Supplementary Information to:

Temperature variability projections remain uncertain after constraining them to best performing Large Ensembles

By Laura Suarez-Gutierrez [1,2,3] & Nicola Maher [4,5]

1. Institute for Atmospheric and Climate Science, ETH Zürich, Zurich, Switzerland
2. Laboratoire des Sciences du Climat et de l'Environnement, Institut Pierre-Simon Laplace, Paris, France
3. Meteorology and Air Quality Group, Wageningen University & Research, Wageningen, the Netherlands
4. Research School of Earth Sciences, The Australian National University, Canberra, Australia
5. ARC Centre of Excellence for Weather of the 21st Century

The Supplementary Information (SI) file provides additional data and analyses supporting the findings of the main manuscript. Section 1 of this SI includes the full account of rank-frequency temperature evaluation results for all 11 models included in the main manuscript. Section 2 describes the specific timing when warming levels are reach in each model. Section 3 provides a performance assessment of the constrained and non-constrained ensembles in capturing future variability change against the best preforming model, CESM2-LE.

1. Rank-Frequency Performance Evaluation of Temperature Variability

This section encloses the full temperature variability rank-frequency evaluation framework results for all 11 single-model initial condition large ensembles (SMILEs) considered in this study globally, against GISTEMPv4 and ERSSTv5 observations over land and ocean regions respectively, both for detrended and non-detrended temperature anomalies. For the purposes of this evaluation, monthly mean temperature anomalies are relative to the period 1961–1990, and model output data are regridded to match the observational grid.

Enclosed figures in this section include:
For each model, first, maps for the rank-frequency evaluation for December, January, February (DJF; top) and June, July, August (JJA; bottom) monthly mean temperature anomalies, followed by time series and rank-histograms for DJF (first) and JJA (second) for spatially aggregated regions. Results are first shown for land regions for detrended data for all models, followed by results for non-detrended data for all models; then for ocean regions following the dame structure. Results are shown for the following 11 SMILEs, in order of appearance: ACCESS-ESM1.5, CanESM2, CanESM5, CESM-LE, CESM2-LE, CSIRO-MK3.6, GFDL-ESM2M, GFDL-SPEAR-MED, MIROC6, MPI-GE5 and MPI-GE6.

These results are separated into the following subsections:

- Detrended Land Surface Air Temperatures
- Non-Detrended Land Surface Air Temperatures
- Detrended Ocean Surface Temperatures
- Non-Detrended Ocean Surface Temperatures

Detrended Land Surface Air Temperatures

Rank-frequency variability evaluation framework for detrended 2m air temperature (TAS) anomalies over land grid cells.

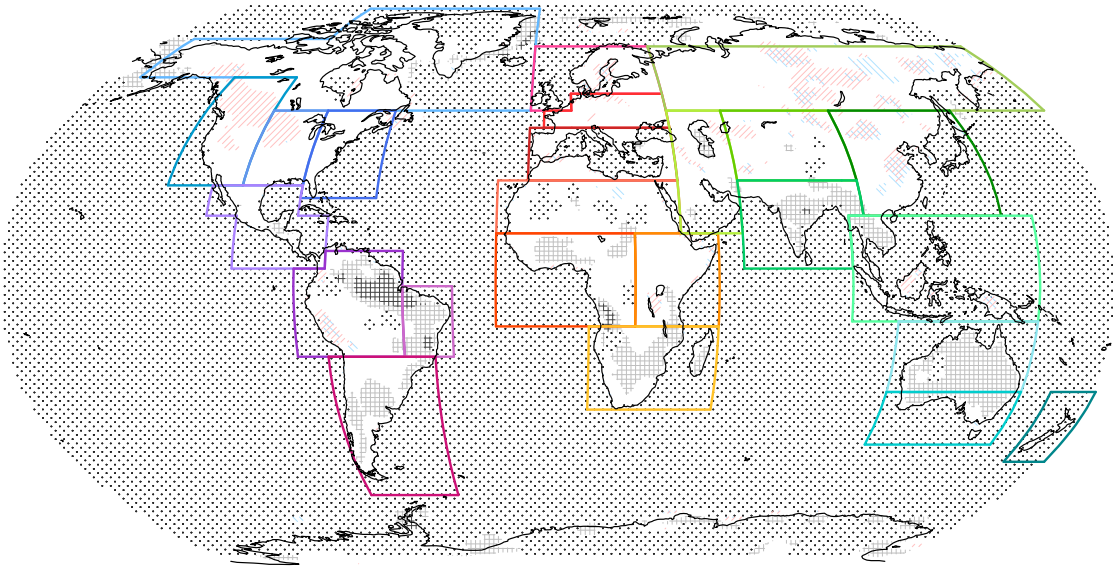
Maps show grid-cell evaluation of the simulated DJF and JJA monthly mean temperature anomalies for the 11 SMILEs included against GISTEMPv4 observations globally. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensemble (12.5th to 87.5th percentiles) for more than 80% of months (light grey) or for more than 95% of months (dark grey). Red and blue shading represents where observations are larger than the ensemble maximum (red) or smaller than the ensemble minimum (blue), respectively, for more than 8% of the months (light red and blue) or for more than 20% of the months (dark red and blue). Dotted areas represent ocean areas or grid cells where observations are missing and are therefore excluded from this analysis. Colored boxes demark the boundaries of each land region assessed. The percentage of assessed grid-cells that present none of these biases, unbiased grid-cells, in given at the top (white area). For each assessed region, models are considered to provide an adequate grid-cell representation when they exhibit more than 50% of unbiased grid cells in the region, fulfilling Criteria 2 of our evaluation framework.

Time series and rank frequency histograms show spatially aggregated DJF and JJA TAS for each land region for all 11 SMILEs. Time series show the ensemble maximum and minimum (color lines) and central 75th percentile ensemble spread (shading) are shown against observations (black dots).

Rank histograms represent the frequency of each place that observations would take in a list of ensemble members ordered by ascending temperature anomaly values. Rank histograms show the observations rank frequency accumulated for 3-rank bins (bars), the running mean rank frequency over a centered n/5 rank window (lines; for 1 to n-1 ranks) and the absolute frequencies of rank 0 and n (crosses), with n the number of ensemble members, for observations (color) and perfect model rank range (gray).

To illustrate how internal variability sampling may affect rank frequencies given the non-infinite record length considered, we include a comparison to the perfect-model rank range, which shows the range of rank frequencies that each ensemble member would take if it were observations (see Methods in main article for further details). If the rank exhibited by observations (colors) is within this perfect-model range (gray; allowing a maximum 10% deviation outside of this range), the rank frequency evaluation shows an adequate model performance, and any deviations from a perfectly frank rank histogram can be assumed to be within the range of deviations that could be caused by internal variability. In the case the observations rank frequency is within the perfect-model range for all rank windows, Criteria 1 of our evaluation framework is met and this is highlighted by a green star at the top right; if not, by a red cross. Percentages at the top left show the frequency of regionally averaged observations occurring above (red) or below (blue) ensemble limits, or clustering within the central 75th percentile range (gray), analogous to the grid-cell evaluation in Criteria 2.

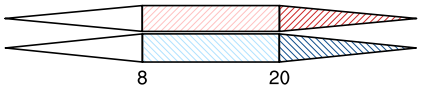
ACCESS vs GISTEMPv4 White Area = 67.4 %



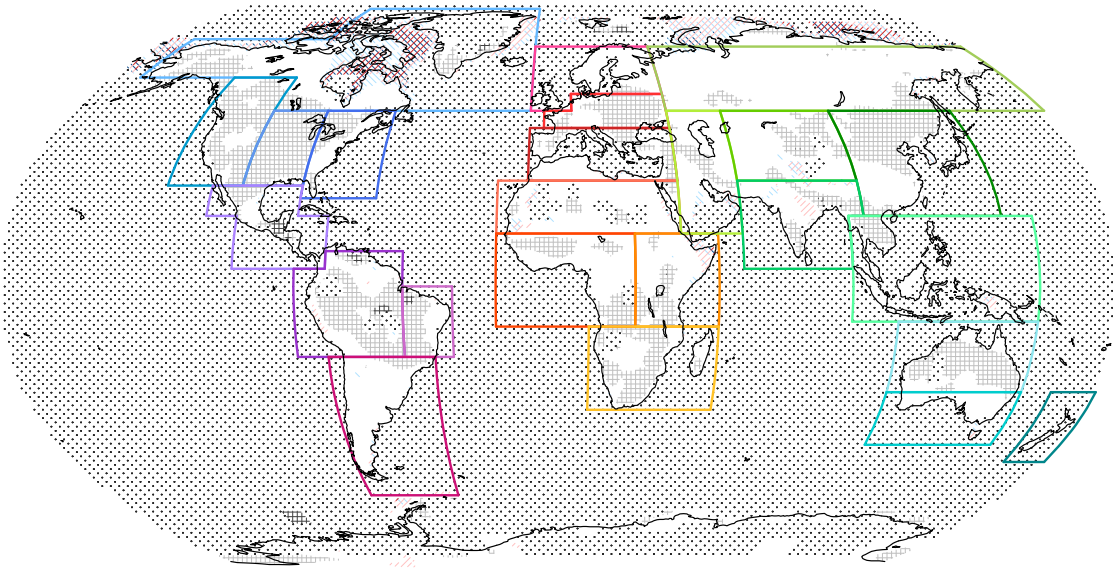
TAS DJF Obs inside central 75th percentile (%)



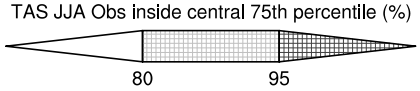
TAS DJF Obs outside ensemble spread (%)



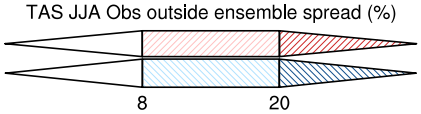
ACCESS vs GISTEMPv4 White Area = 57.1 %



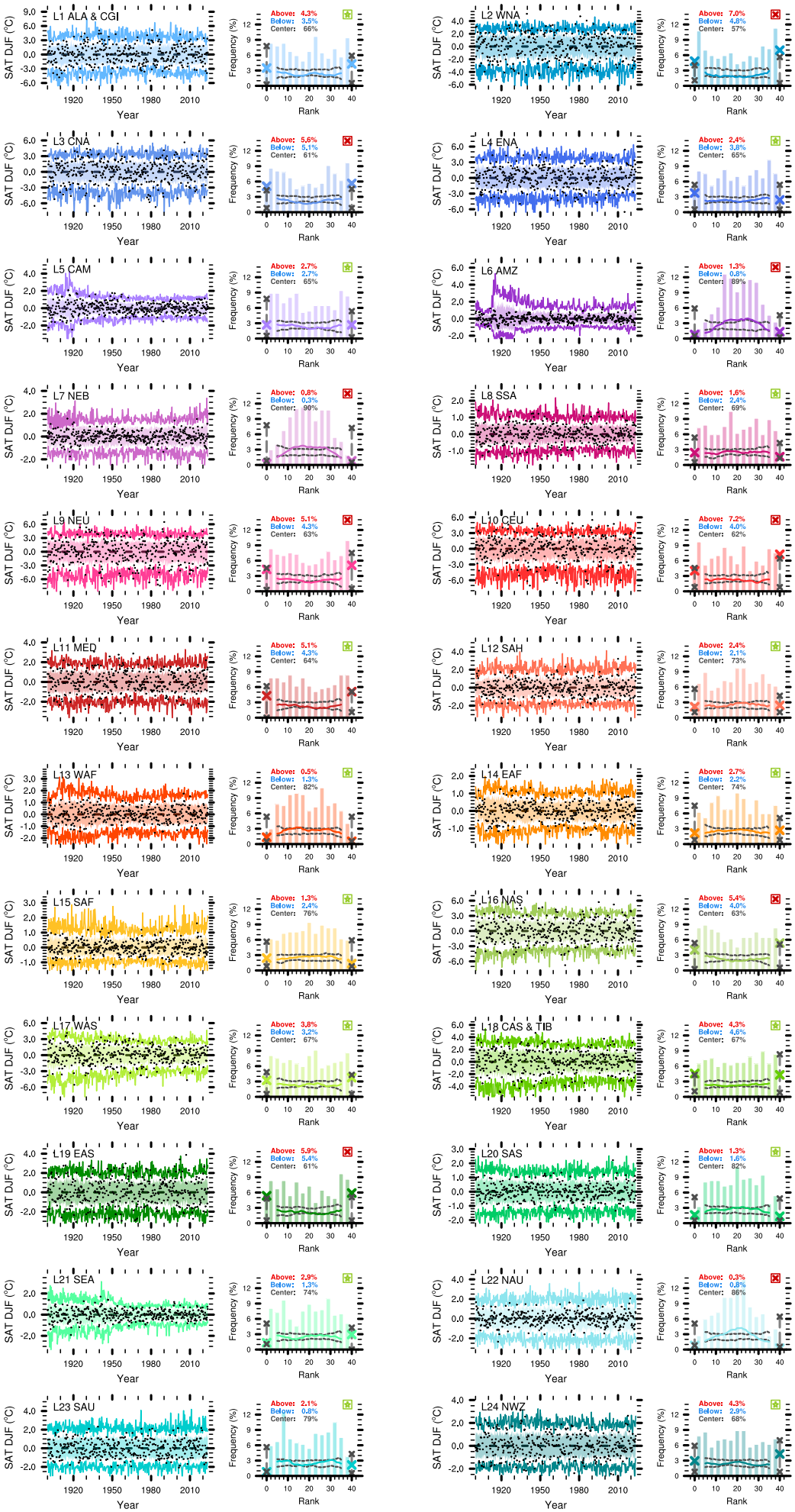
TAS JJA Obs inside central 75th percentile (%)



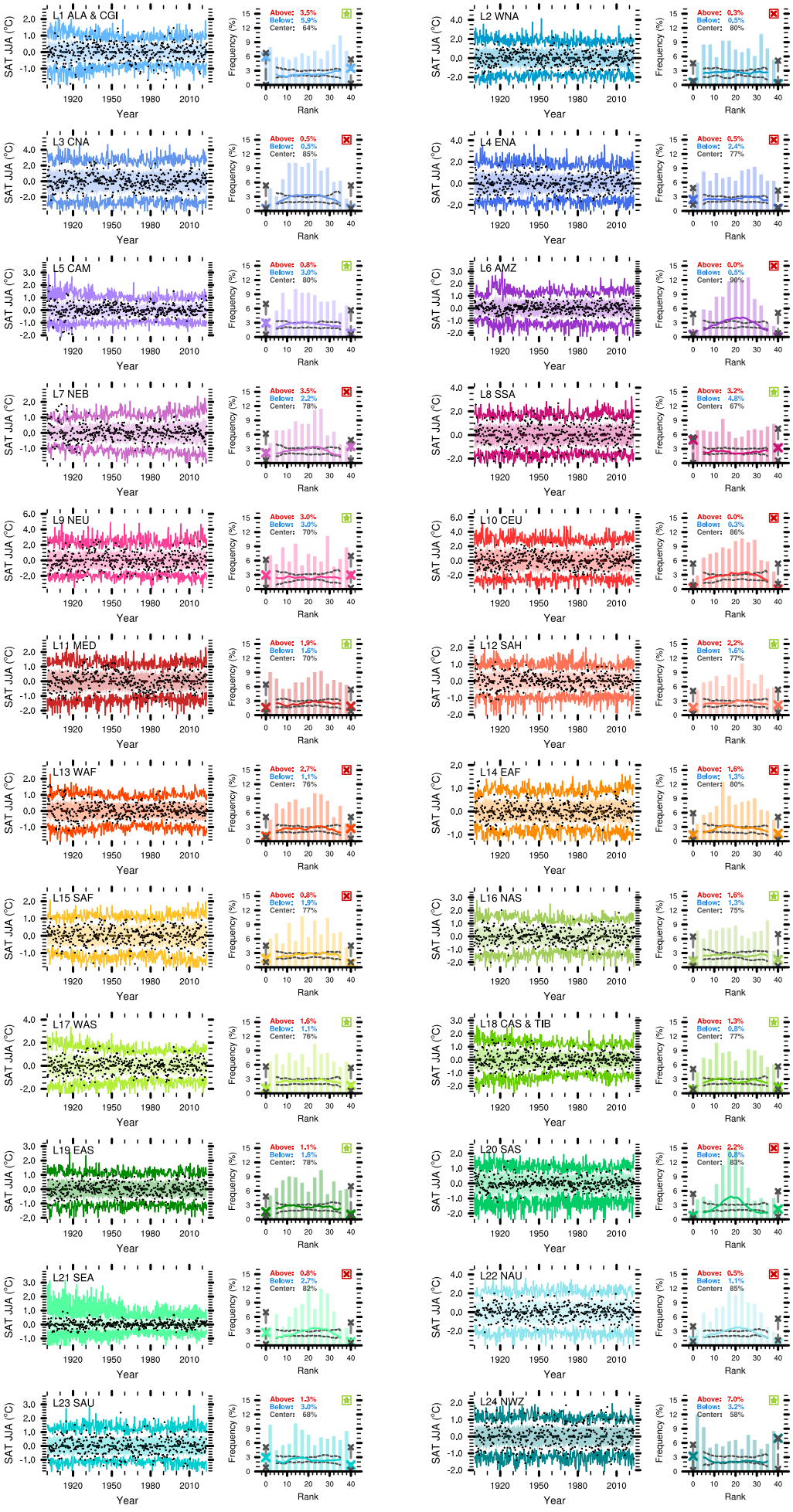
TAS JJA Obs outside ensemble spread (%)



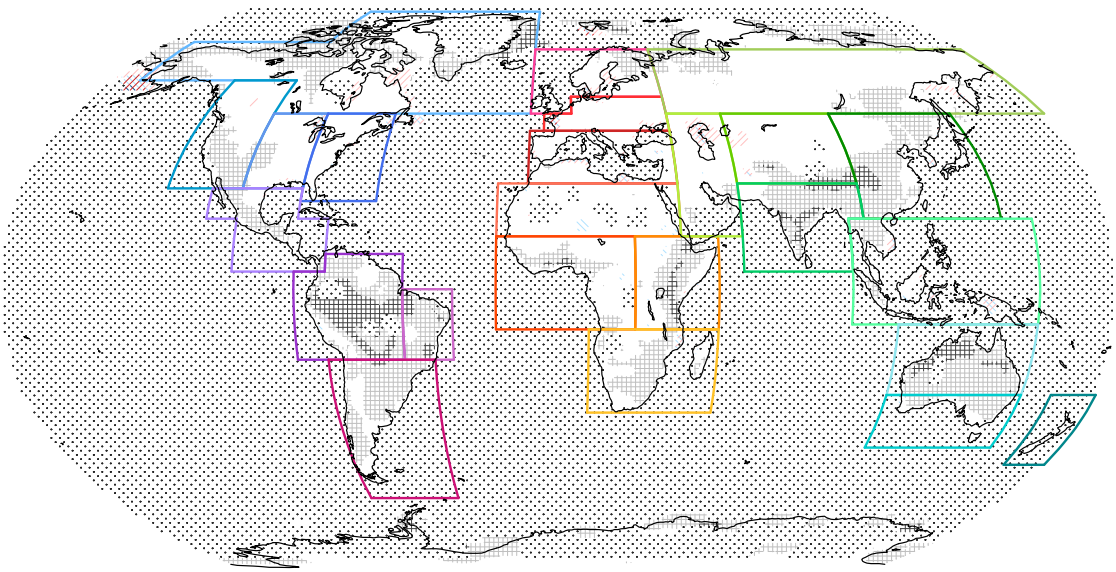
ACCESS vs GISTEMPv4 TAS DJF



ACCESS vs GISTEMPv4 TAS JJA



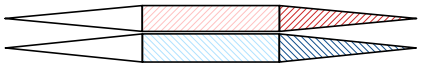
CanESM2 vs GISTEMPv4 White Area = 53.9 %



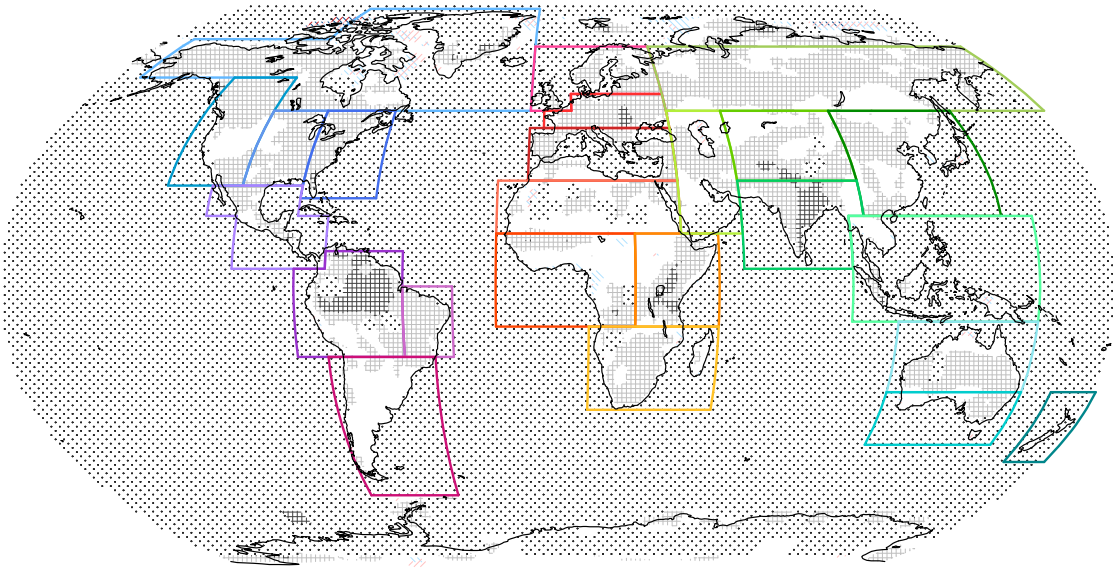
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



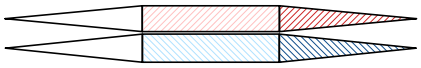
CanESM2 vs GISTEMPv4 White Area = 40.4 %



