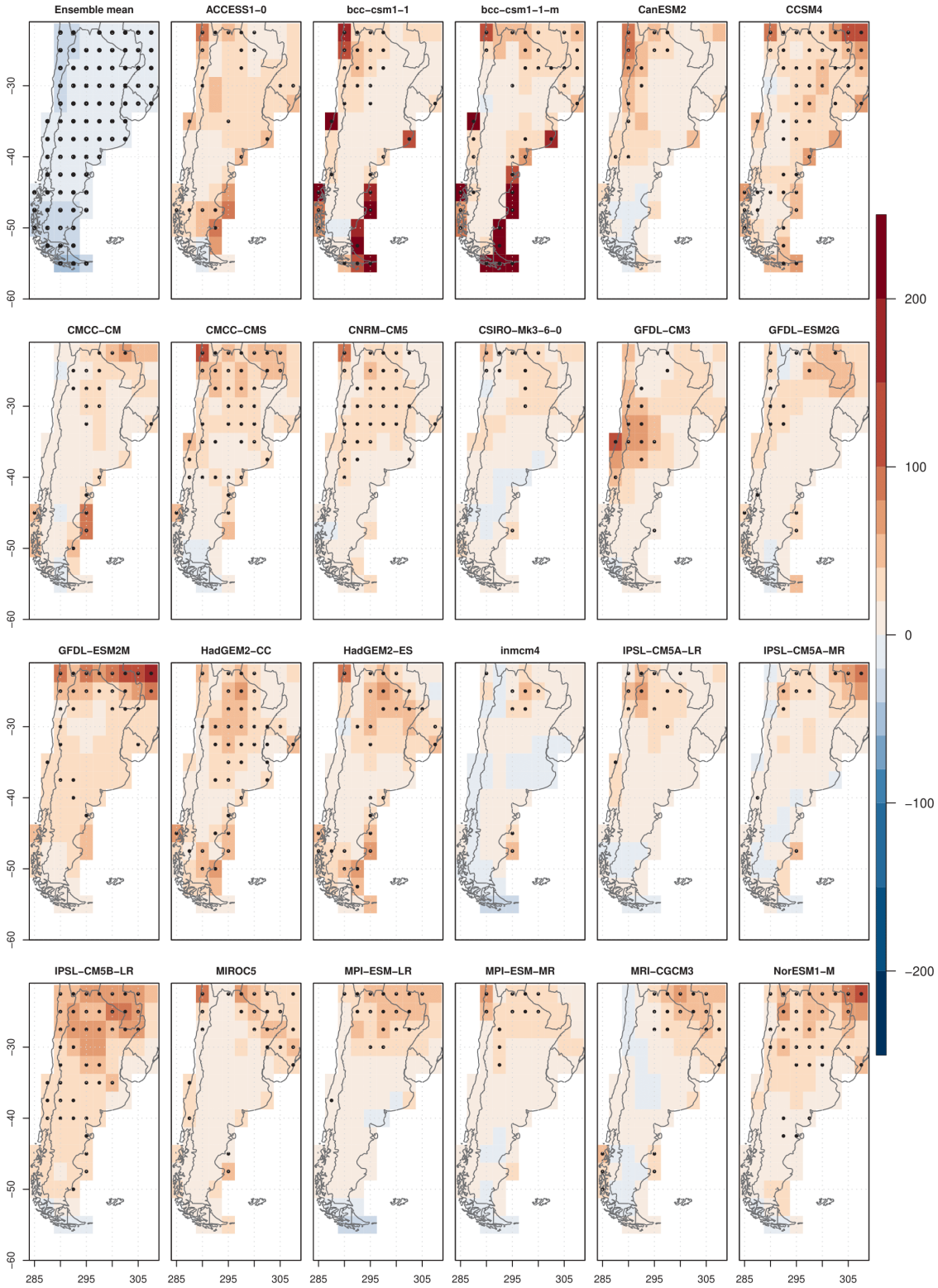

Figure S5: Difference between the observed and modeled variance for JJA TN90p. Significant differences are showed with stippling.