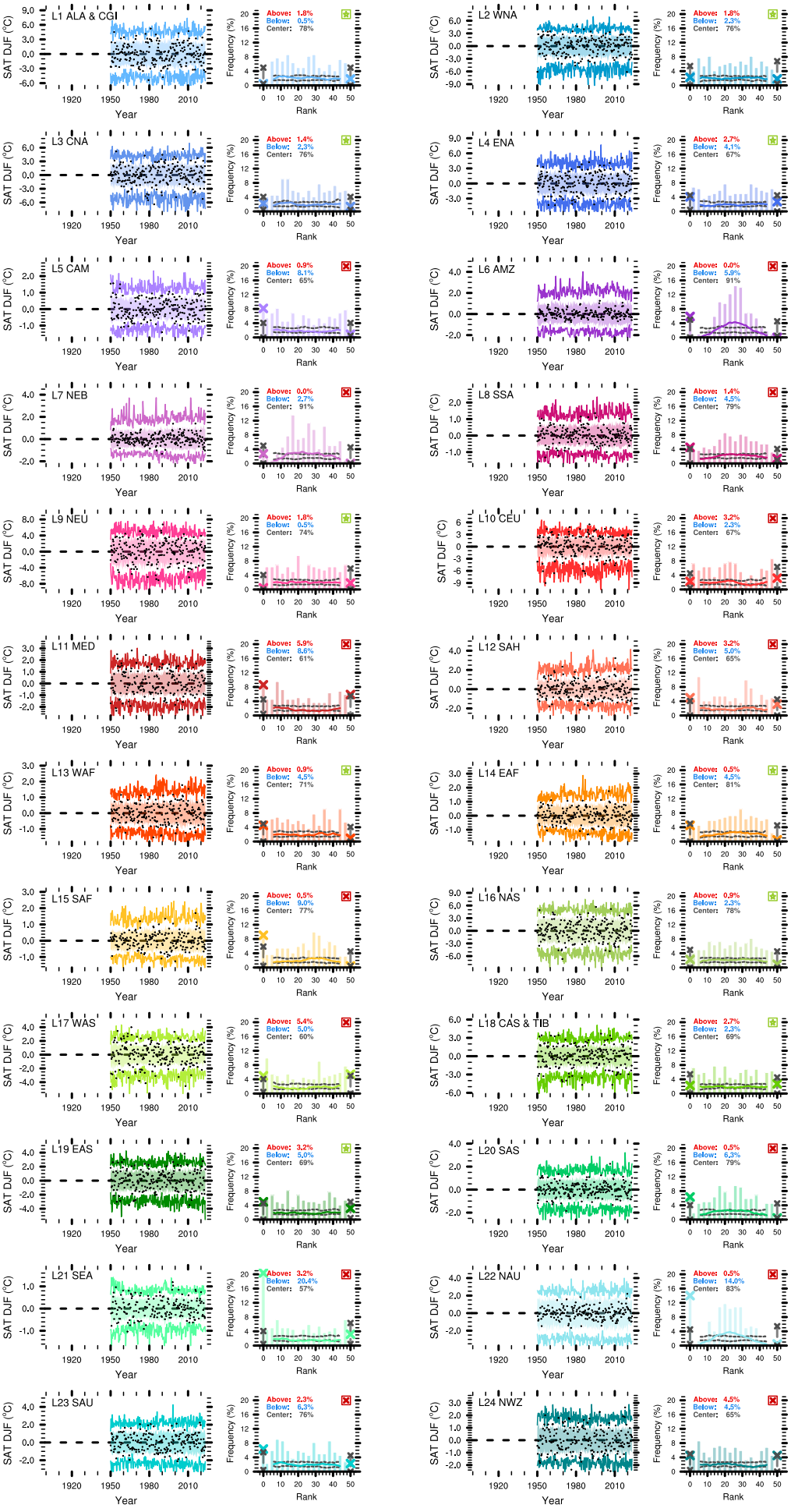
TAS JJA Obs inside central 75th percentile (%)



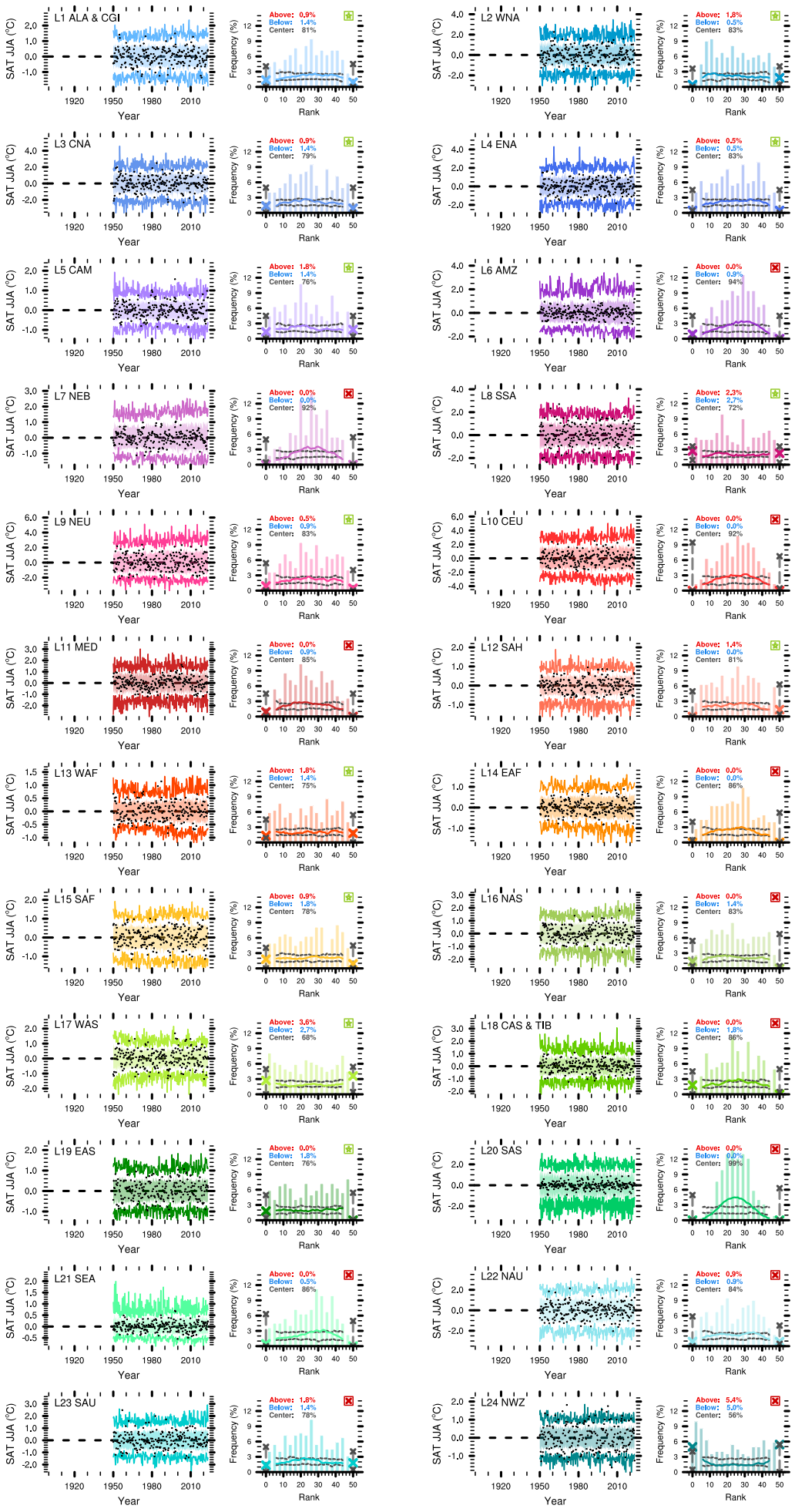
TAS JJA Obs outside ensemble spread (%)



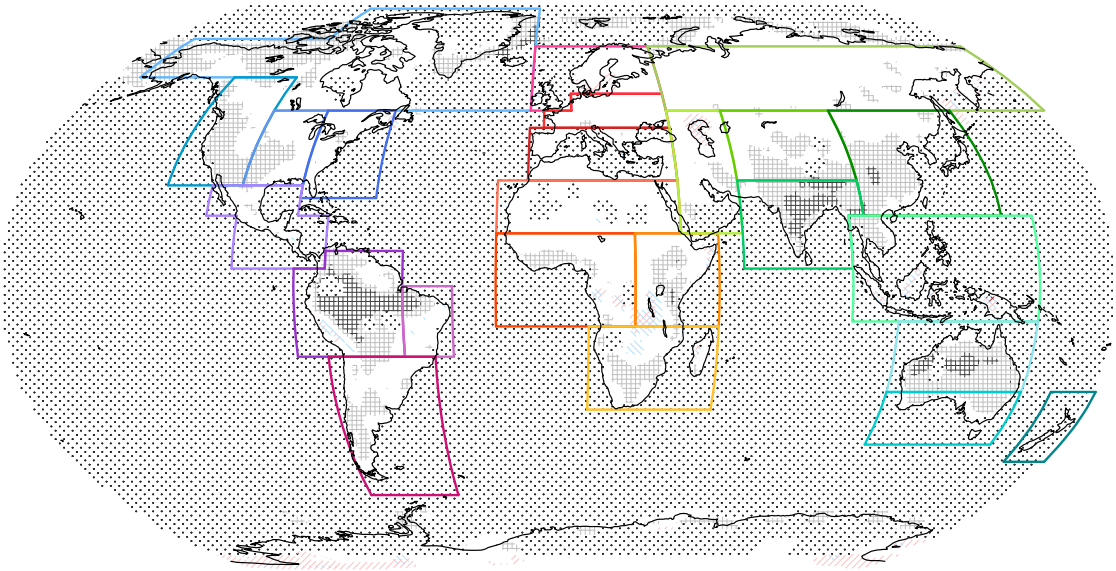
CanESM2 vs GISTEMPv4 TAS DJF



CanESM2 vs GISTEMPv4 TAS JJA



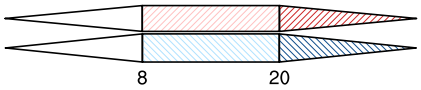
CanESM5 vs GISTEMPv4 White Area = 60.0 %



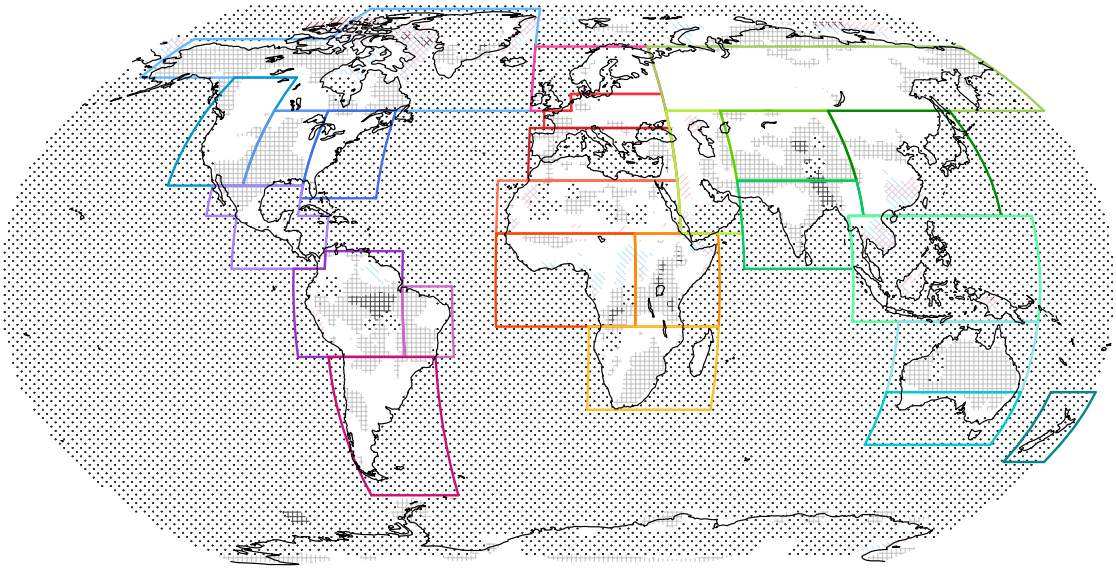
TAS DJF Obs inside central 75th percentile (%)



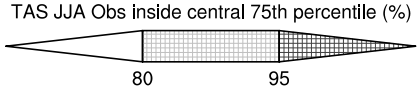
TAS DJF Obs outside ensemble spread (%)



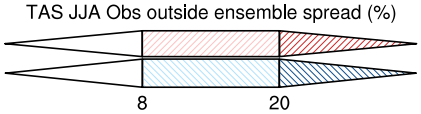
CanESM5 vs GISTEMPv4 White Area = 56.2 %



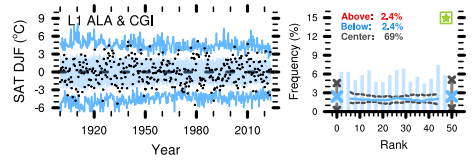
TAS JJA Obs inside central 75th percentile (%)

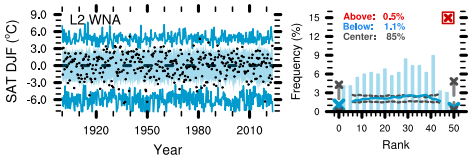


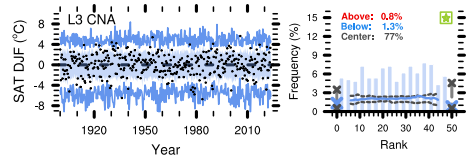
TAS JJA Obs outside ensemble spread (%)

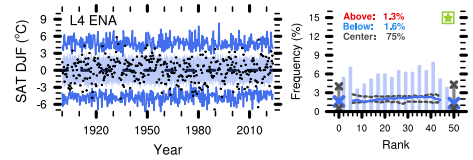


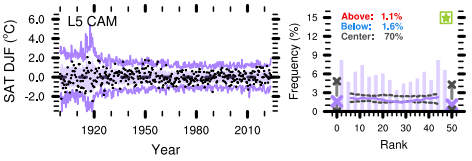
CanESM5 vs GISTEMPv4 TAS DJF

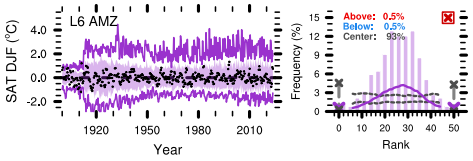


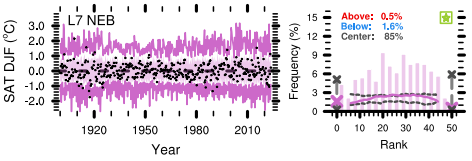


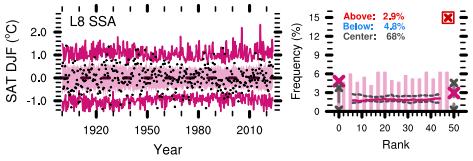


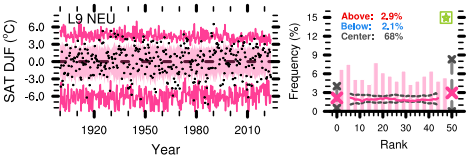


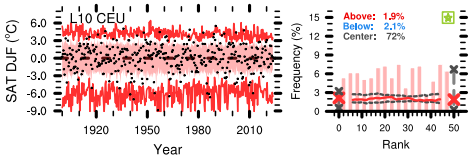


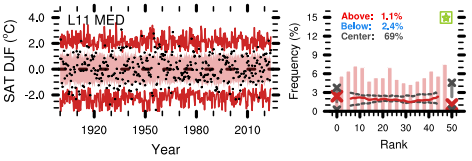


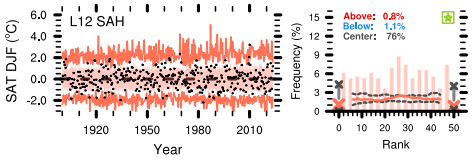


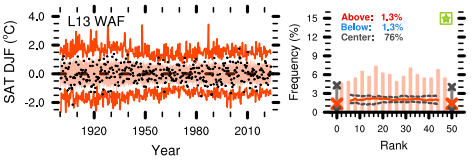


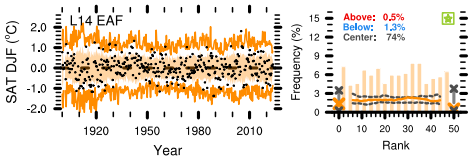


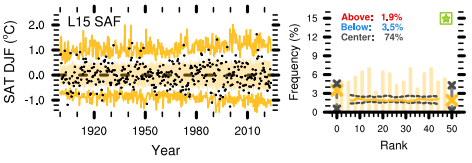


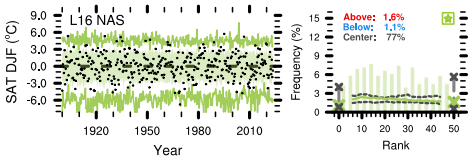


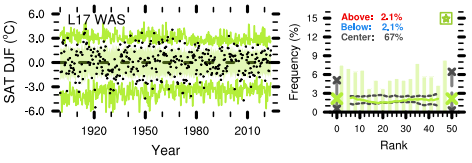


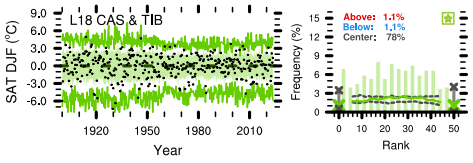


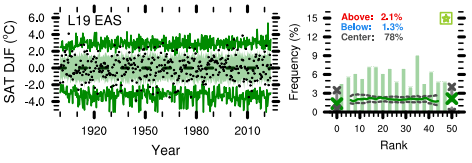


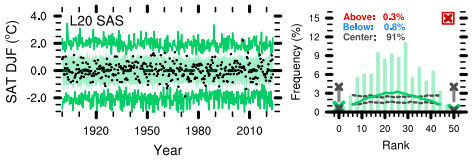


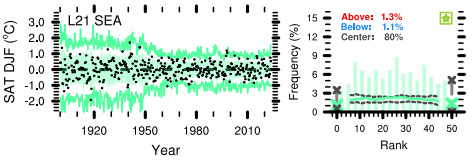


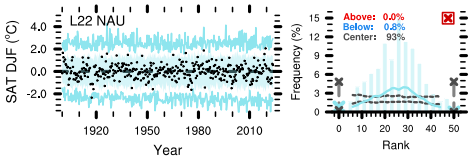


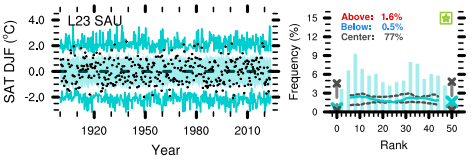


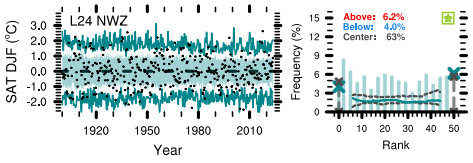




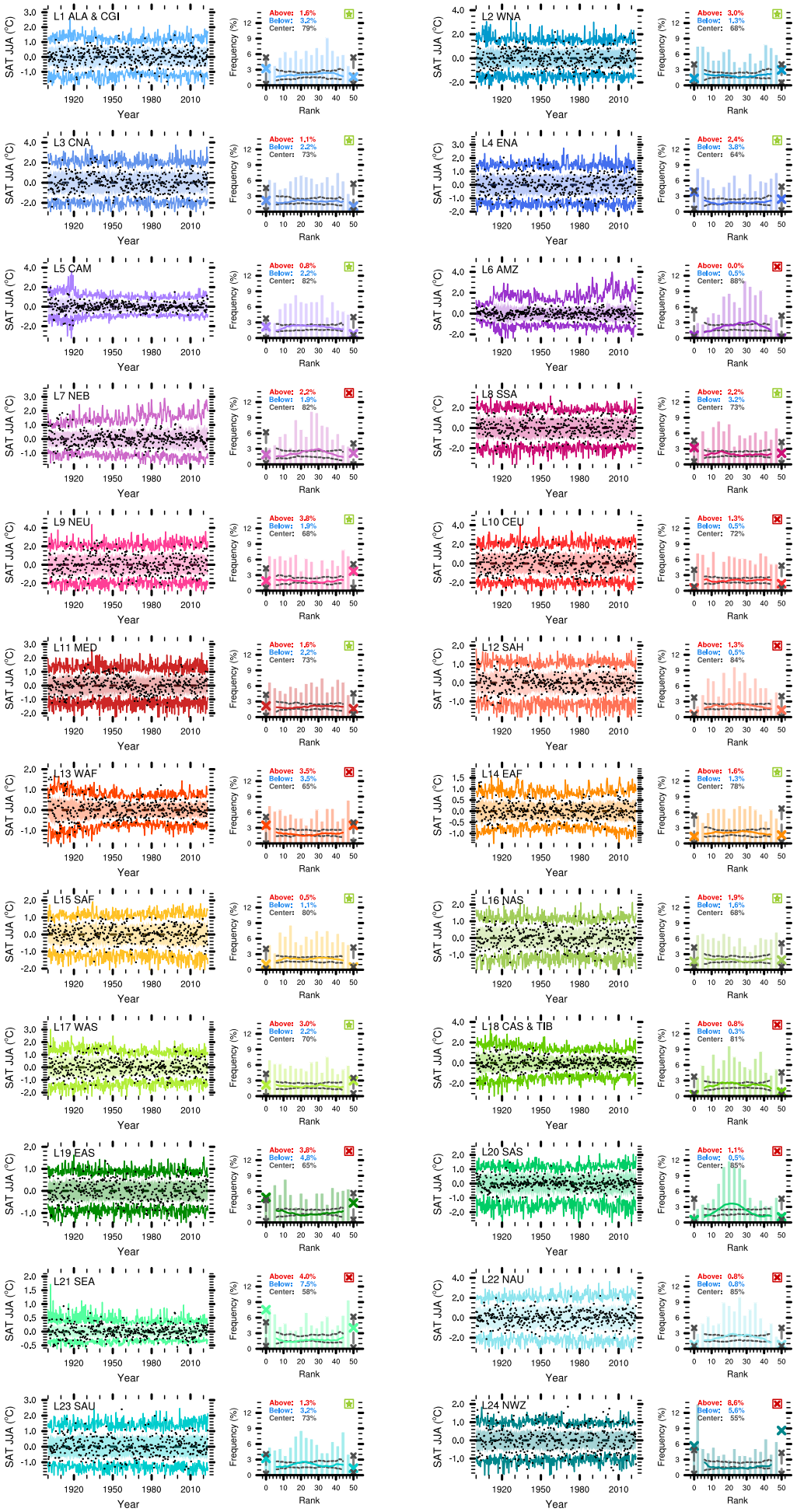




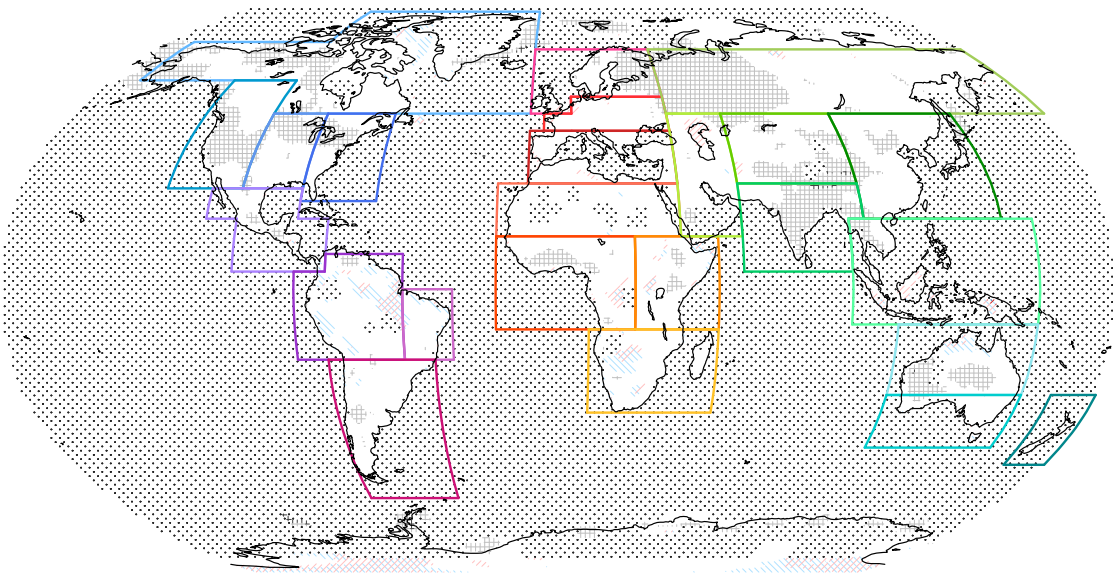




CanESM5 vs GISTEMPv4 TAS JJA



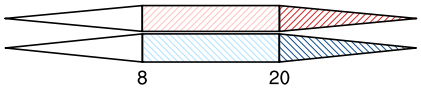
CESM-LE vs GISTEMPv4 White Area = 67.6 %



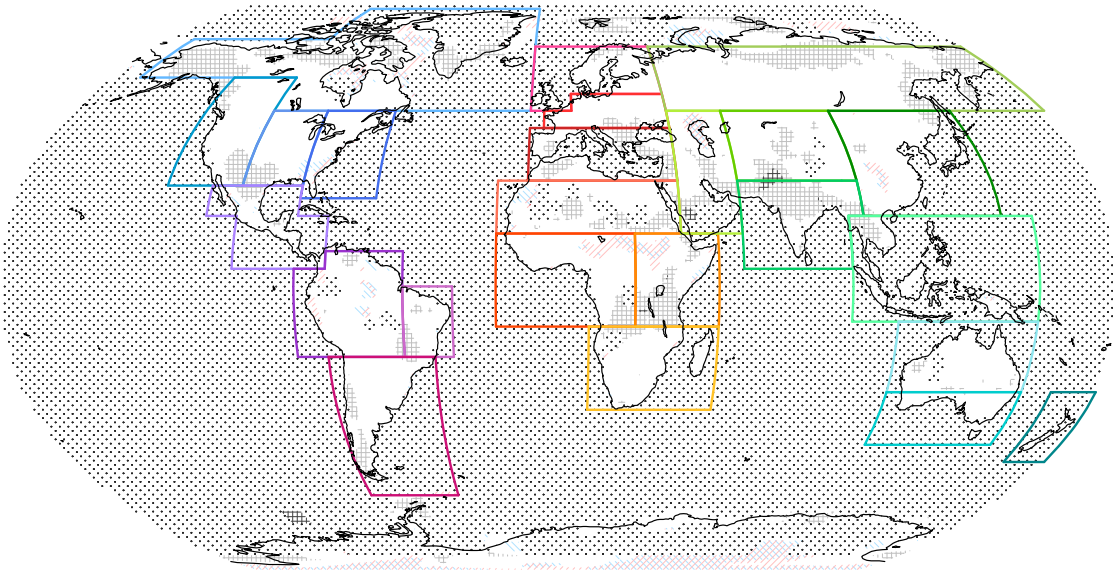
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



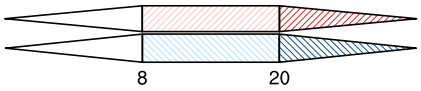
CESM-LE vs GISTEMPv4 White Area = 71.9 %



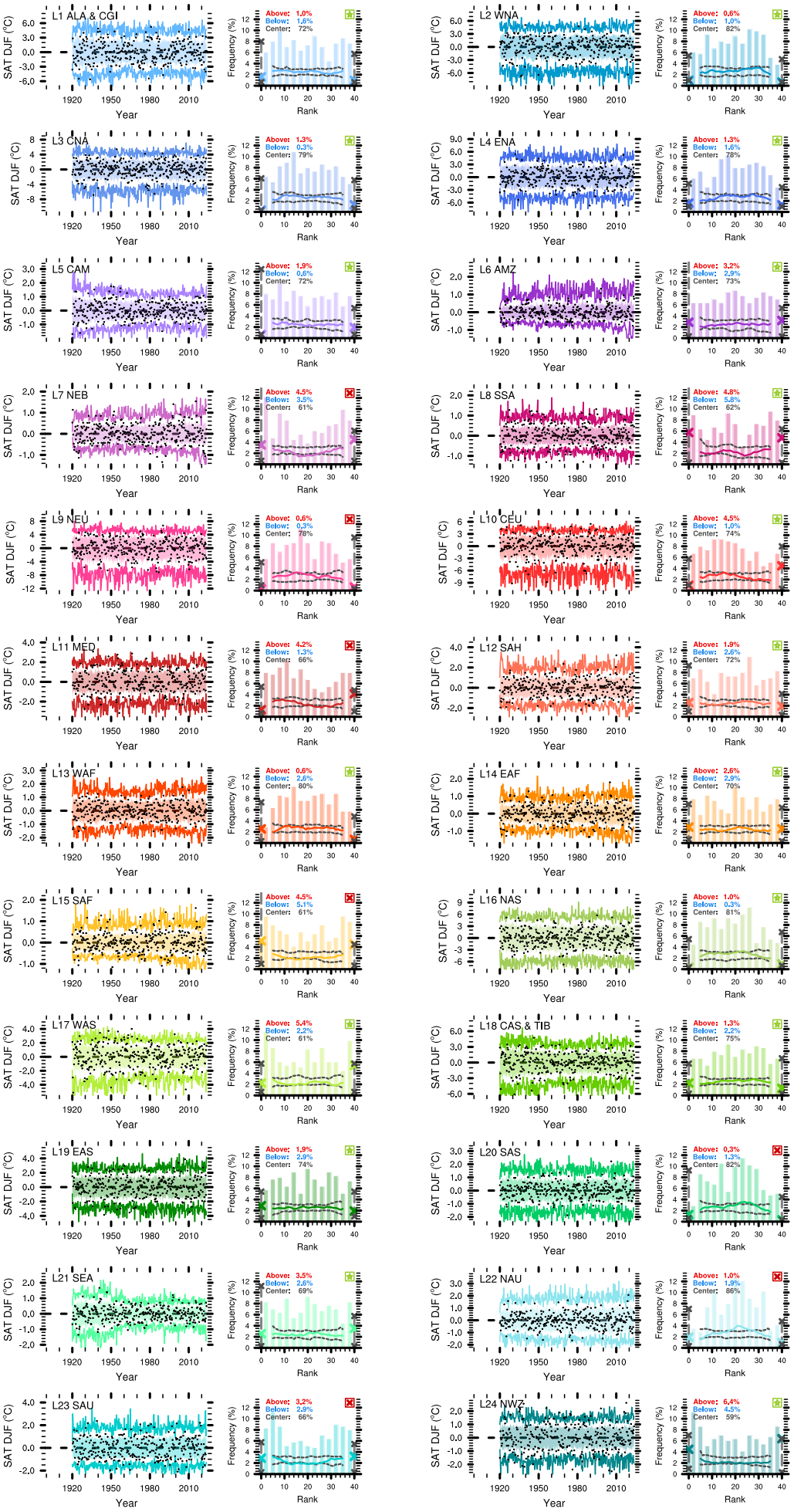
TAS JJA Obs inside central 75th percentile (%)



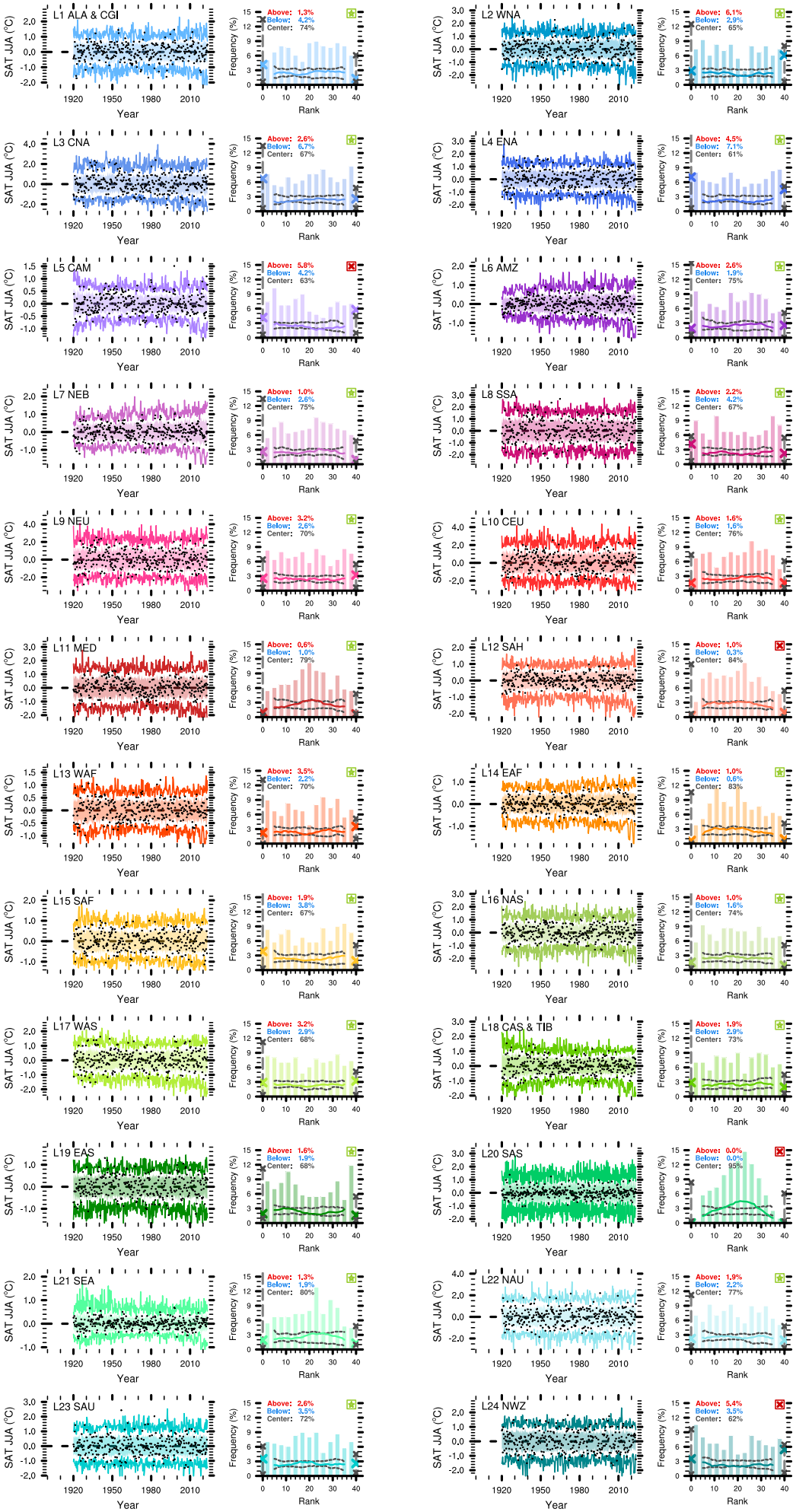
TAS JJA Obs outside ensemble spread (%)



CESM-LE vs GISTEMPv4 TAS DJF

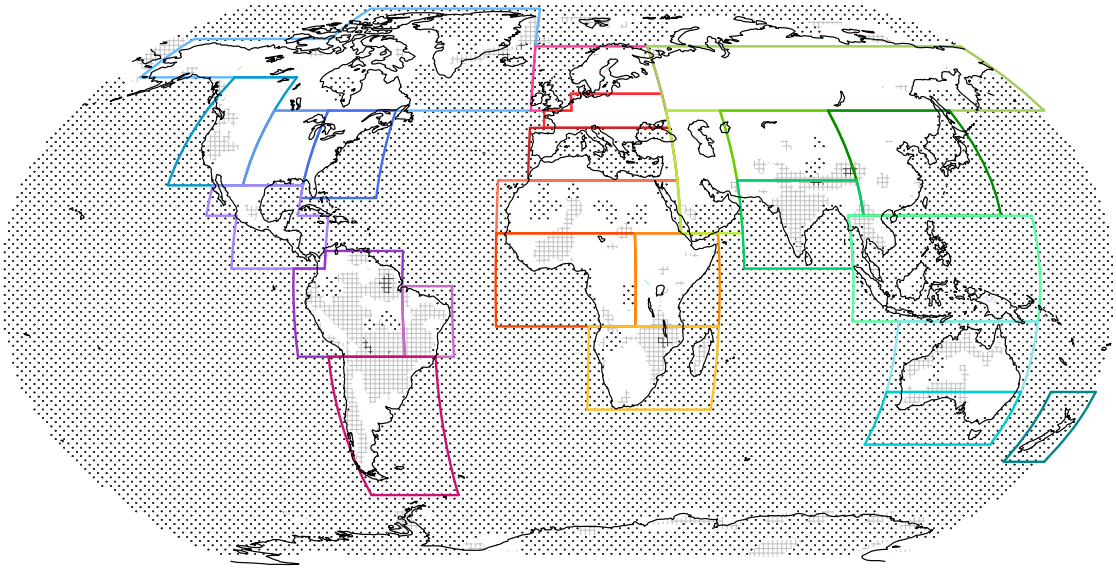


CESM-LE vs GISTEMPv4 TAS JJA



CESM2-LE vs GISTEMPv4

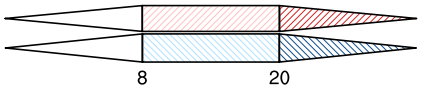
White Area = 79.0 %



TAS DJF Obs inside central 75th percentile (%)

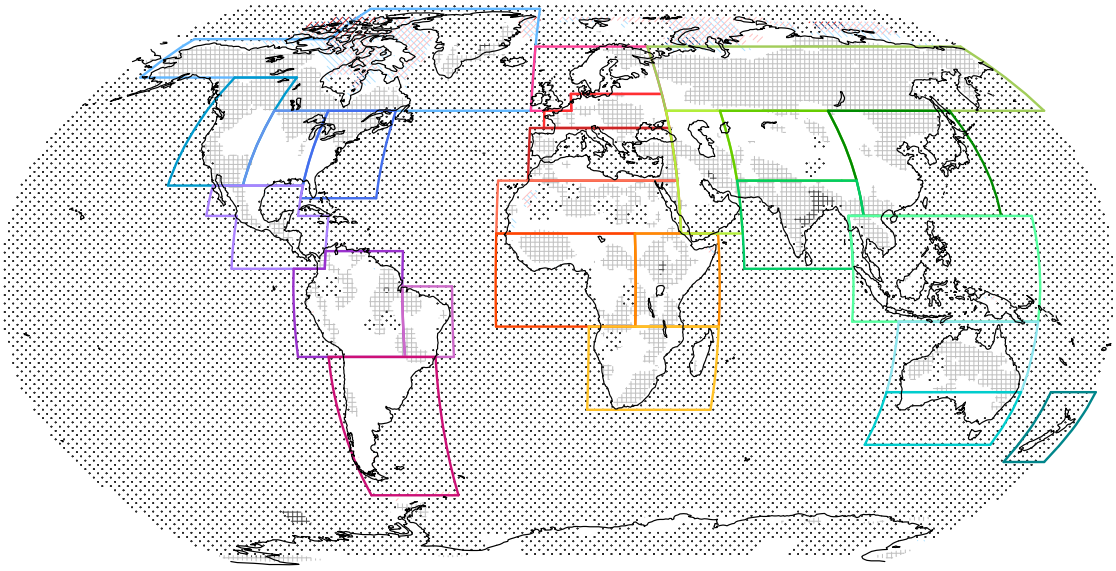


TAS DJF Obs outside ensemble spread (%)



CESM2-LE vs GISTEMPv4

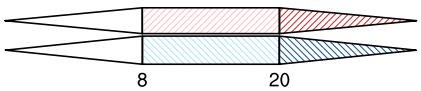
White Area = 54.8 %



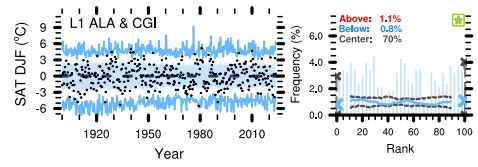
TAS JJA Obs inside central 75th percentile (%)

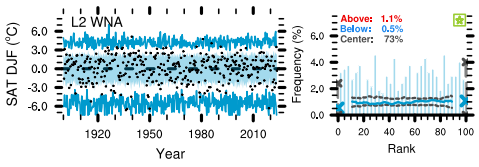


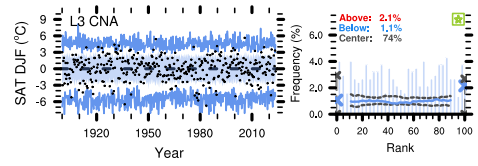
TAS JJA Obs outside ensemble spread (%)