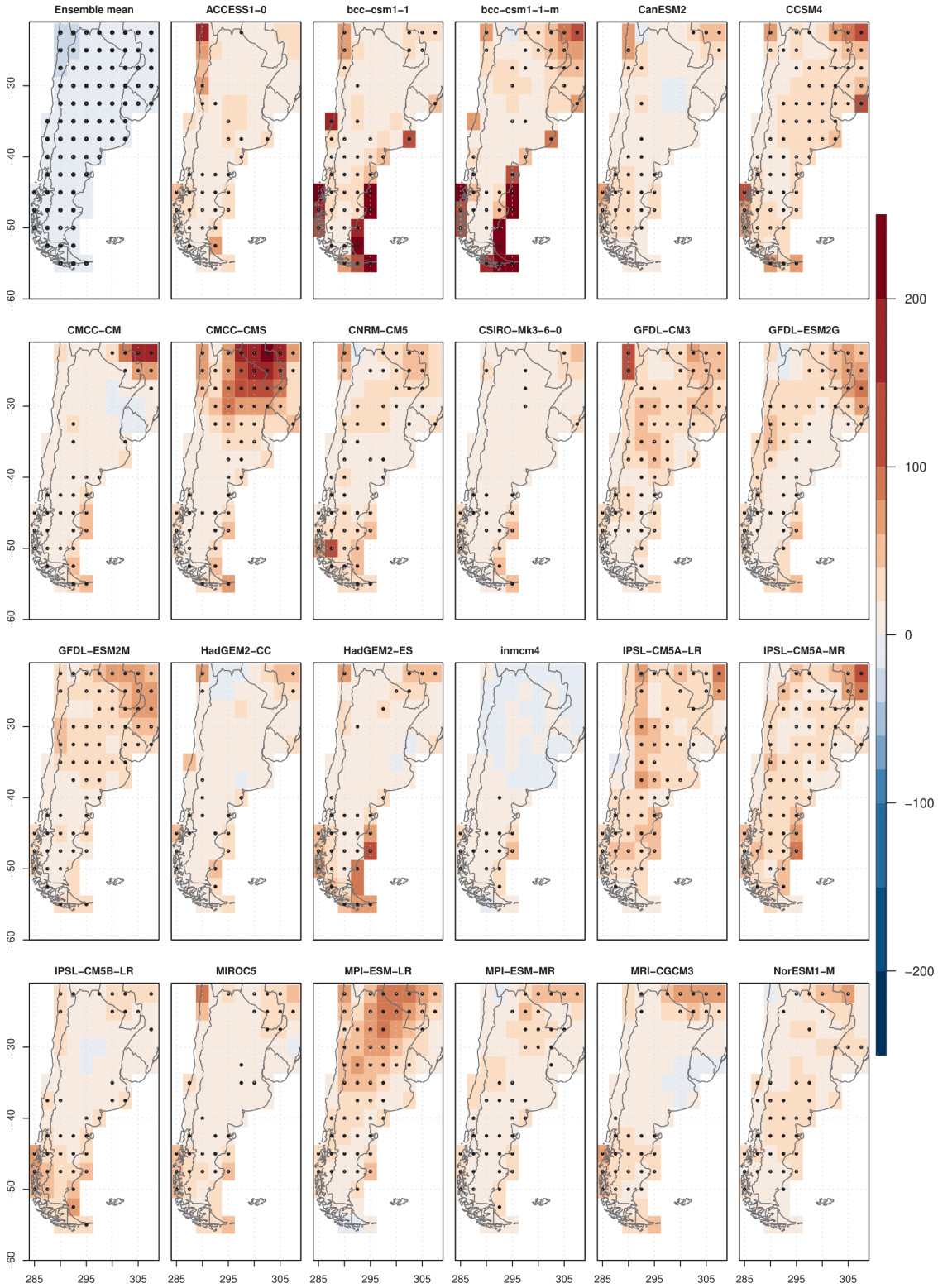

Figure S6: Idem figure 13 for SON TN90p.

References:

Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23:351–361 https://doi.org/10.1080/00401706.1981.10487680

Fligner MA, Killeen TJ (1976) Distribution-free two-sample tests for scale. J Am Stat Assoc 71:210–213 https://doi.org/10.1080/01621459.1976.10481517