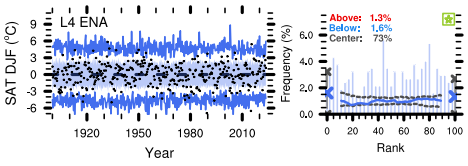


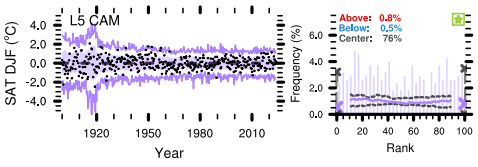
CESM2-LE vs GISTEMPv4 TAS DJF

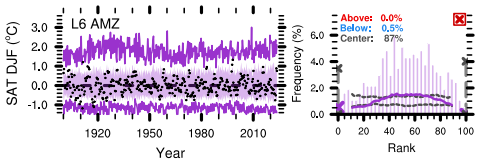


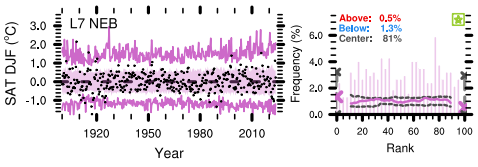


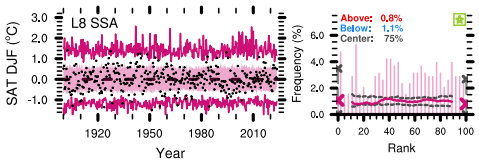


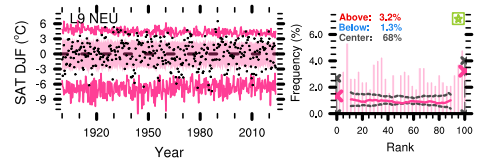


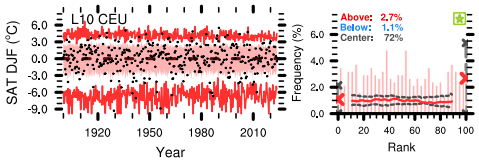


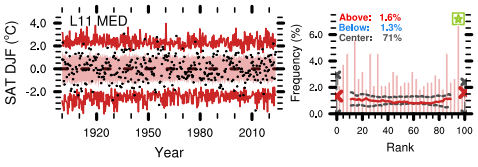


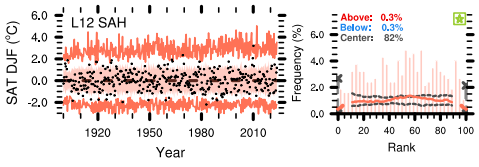


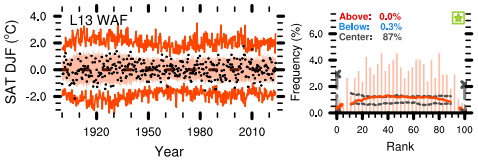


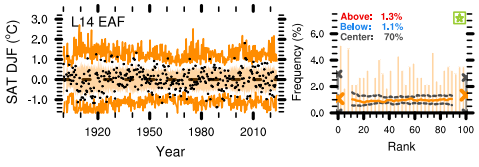


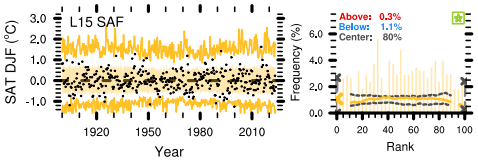


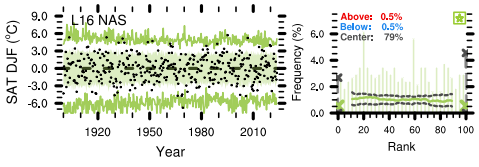


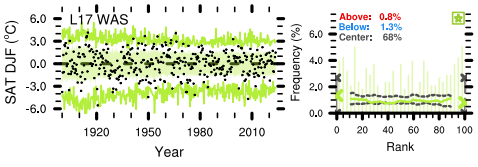


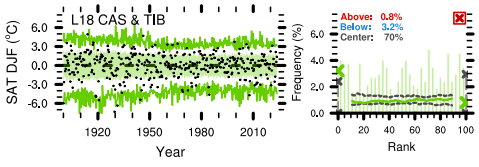


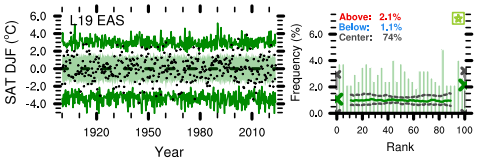


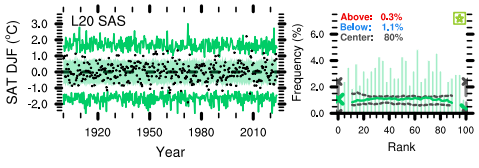


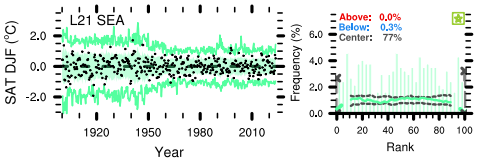


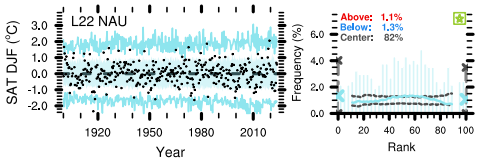


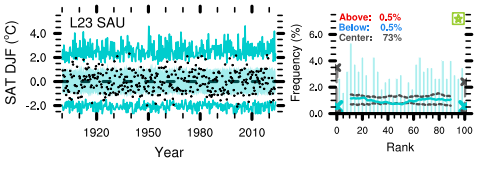


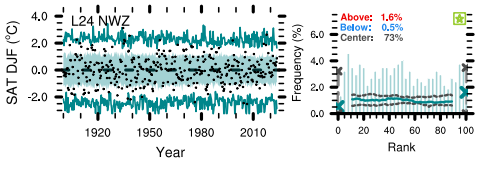




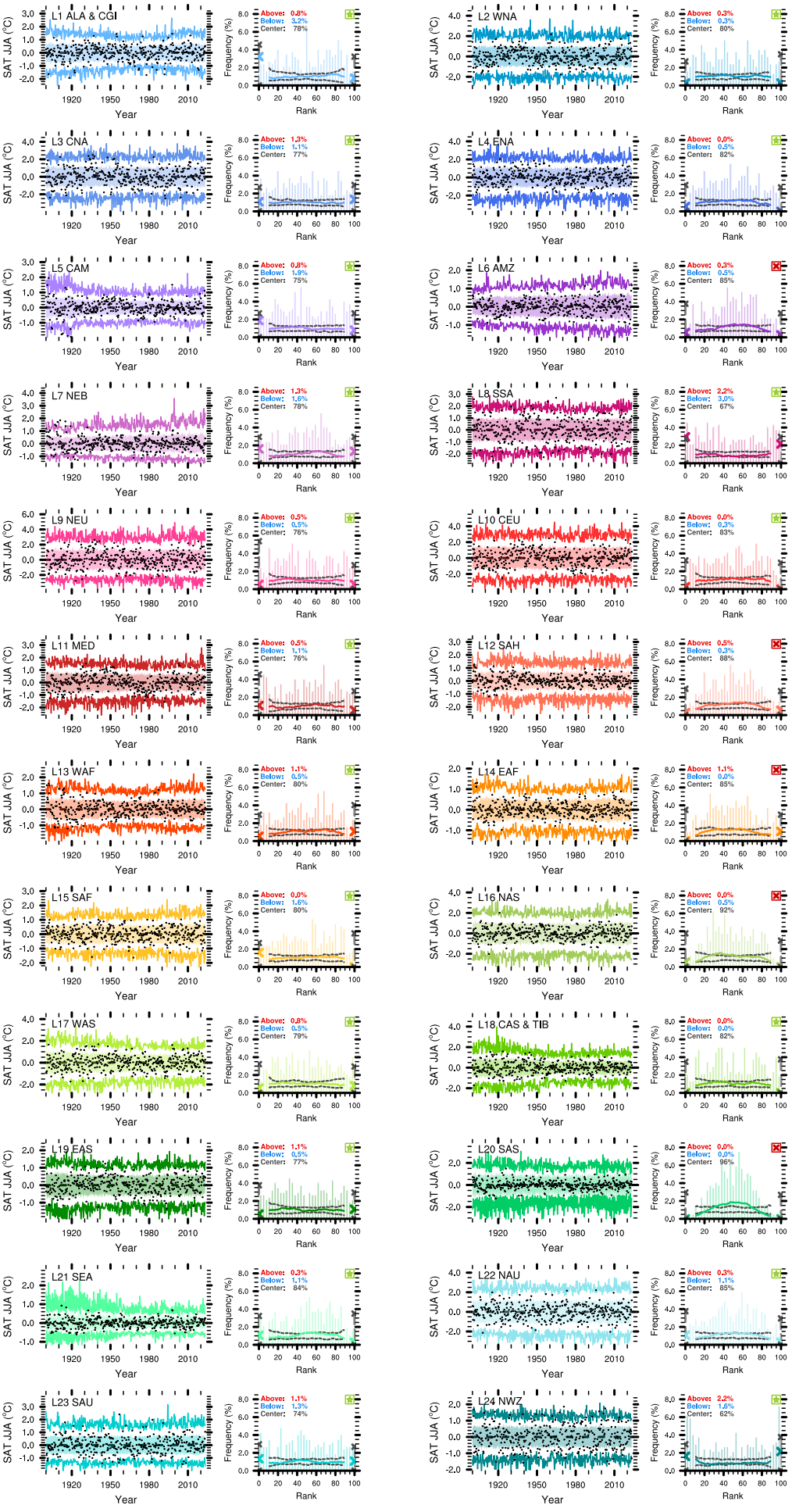






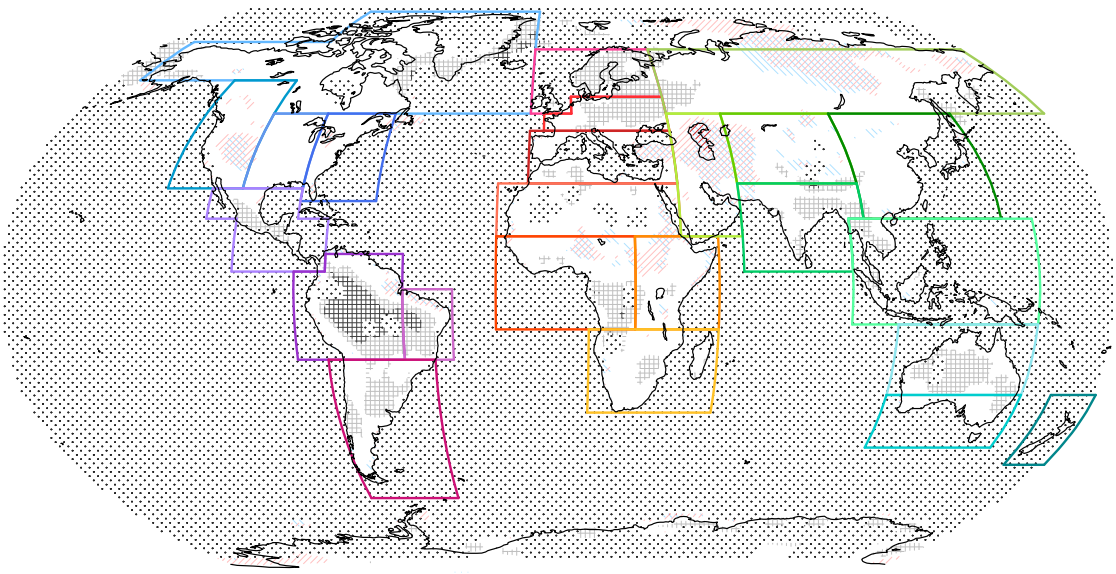


CESM2-LE vs GISTEMPv4 TAS JJA



CSIRO-Mk360 vs GISTEMPv4

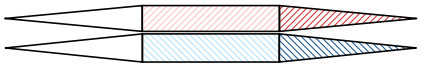
White Area = 62.0 %



TAS DJF Obs inside central 75th percentile (%)

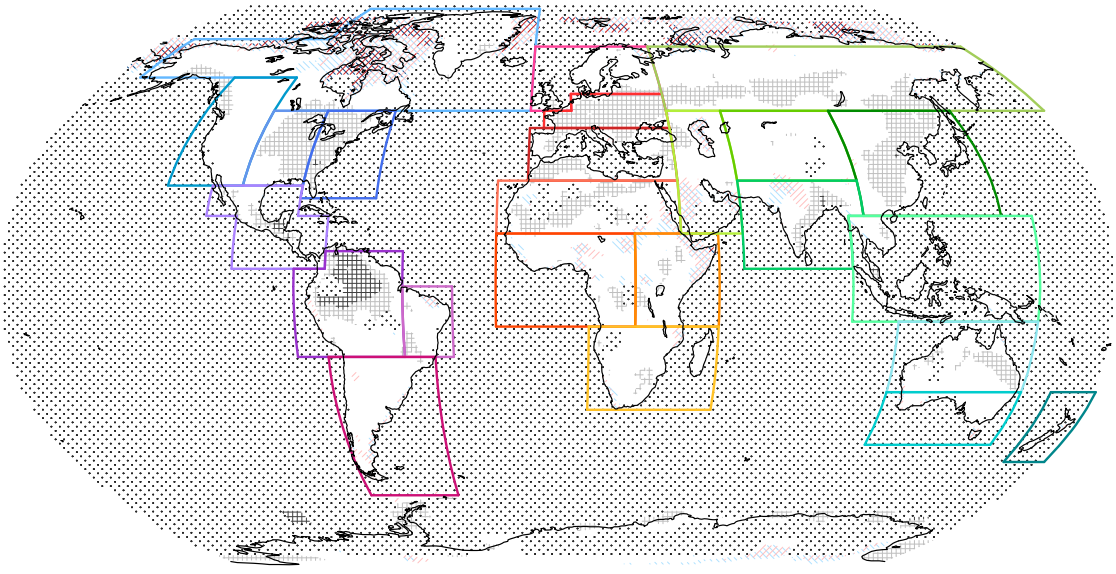


TAS DJF Obs outside ensemble spread (%)



CSIRO-Mk360 vs GISTEMPv4

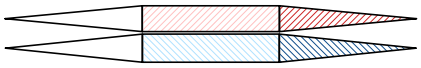
White Area = 60.2 %



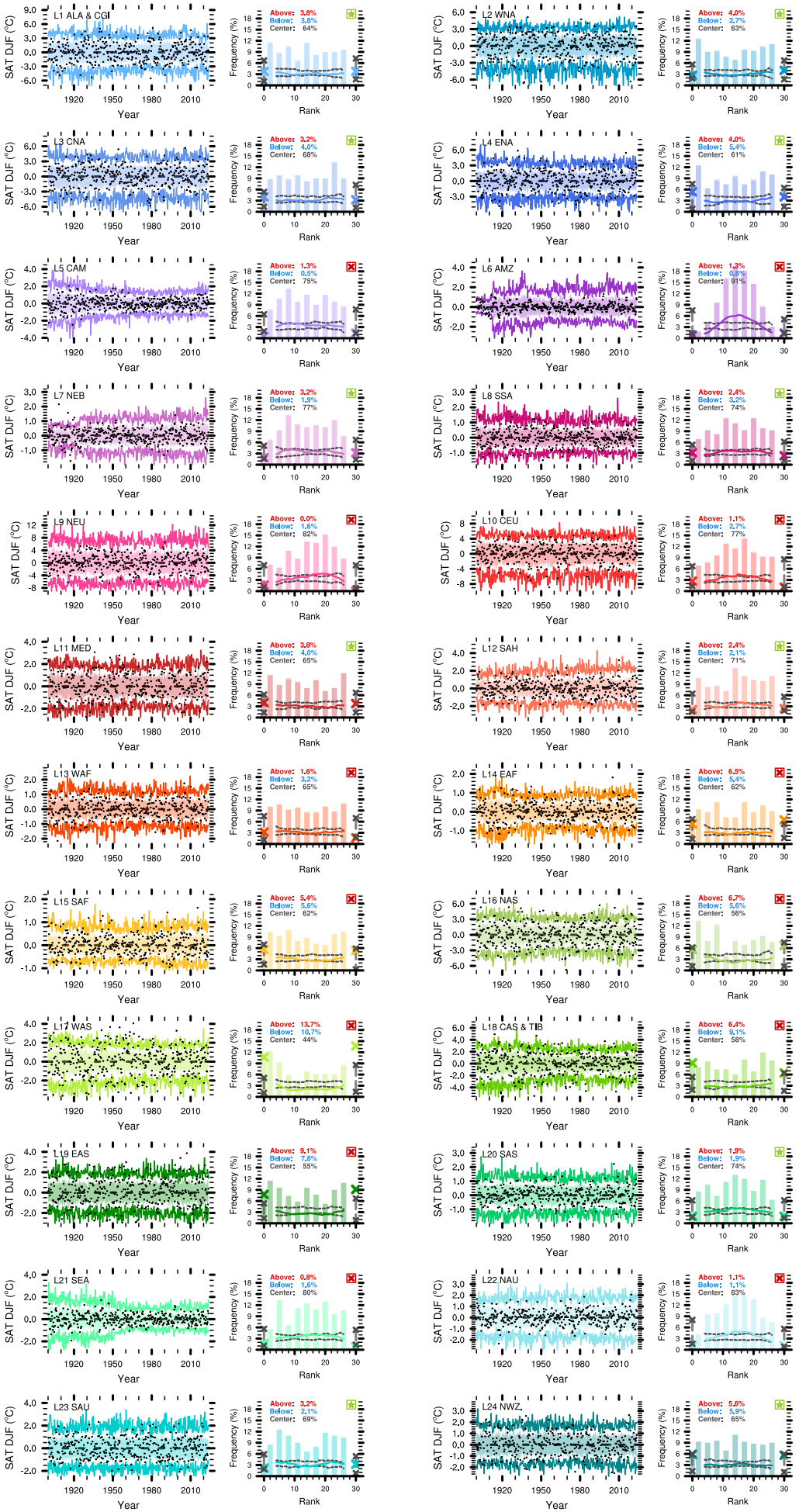
TAS JJA Obs inside central 75th percentile (%)



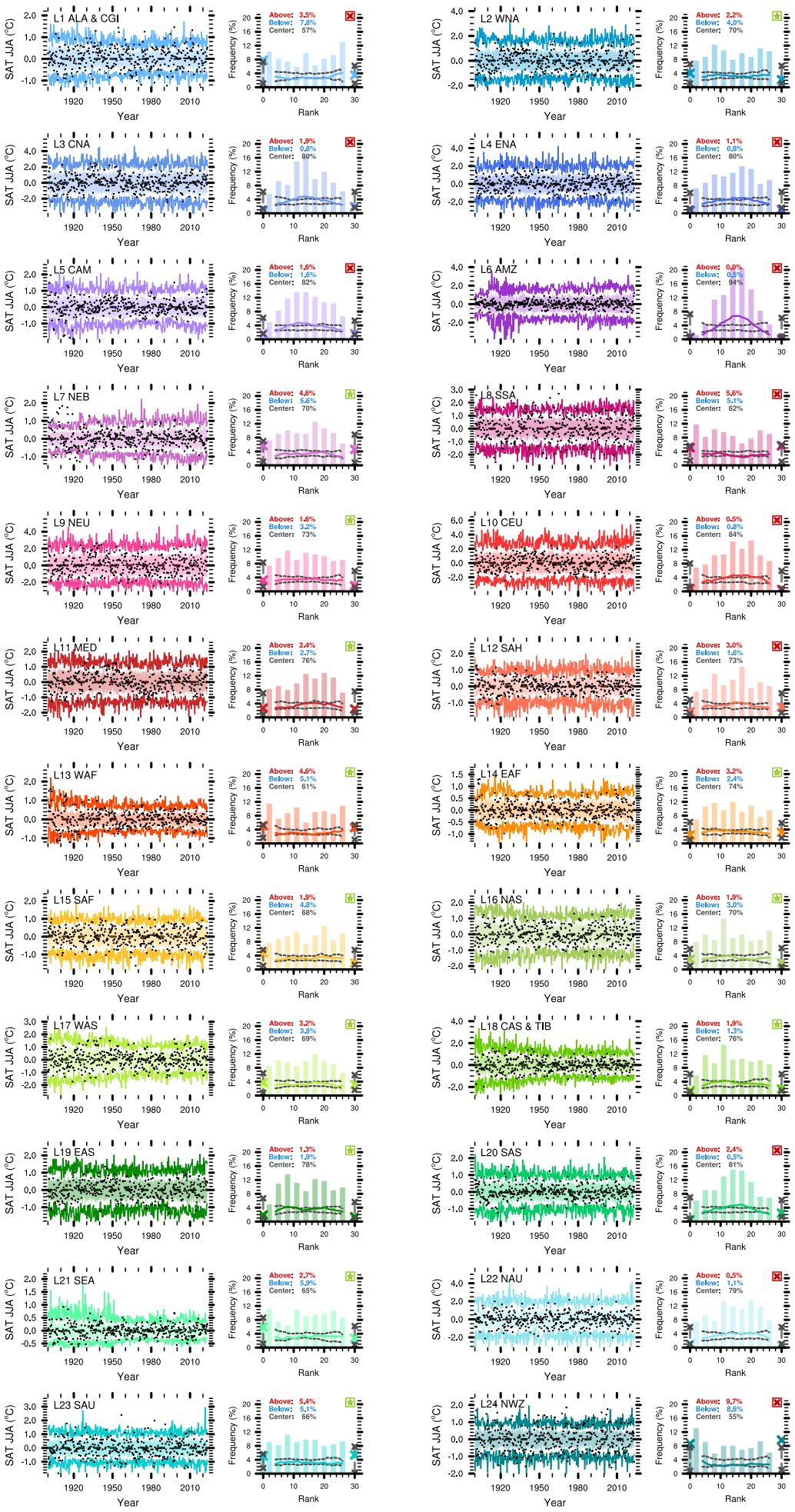
TAS JJA Obs outside ensemble spread (%)



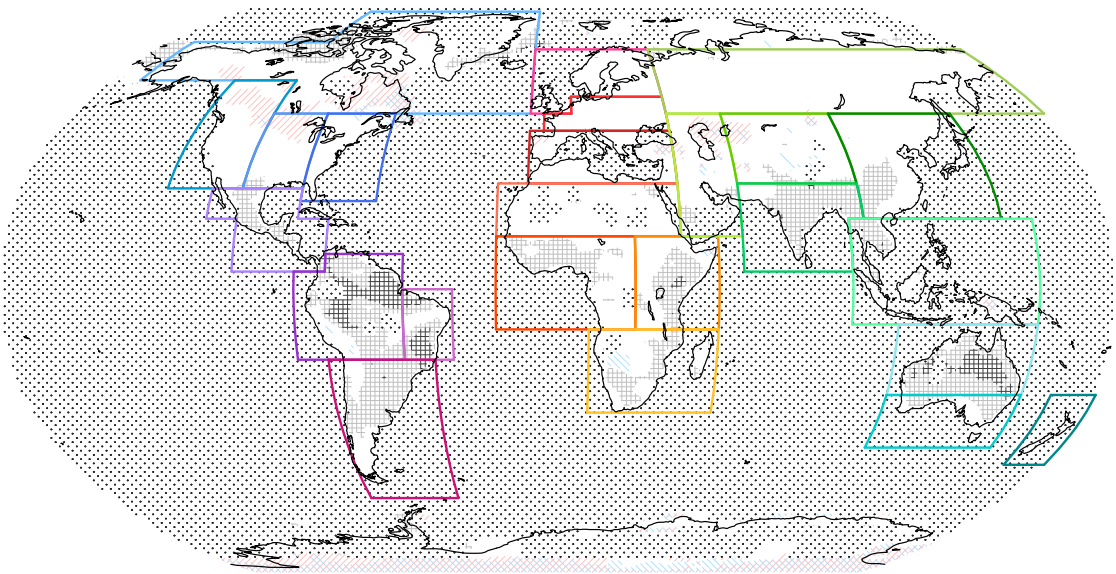
CSIRO-Mk360 vs GISTEMPv4 TAS DJF



CSIRO-Mk360 vs GISTEMPv4 TAS JJA



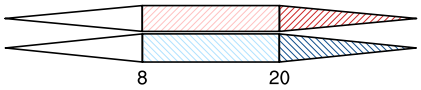
GFDL-ESM2M vs GISTEMPv4 White Area = 63.5 %



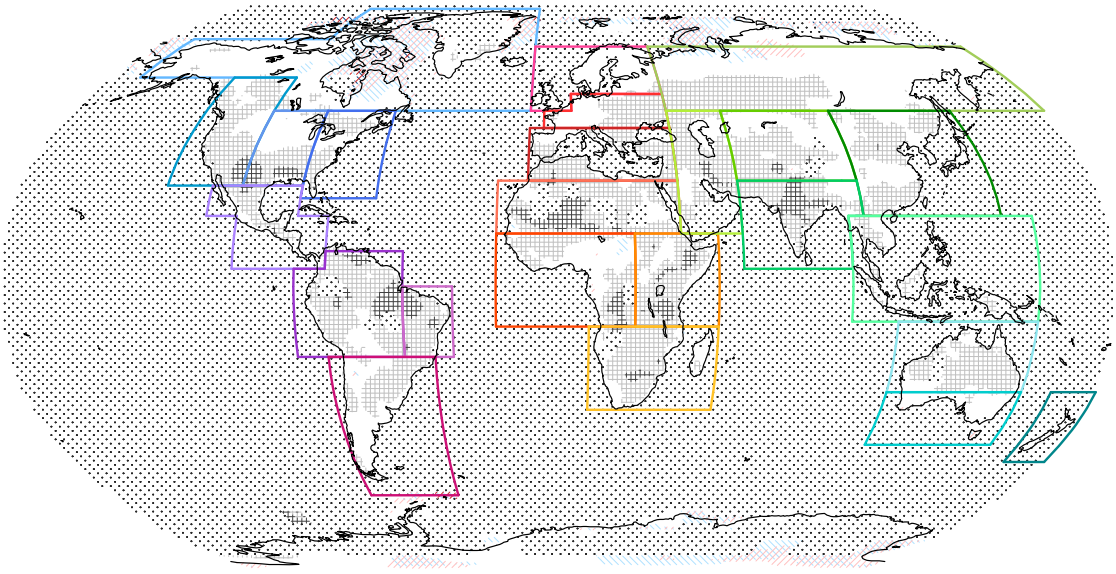
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



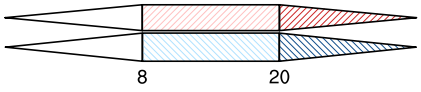
GFDL-ESM2M vs GISTEMPv4 White Area = 42.0 %



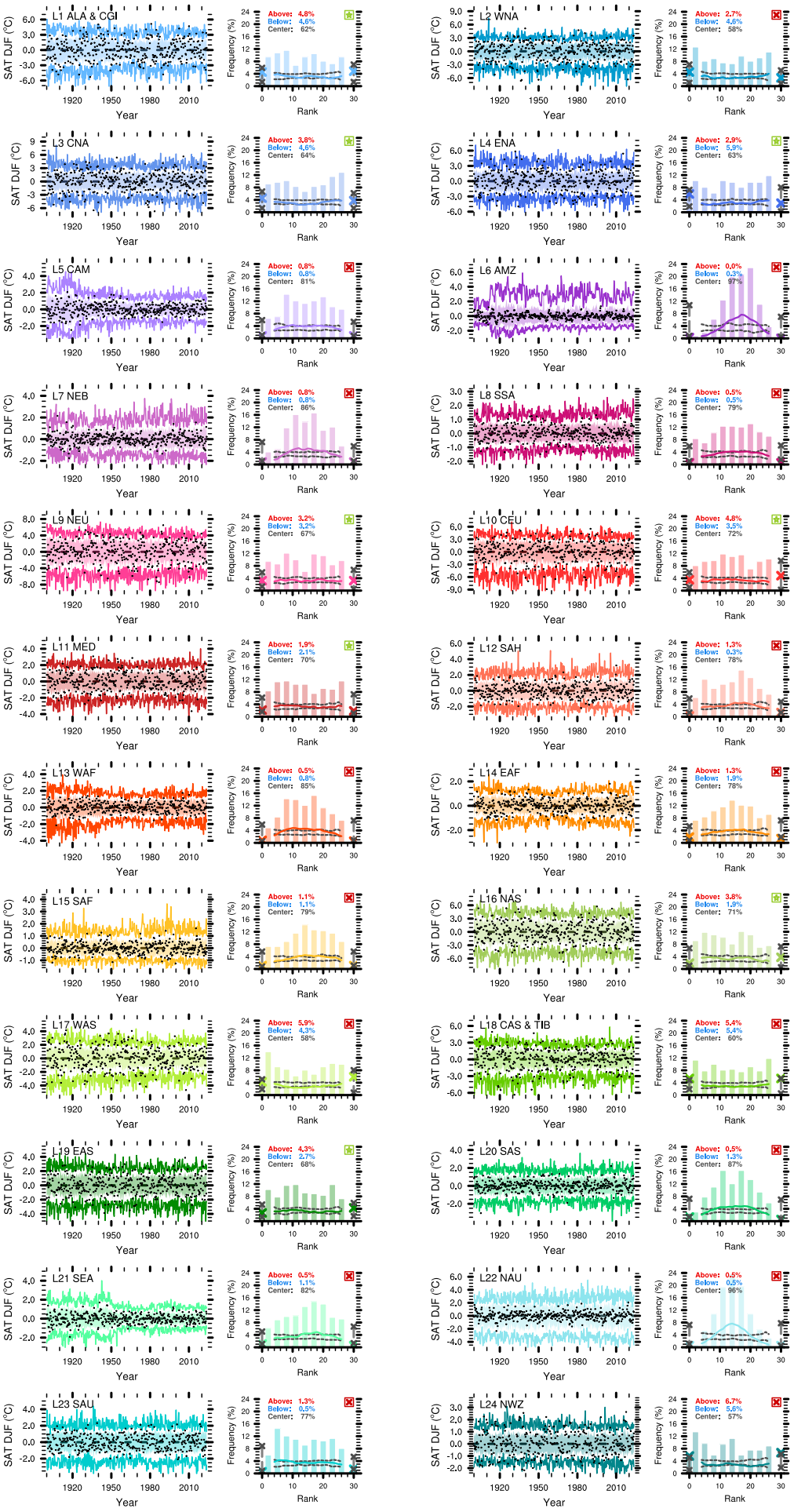
TAS JJA Obs inside central 75th percentile (%)



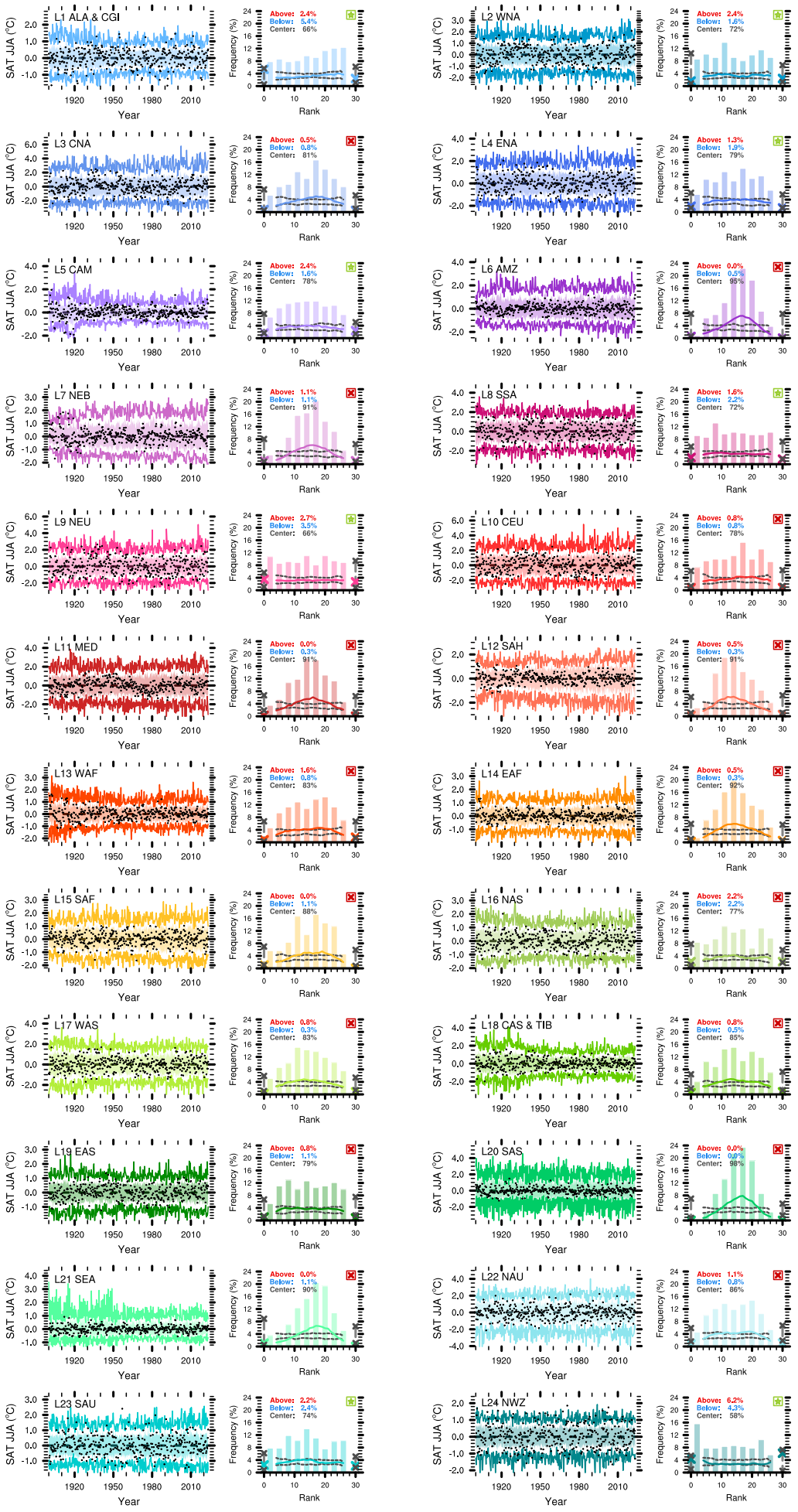
TAS JJA Obs outside ensemble spread (%)



GFDL-ESM2M vs GISTEMPv4 TAS DJF

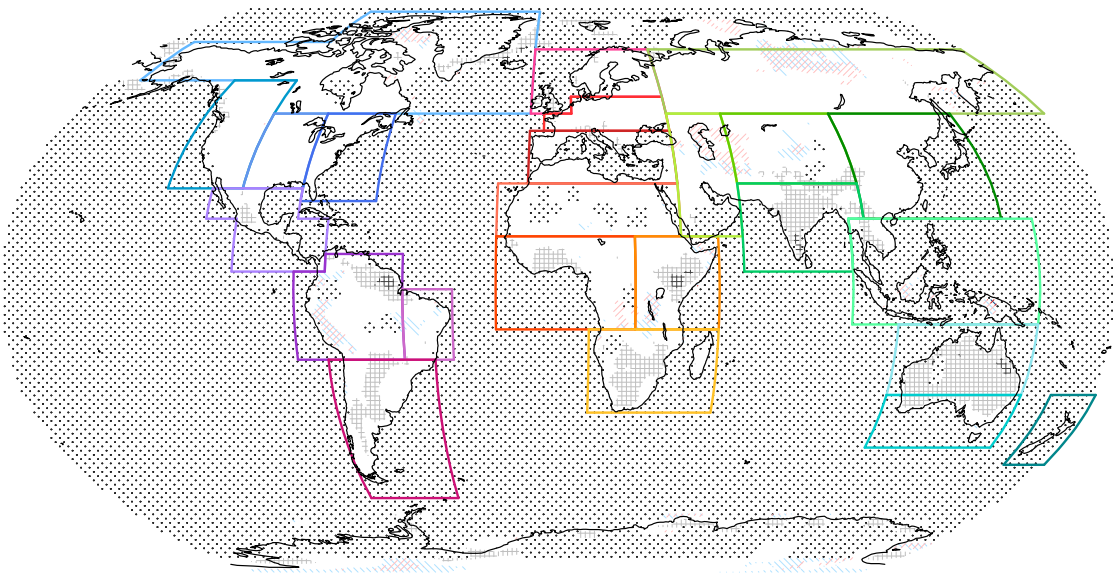


GFDL-ESM2M vs GISTEMPv4 TAS JJA



GFDL-SPEAR vs GISTEMPv4

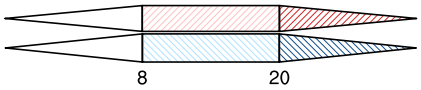
White Area = 74.3 %



TAS DJF Obs inside central 75th percentile (%)

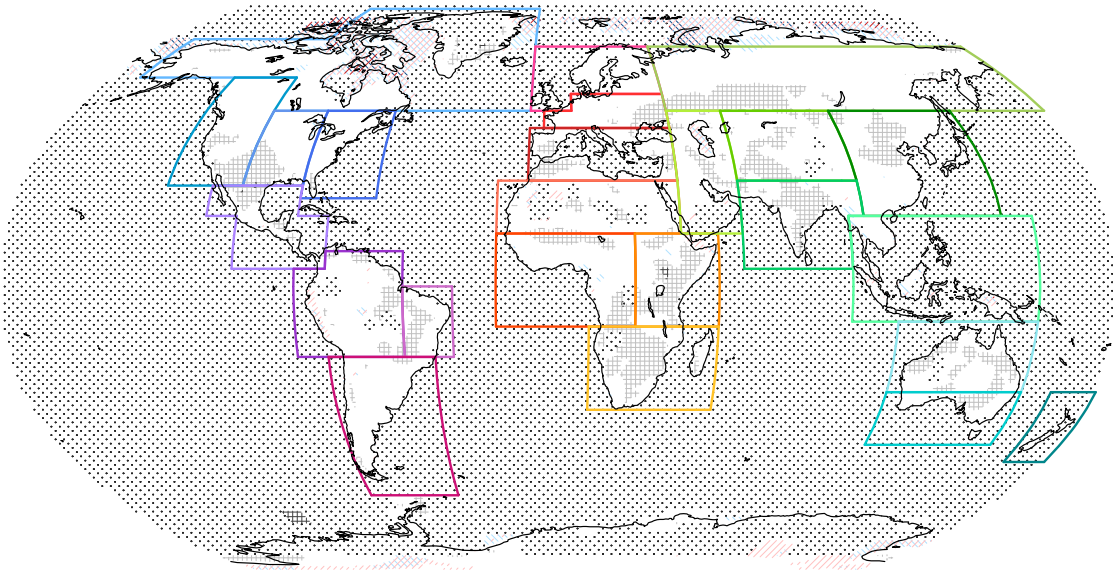


TAS DJF Obs outside ensemble spread (%)



GFDL-SPEAR vs GISTEMPv4

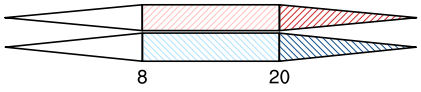
White Area = 68.1 %



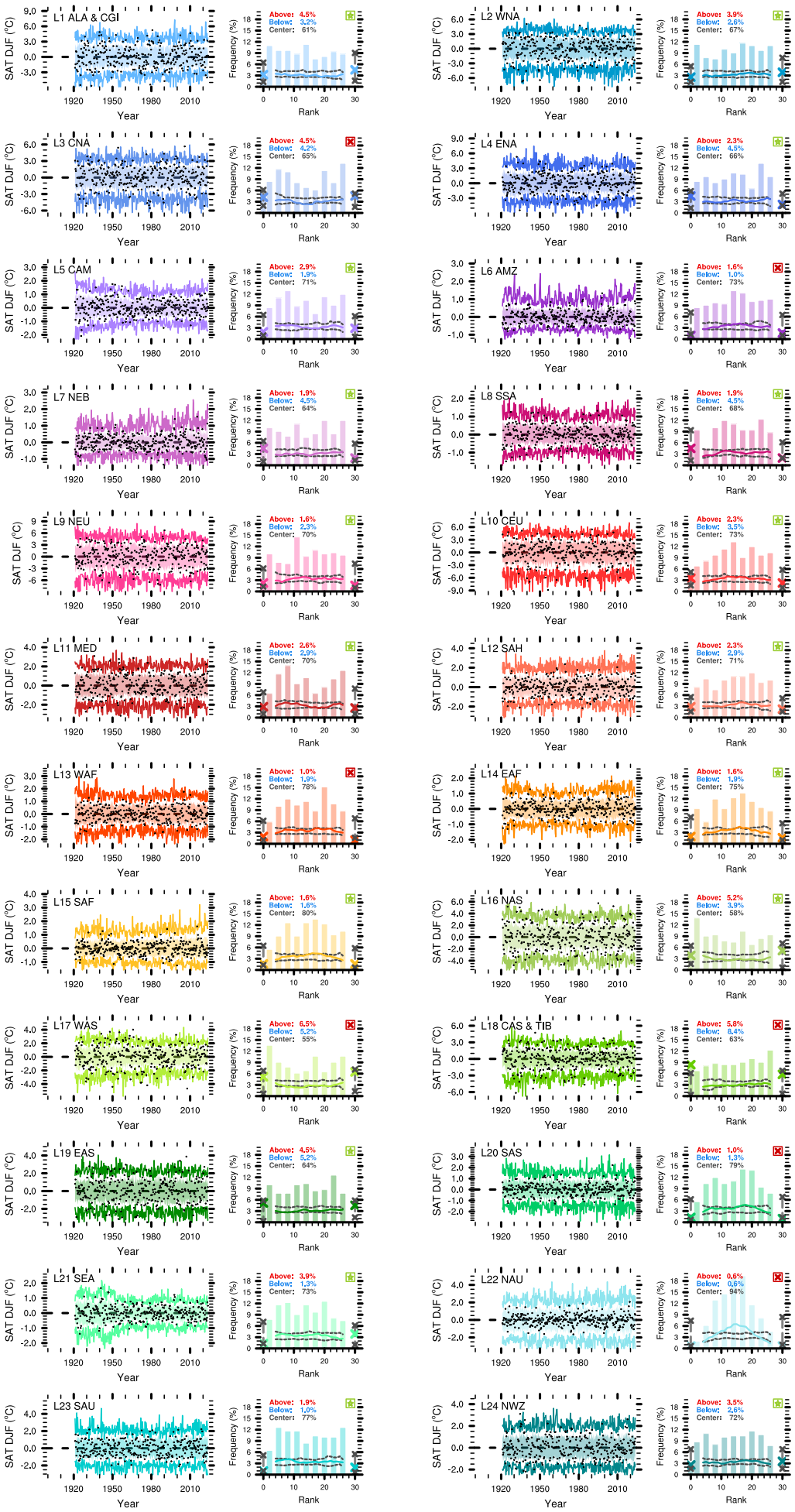
TAS JJA Obs inside central 75th percentile (%)



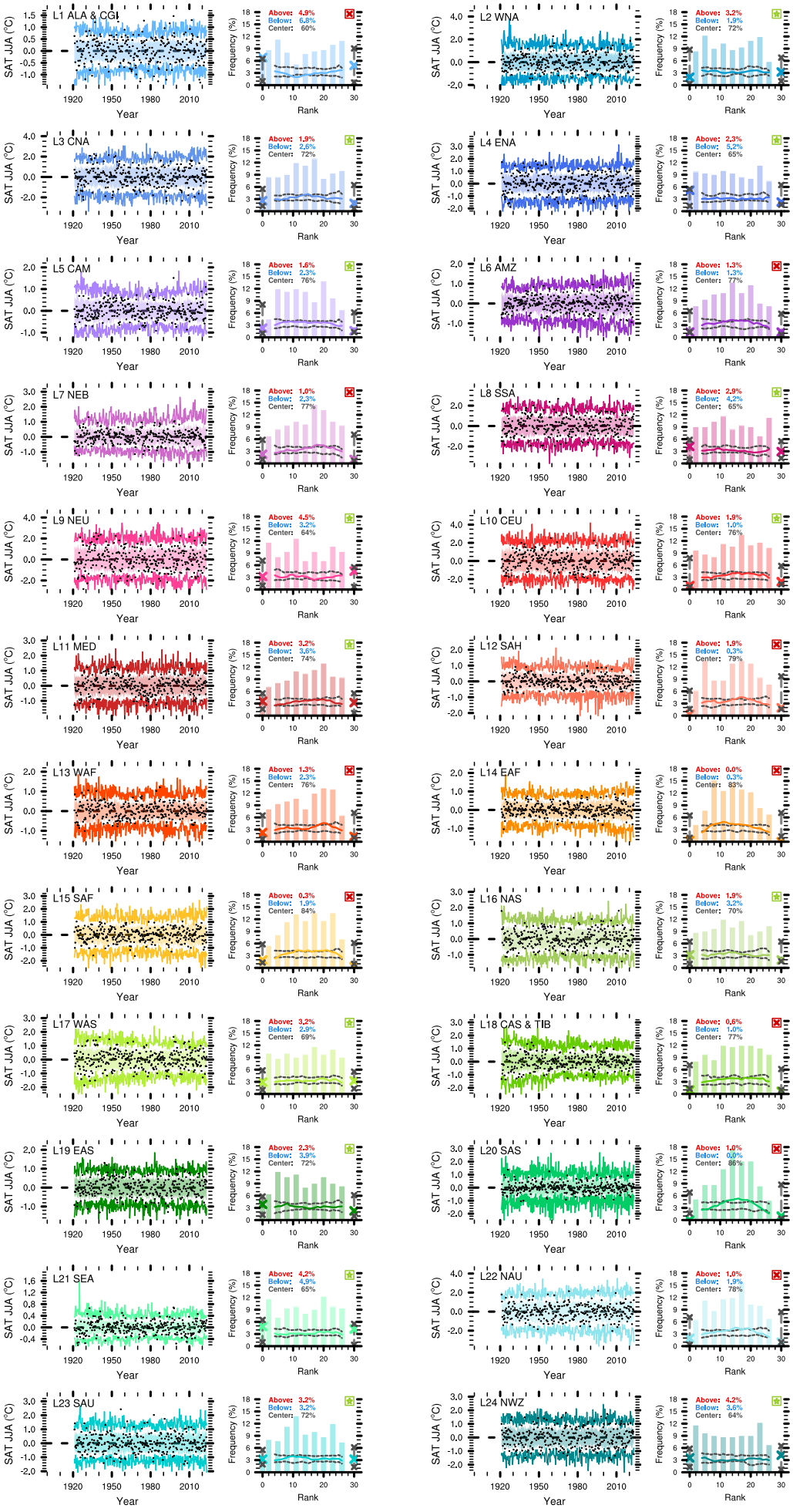
TAS JJA Obs outside ensemble spread (%)



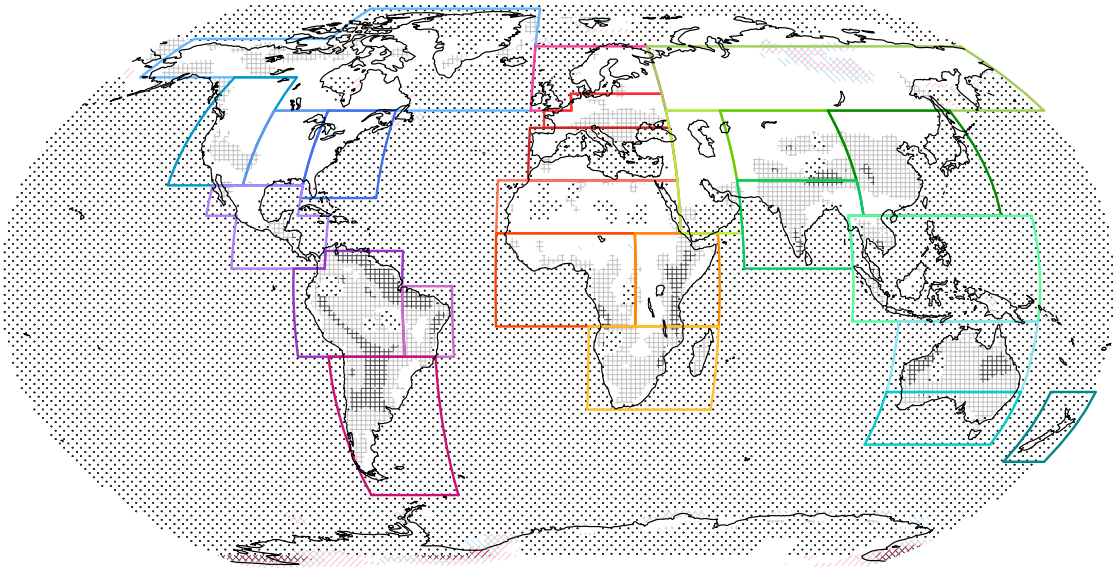
GFDL-SPEAR vs GISTEMPv4 TAS DJF



GFDL-SPEAR vs GISTEMPv4 TAS JJA



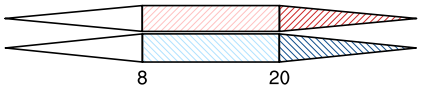
MIROC6 vs GISTEMPv4 White Area = 58.6 %



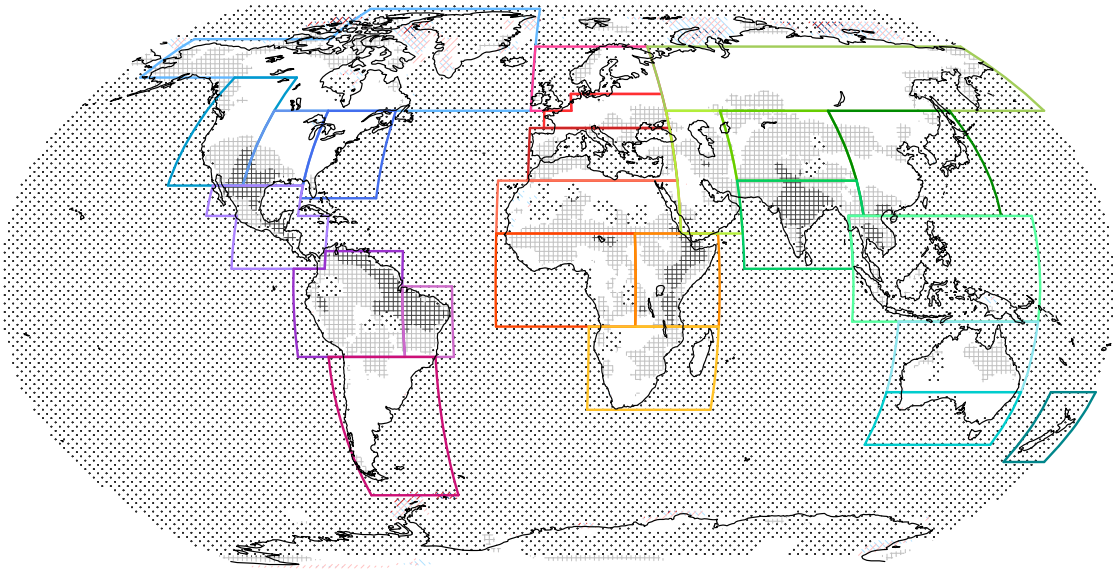
TAS DJF Obs inside central 75th percentile (%)



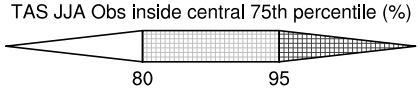
TAS DJF Obs outside ensemble spread (%)



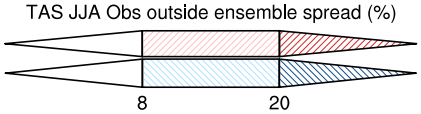
MIROC6 vs GISTEMPv4 White Area = 52.5 %



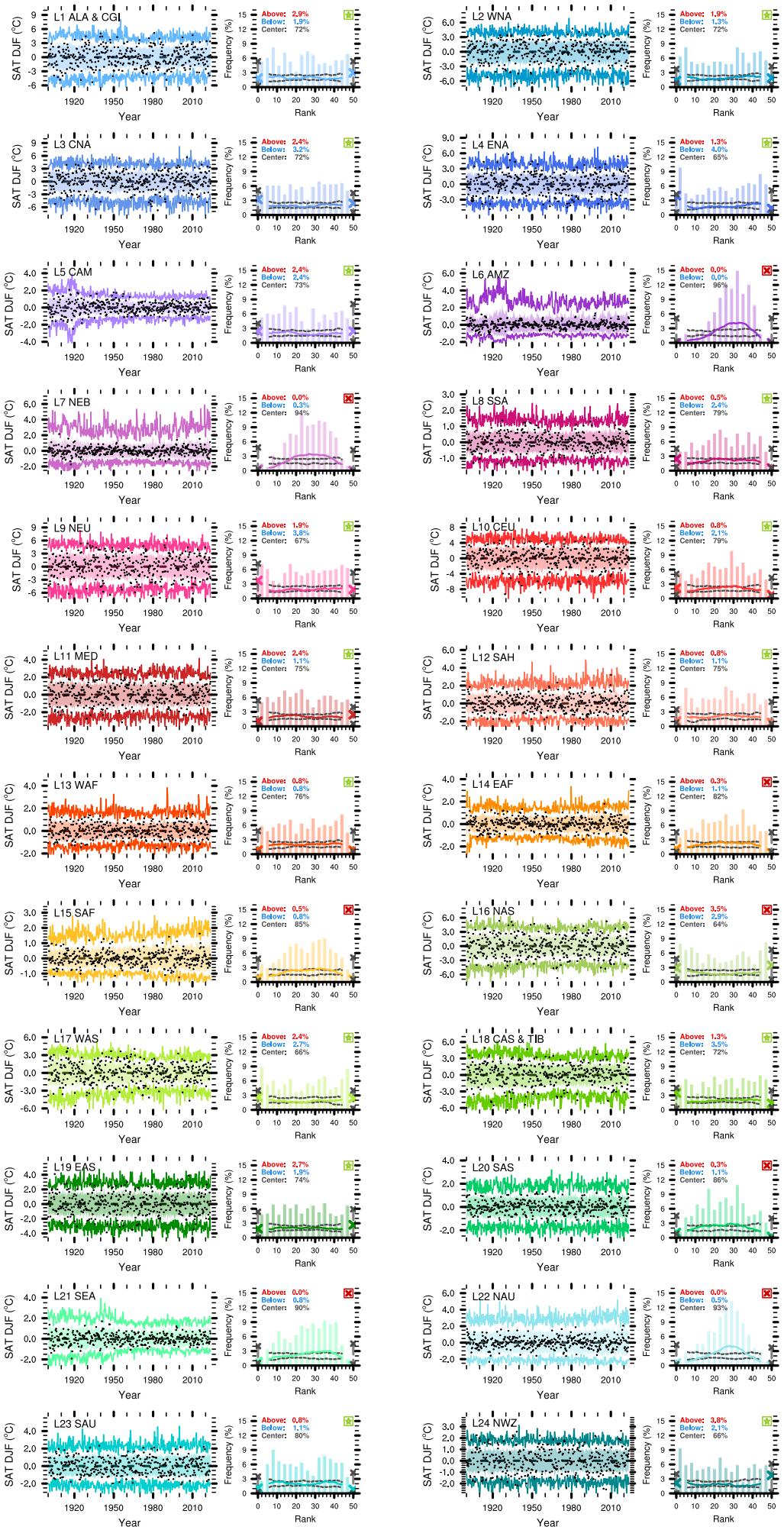
TAS JJA Obs inside central 75th percentile (%)



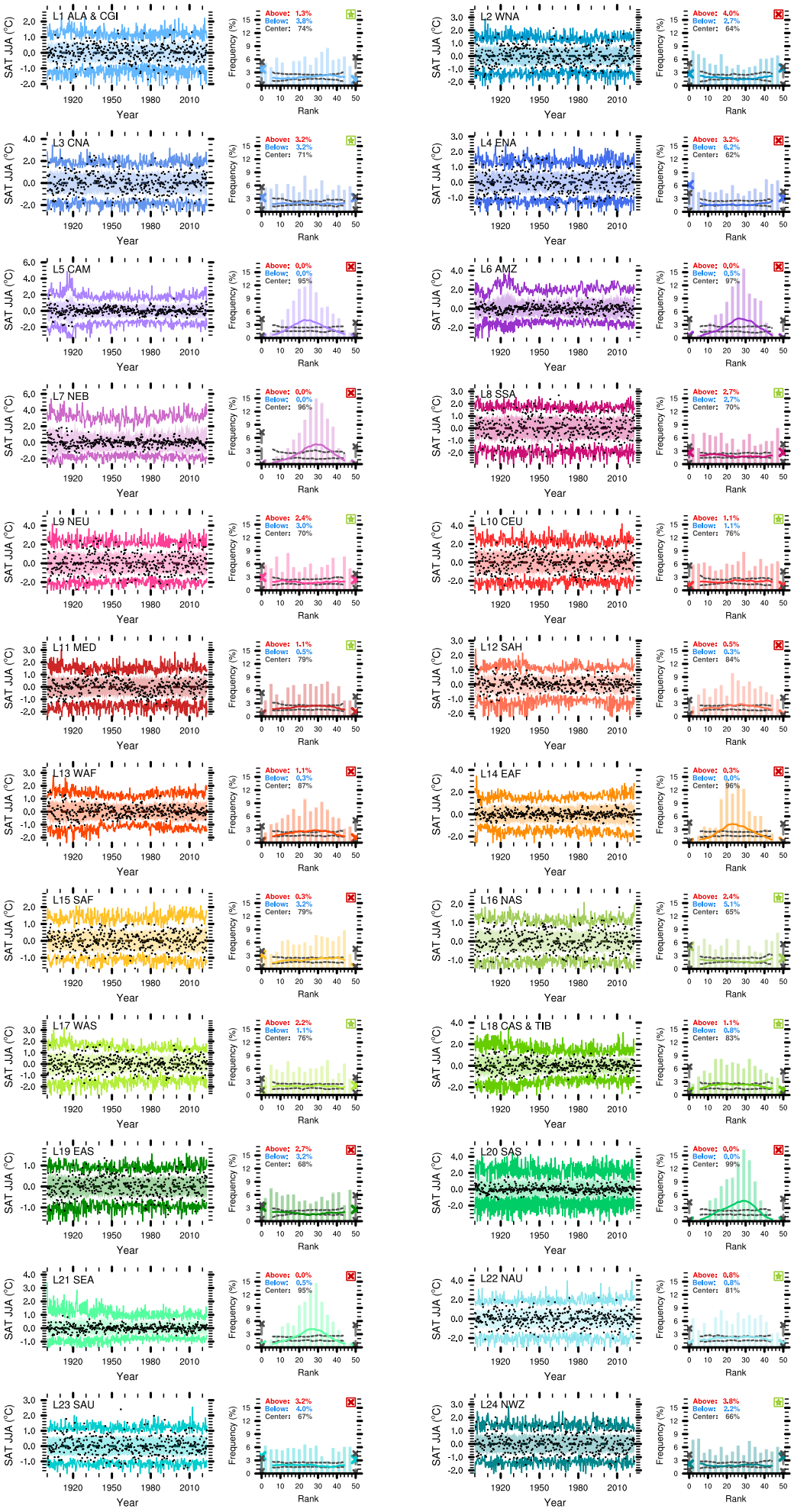
TAS JJA Obs outside ensemble spread (%)



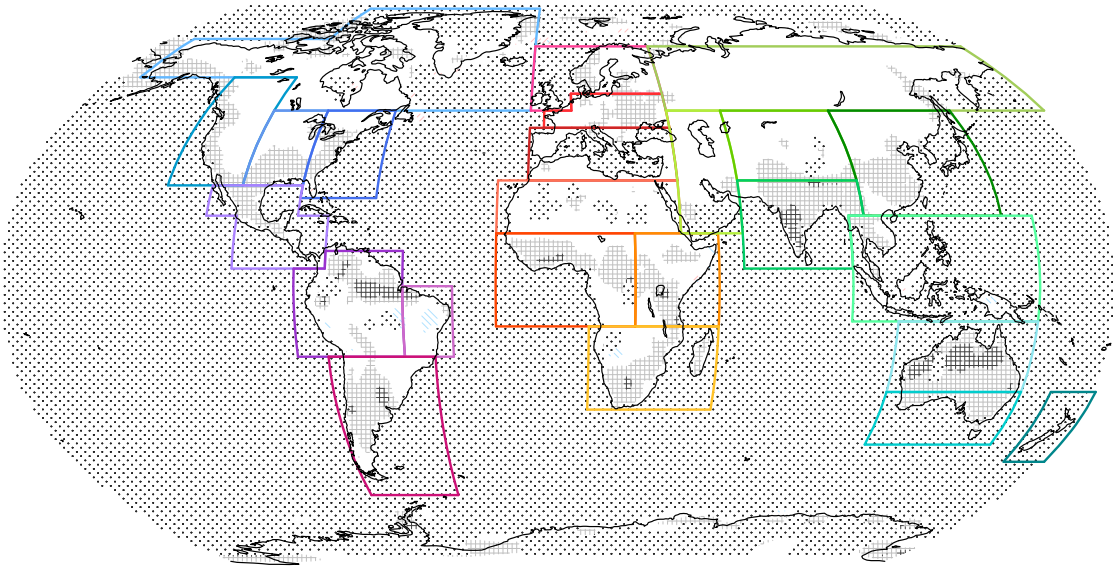
MIROC6 vs GISTEMPv4 TAS DJF



MIROC6 vs GISTEMPv4 TAS JJA



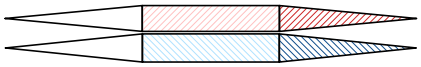
MPI-GE5 vs GISTEMPv4 White Area = 65.6 %



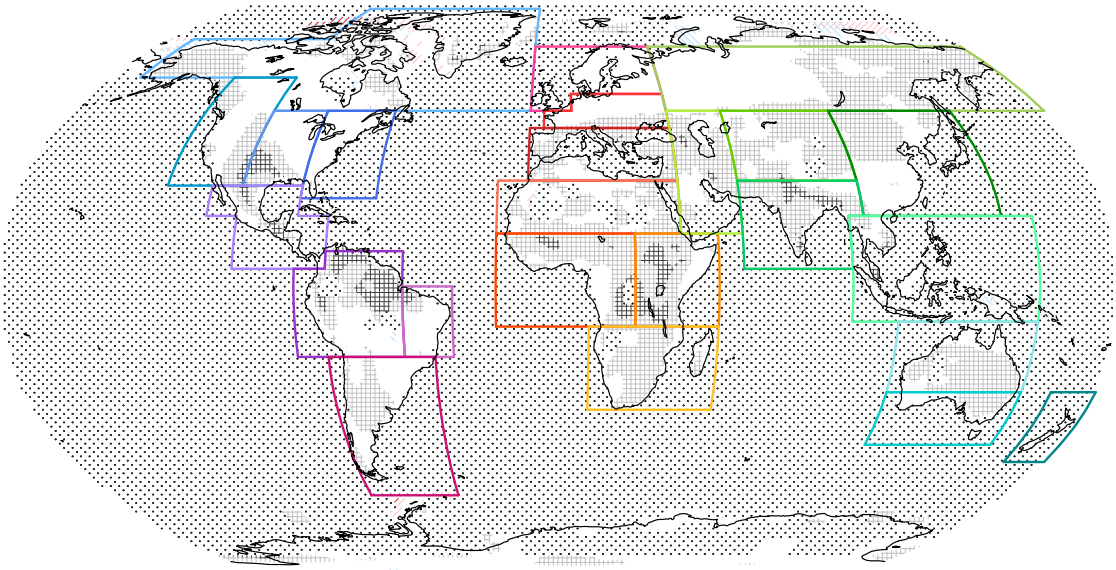
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



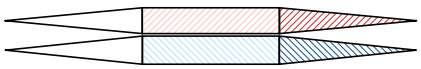
MPI-GE5 vs GISTEMPv4 White Area = 51.0 %



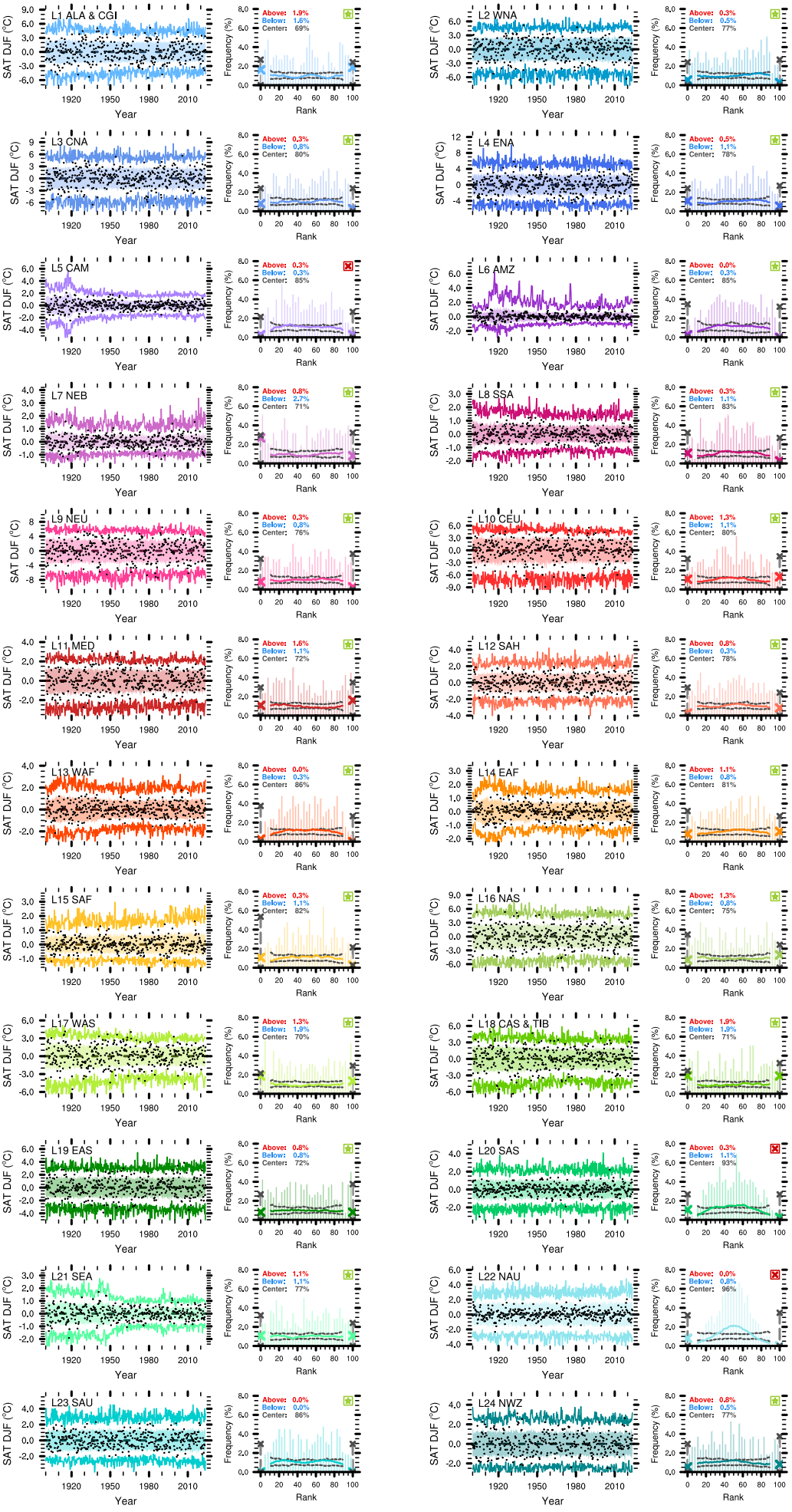
TAS JJA Obs inside central 75th percentile (%)



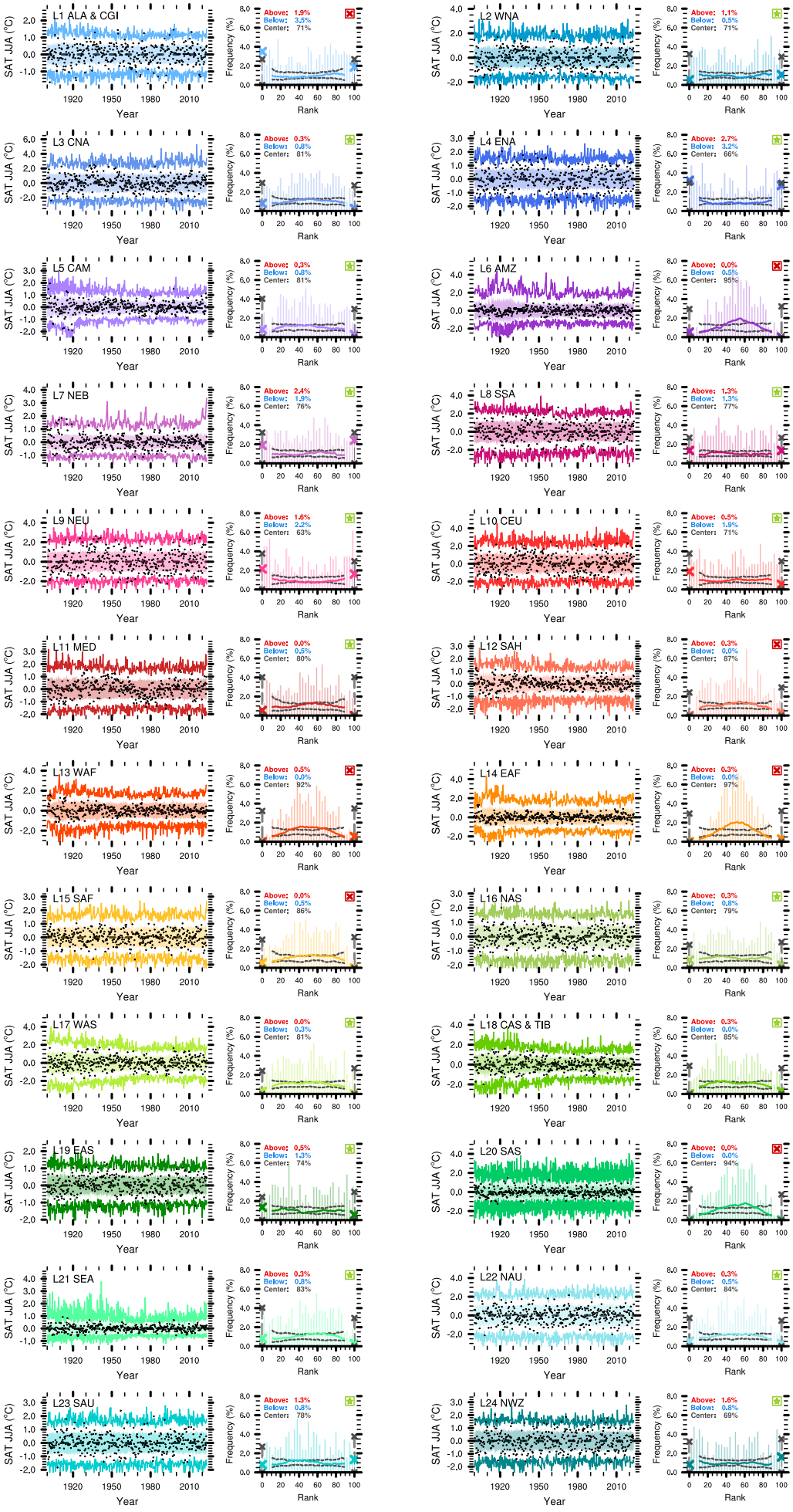
TAS JJA Obs outside ensemble spread (%)



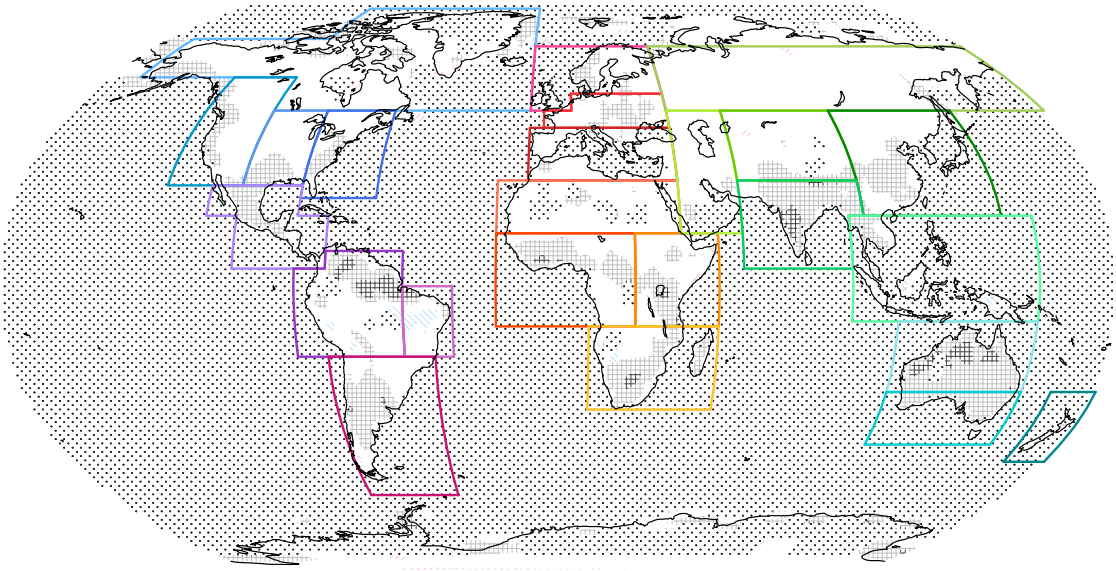
MPI-GE5 vs GISTEMPv4 TAS DJF



MPI-GE5 vs GISTEMPv4 TAS JJA



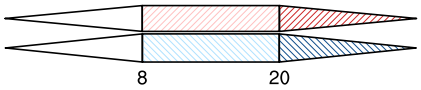
MPI-GE6 vs GISTEMPv4 White Area = 68.0 %



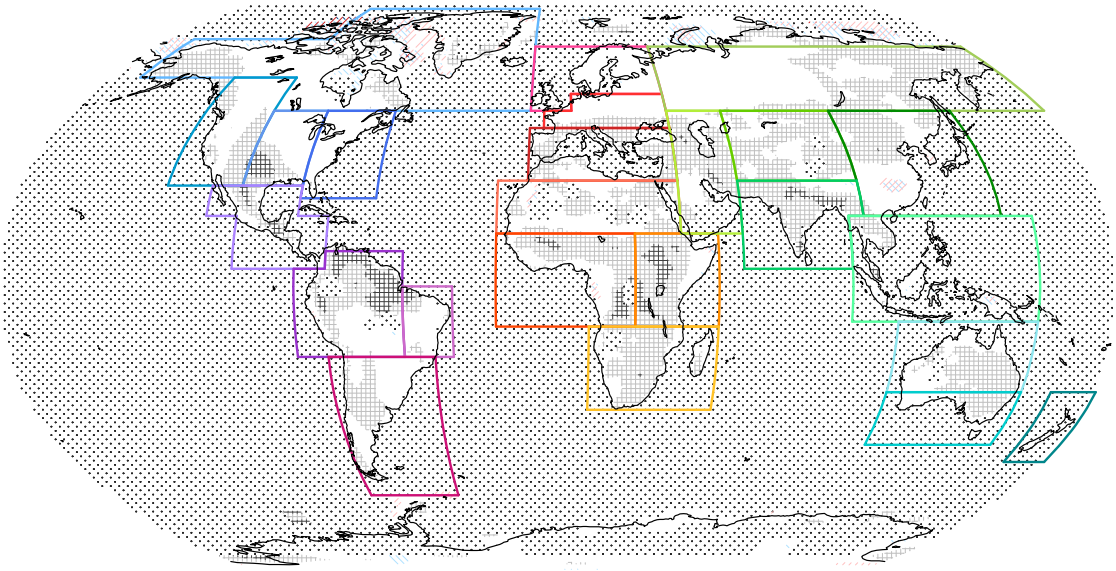
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



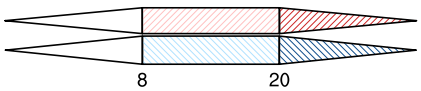
MPI-GE6 vs GISTEMPv4 White Area = 50.7 %



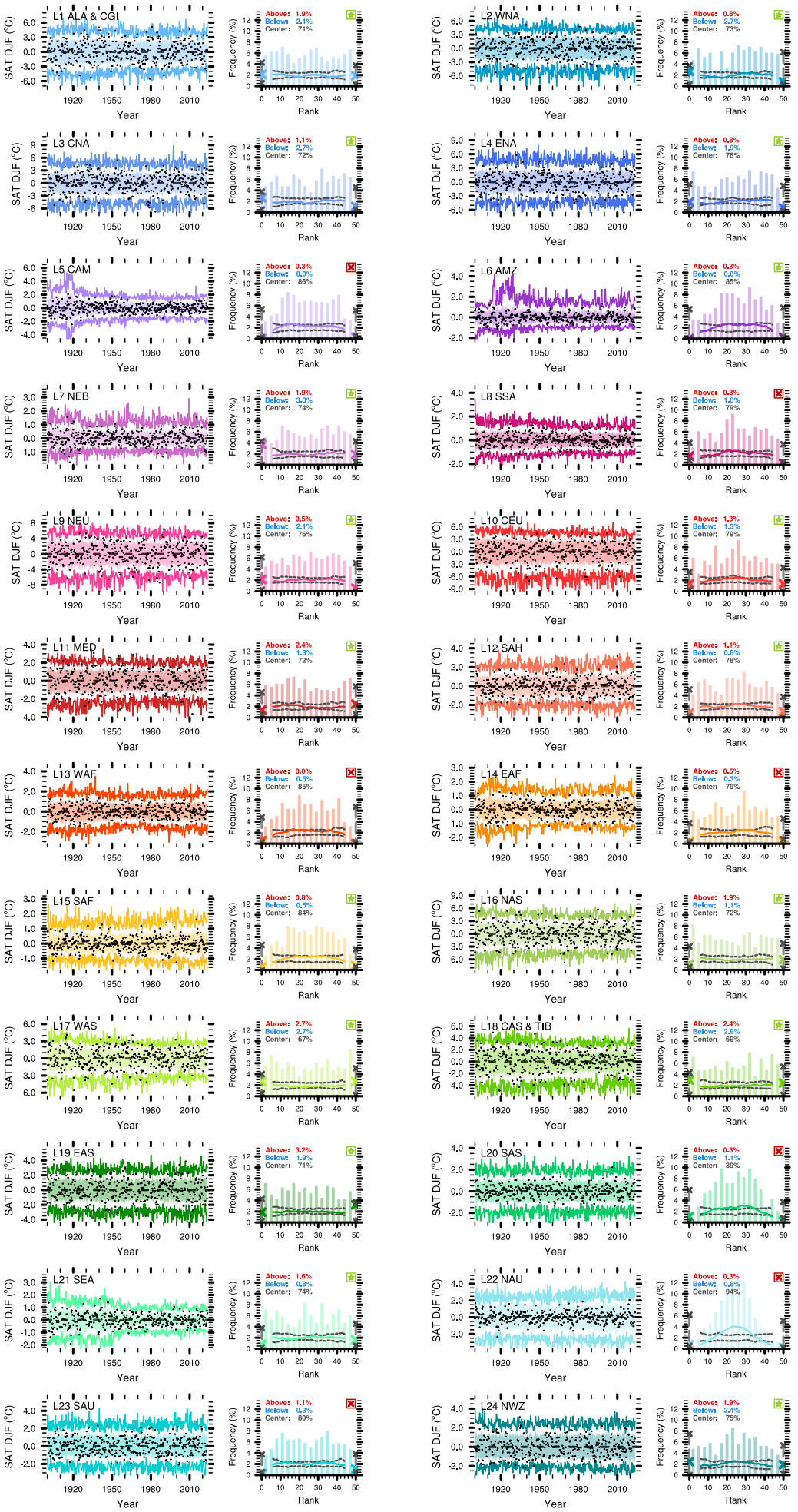
TAS JJA Obs inside central 75th percentile (%)



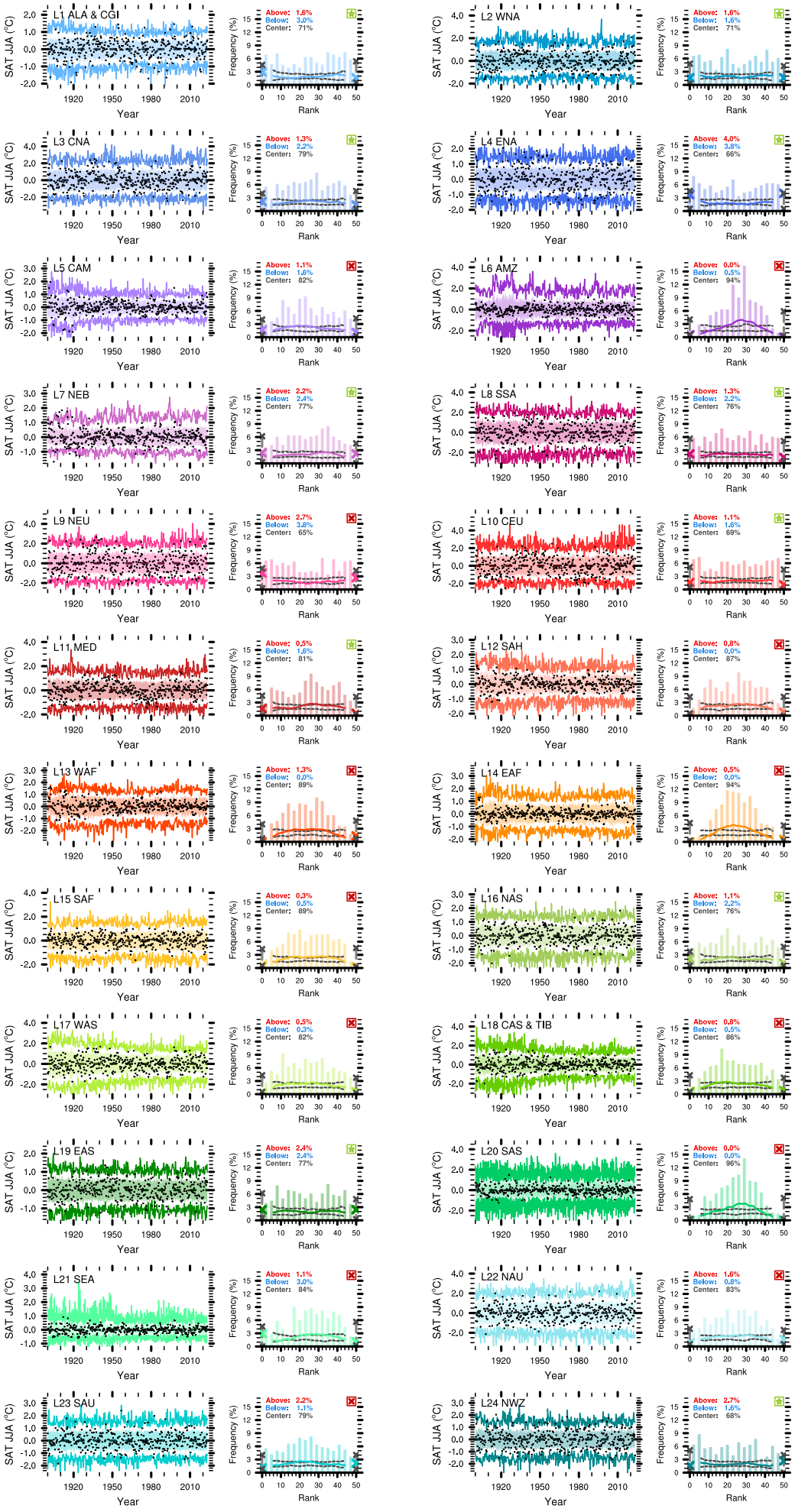
TAS JJA Obs outside ensemble spread (%)



MPI-GE6 vs GISTEMPv4 TAS DJF



MPI-GE6 vs GISTEMPv4 TAS JJA



Non-Detrended Land Surface Air Temperatures

Rank-frequency variability evaluation framework for non-detrended 2m air temperature (TAS) anomalies over land grid cells.

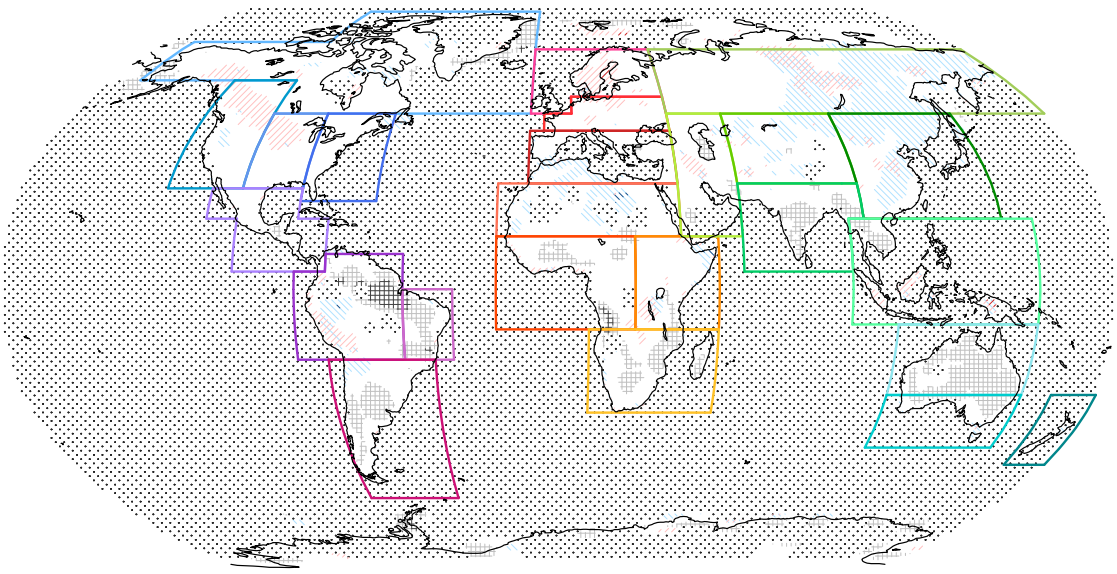
Maps show grid-cell evaluation of the simulated DJF and JJA monthly mean temperature anomalies for the 11 SMILEs included against GISTEMPv4 observations globally. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensemble (12.5th to 87.5th percentiles) for more than 80% of months (light grey) or for more than 95% of months (dark grey). Red and blue shading represents where observations are larger than the ensemble maximum (red) or smaller than the ensemble minimum (blue), respectively, for more than 8% of the months (light red and blue) or for more than 20% of the months (dark red and blue). Dotted areas represent ocean areas or grid cells where observations are missing and are therefore excluded from this analysis. Colored boxes demark the boundaries of each land region assessed. The percentage of assessed grid-cells that present none of these biases, unbiased grid-cells, in given at the top (white area). For each assessed region, models are considered to provide an adequate grid-cell representation when they exhibit more than 50% of unbiased grid cells in the region, fulfilling Criteria 2 of our evaluation framework.

Time series and rank frequency histograms show spatially aggregated DJF and JJA TAS for each land region for all 11 SMILEs. Time series show the ensemble maximum and minimum (color lines) and central 75th percentile ensemble spread (shading) are shown against observations (black dots).

Rank histograms represent the frequency of each place that observations would take in a list of ensemble members ordered by ascending temperature anomaly values. Rank histograms show the observations rank frequency accumulated for 3-rank bins (bars), the running mean rank frequency over a centered n/5 rank window (lines; for 1 to n-1 ranks) and the absolute frequencies of rank 0 and n (crosses), with n the number of ensemble members, for observations (color) and perfect model rank range (gray).

To illustrate how internal variability sampling may affect rank frequencies given the non-infinite record length considered, we include a comparison to the perfect-model rank range, which shows the range of rank frequencies that each ensemble member would take if it were observations (see Methods in main article for further details). If the rank exhibited by observations (colors) is within this perfect-model range (gray; allowing a maximum 10% deviation outside of this range), the rank frequency evaluation shows an adequate model performance, and any deviations from a perfectly frank rank histogram can be assumed to be within the range of deviations that could be caused by internal variability. In the case the observations rank frequency is within the perfect-model range for all rank windows, Criteria 1 of our evaluation framework is met and this is highlighted by a green star at the top right; if not, by a red cross. Percentages at the top left show the frequency of regionally averaged observations occurring above (red) or below (blue) ensemble limits, or clustering within the central 75th percentile range (gray), analogous to the grid-cell evaluation in Criteria 2.

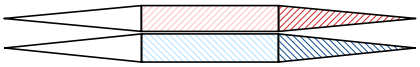
ACCESS vs GISTEMPv4 White Area = 61.3 %



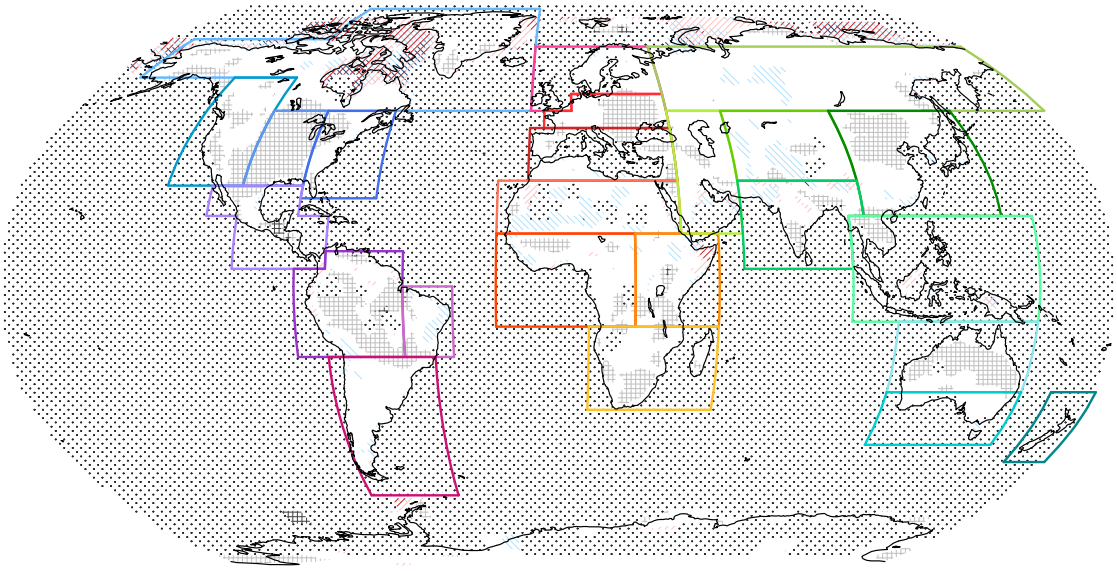
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



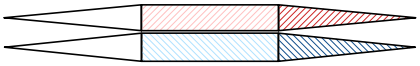
ACCESS vs GISTEMPv4 White Area = 59.0 %



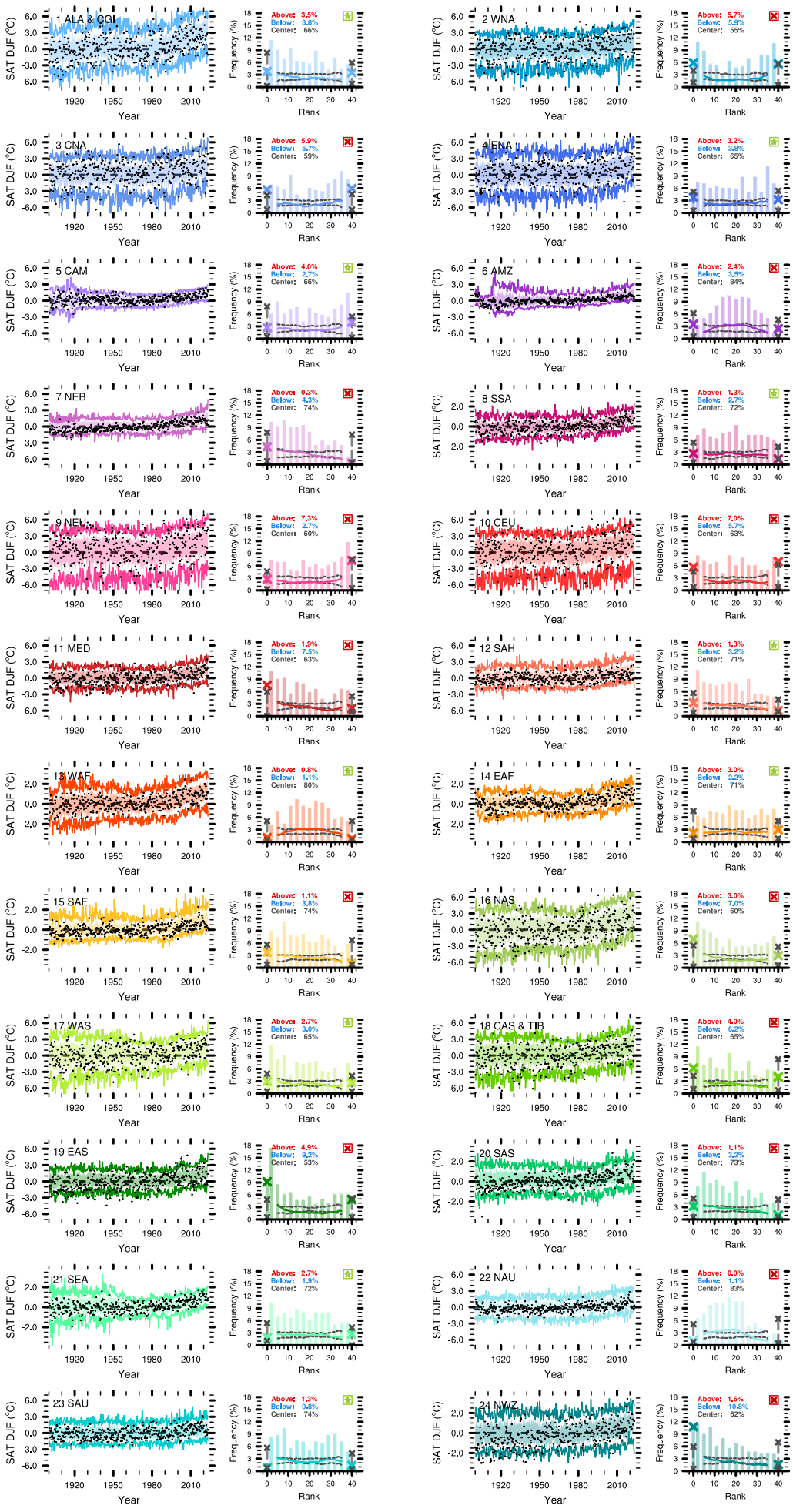
TAS JJA Obs inside central 75th percentile (%)



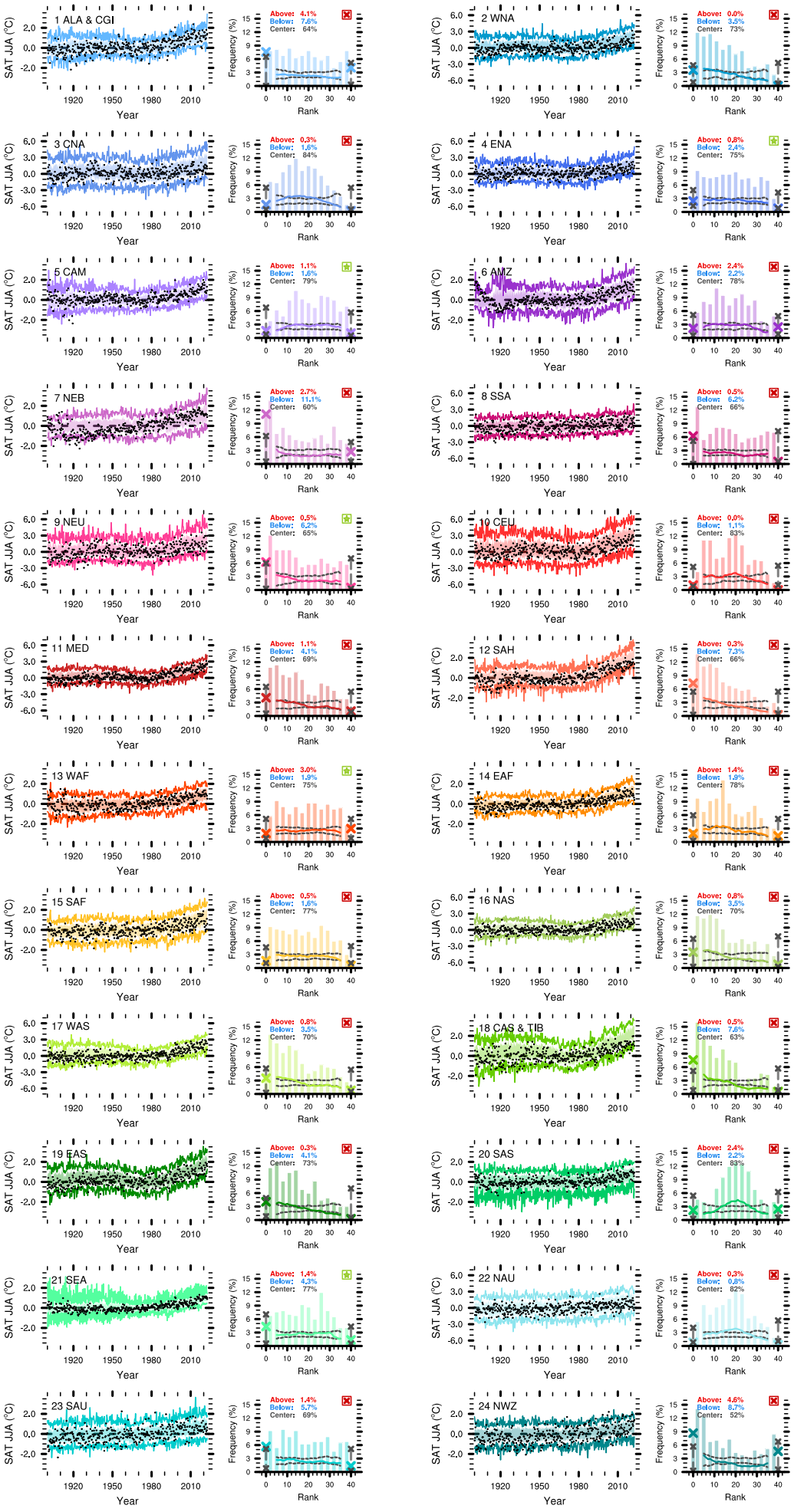
TAS JJA Obs outside ensemble spread (%)



ACCESS vs GISTEMPv4 TAS DJF

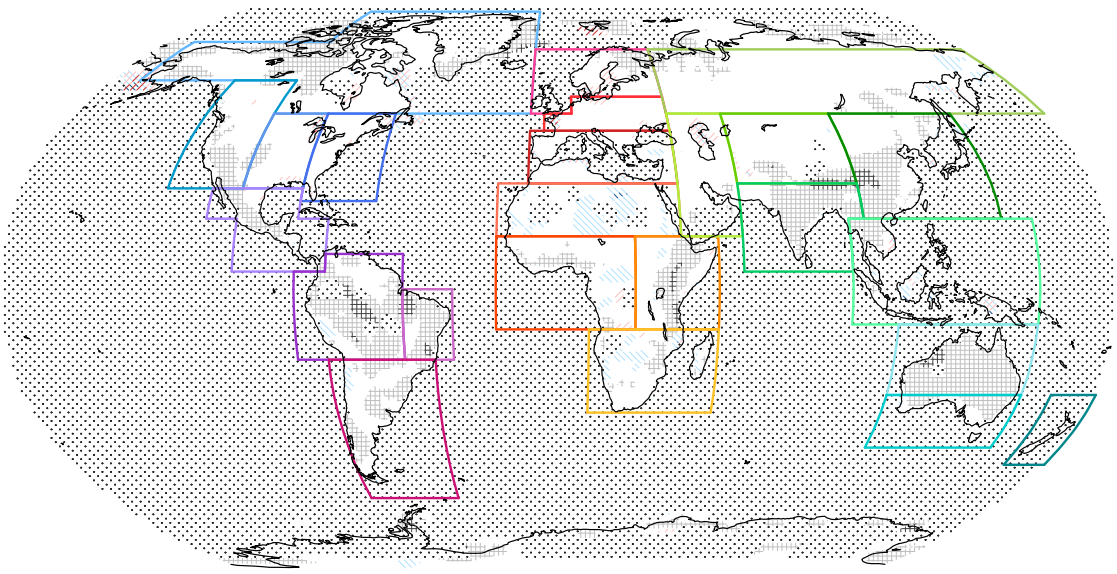


ACCESS vs GISTEMPv4 TAS JJA



CanESM2 vs GISTEMPv4

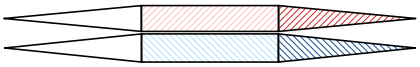
White Area = 57.7 %



TAS DJF Obs inside central 75th percentile (%)

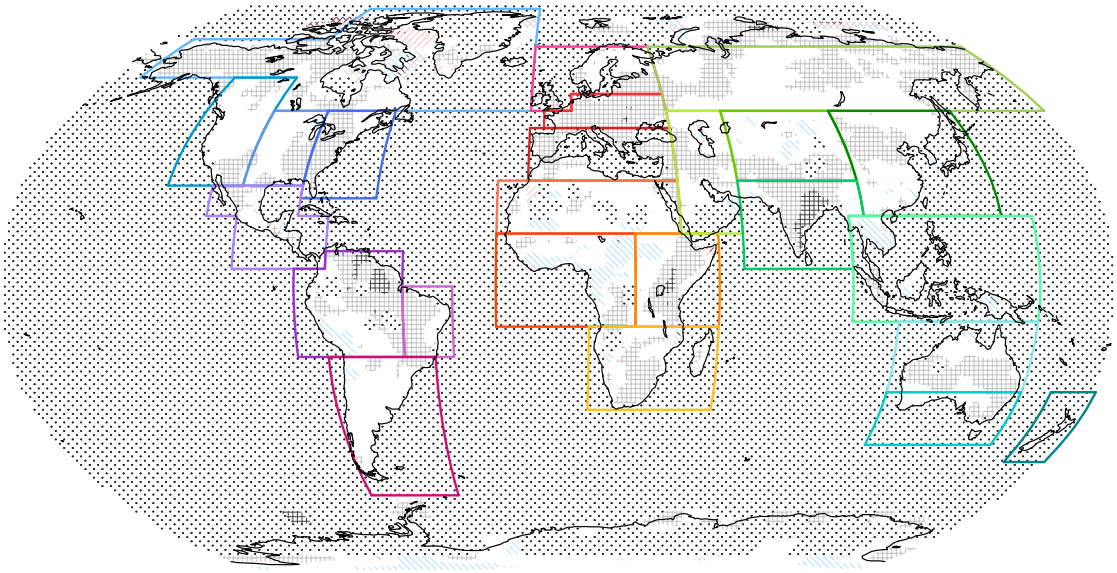


TAS DJF Obs outside ensemble spread (%)

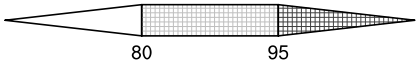


CanESM2 vs GISTEMPv4

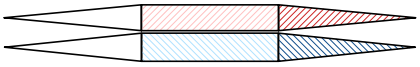
White Area = 50.9 %



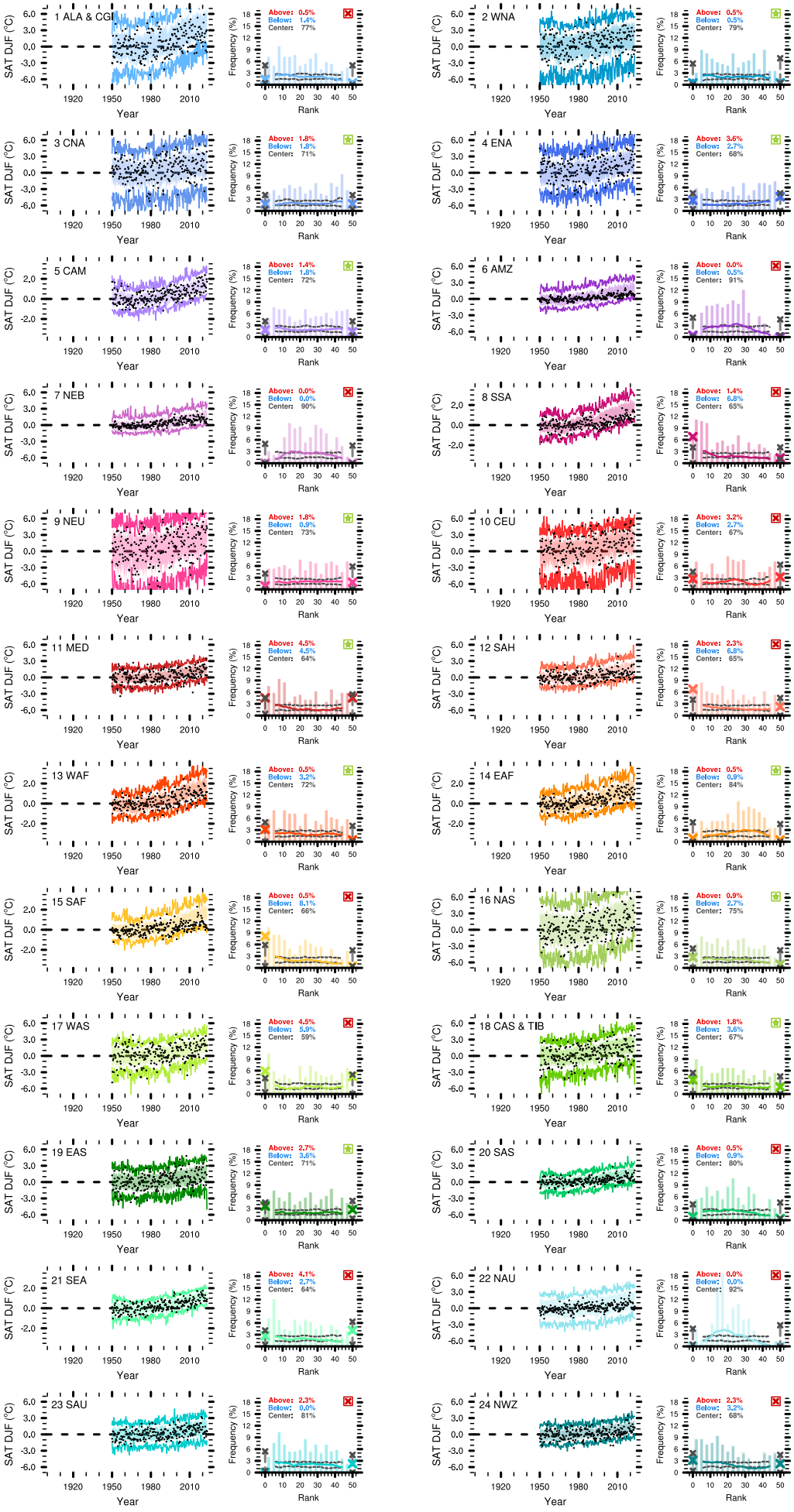
TAS JJA Obs inside central 75th percentile (%)



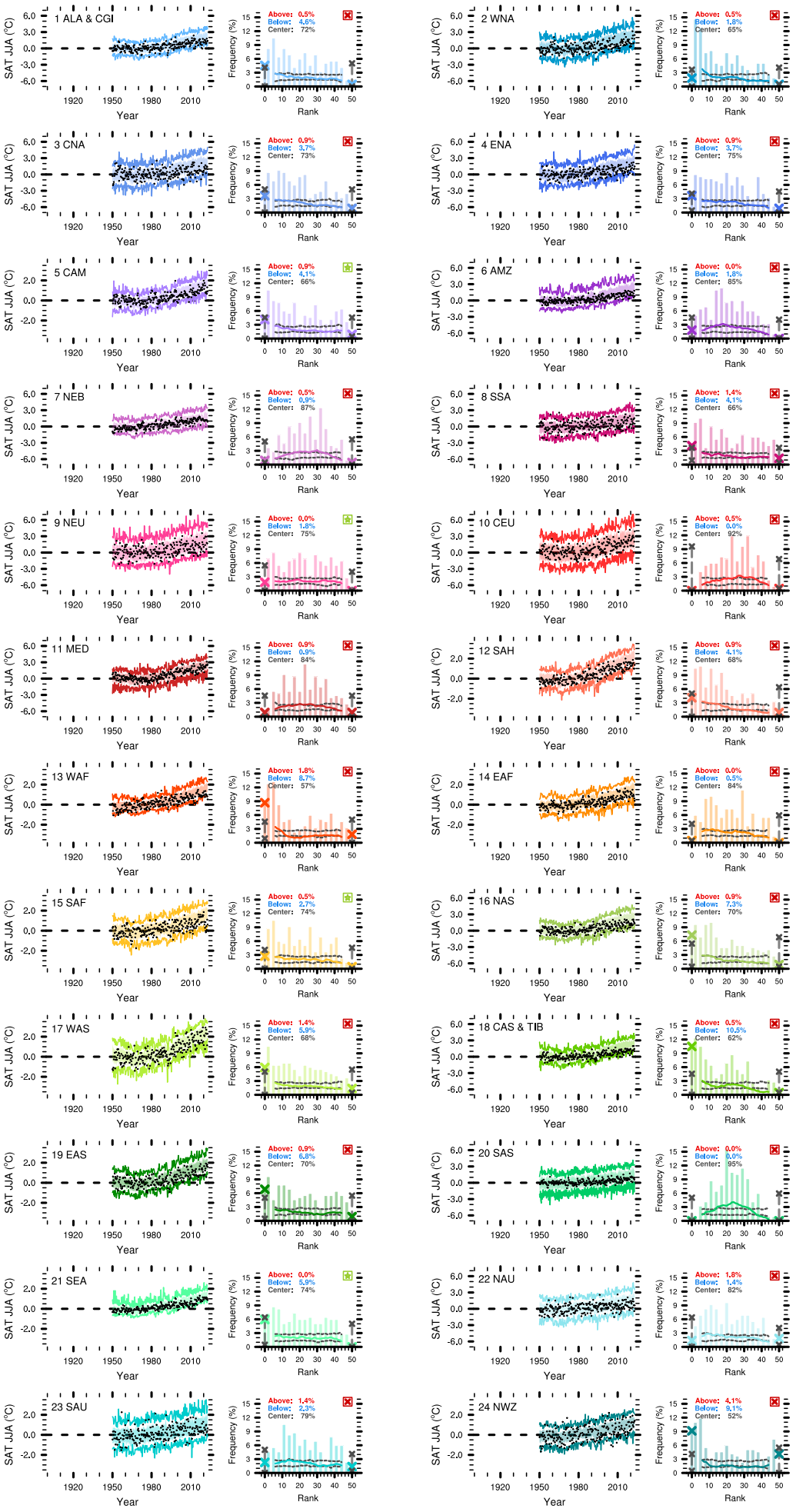
TAS JJA Obs outside ensemble spread (%)



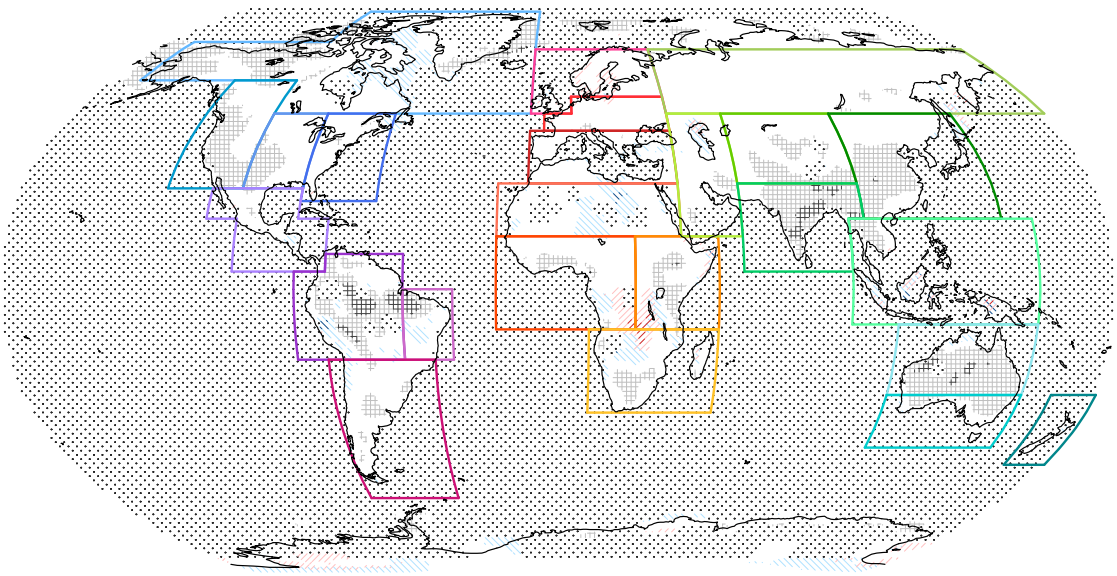
CanESM2 vs GISTEMPv4 TAS DJF



CanESM2 vs GISTEMPv4 TAS JJA



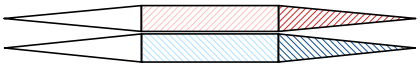
CanESM5 vs GISTEMPv4 White Area = 60.2 %



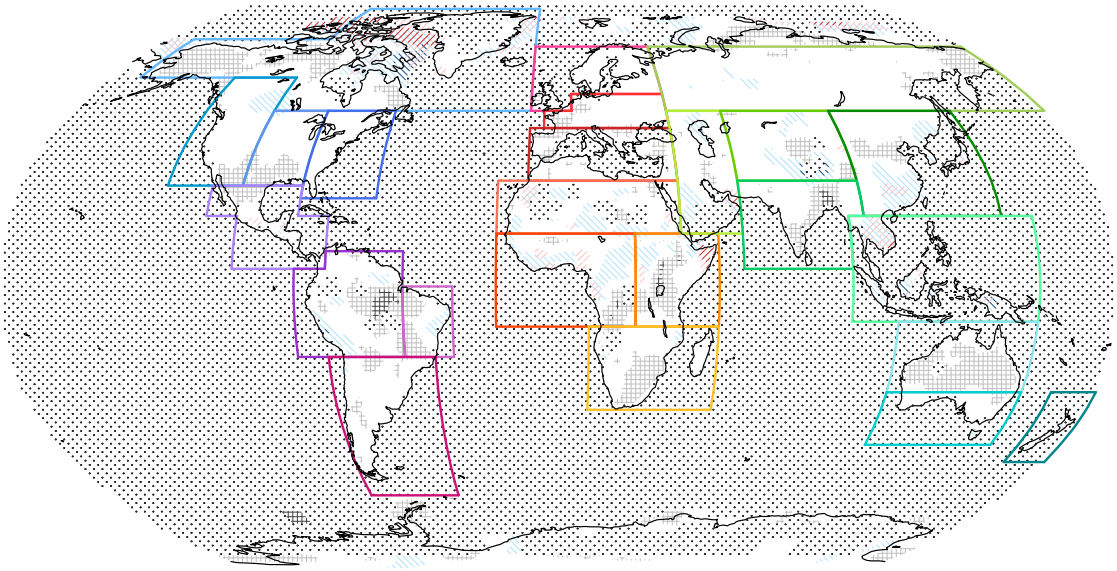
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



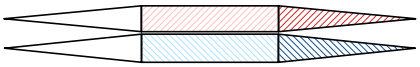
CanESM5 vs GISTEMPv4 White Area = 61.0 %



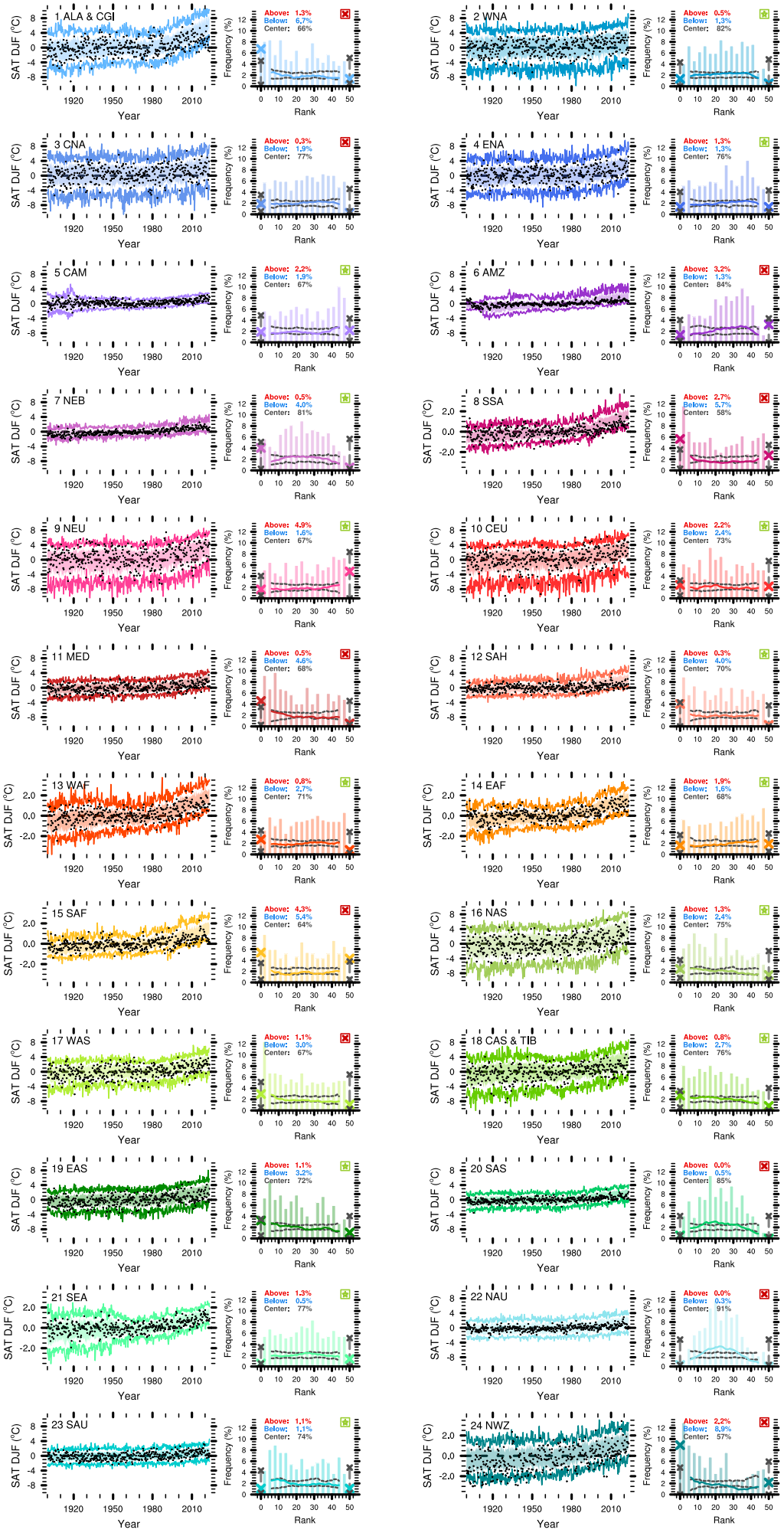
TAS JJA Obs inside central 75th percentile (%)



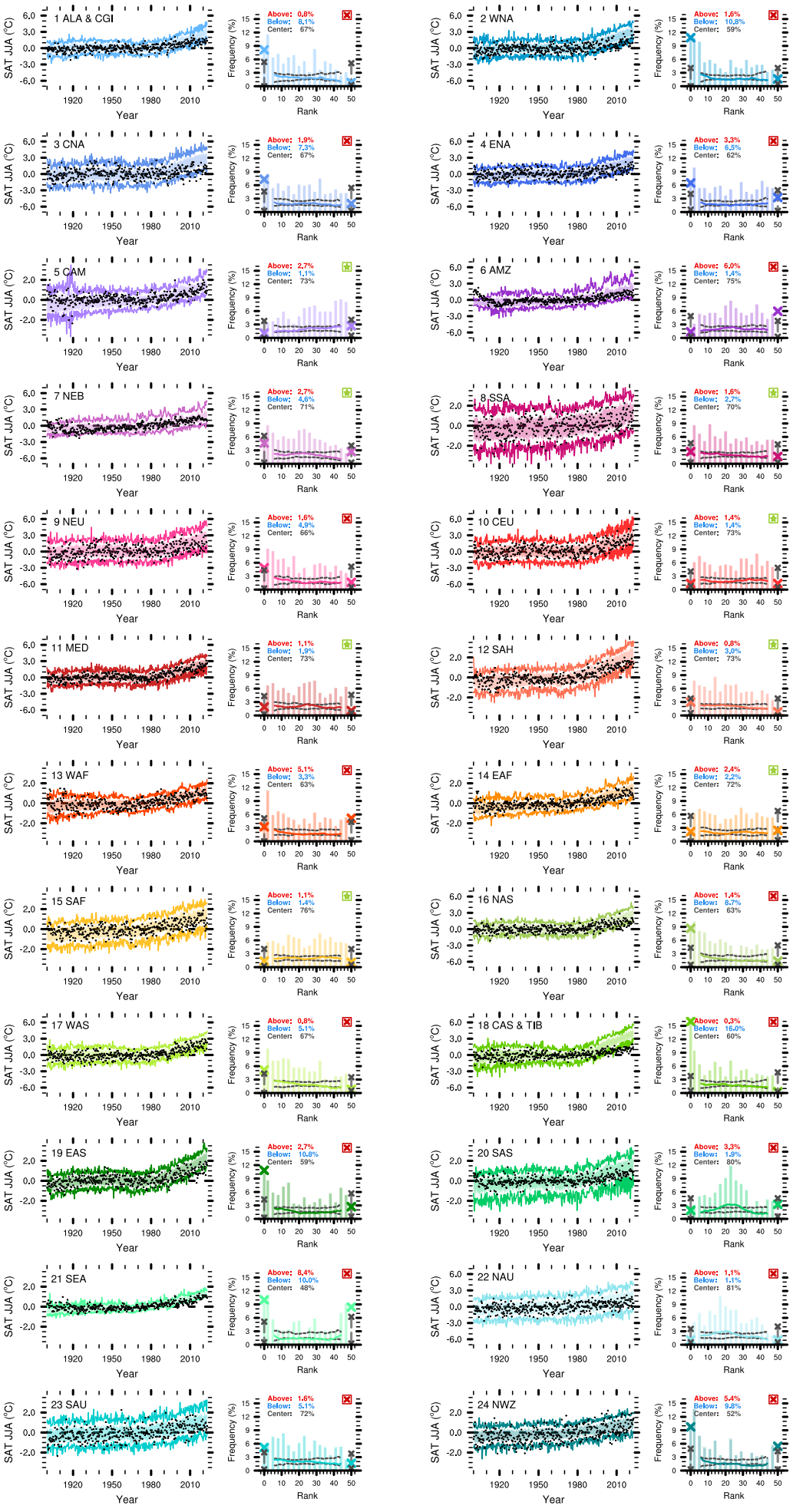
TAS JJA Obs outside ensemble spread (%)



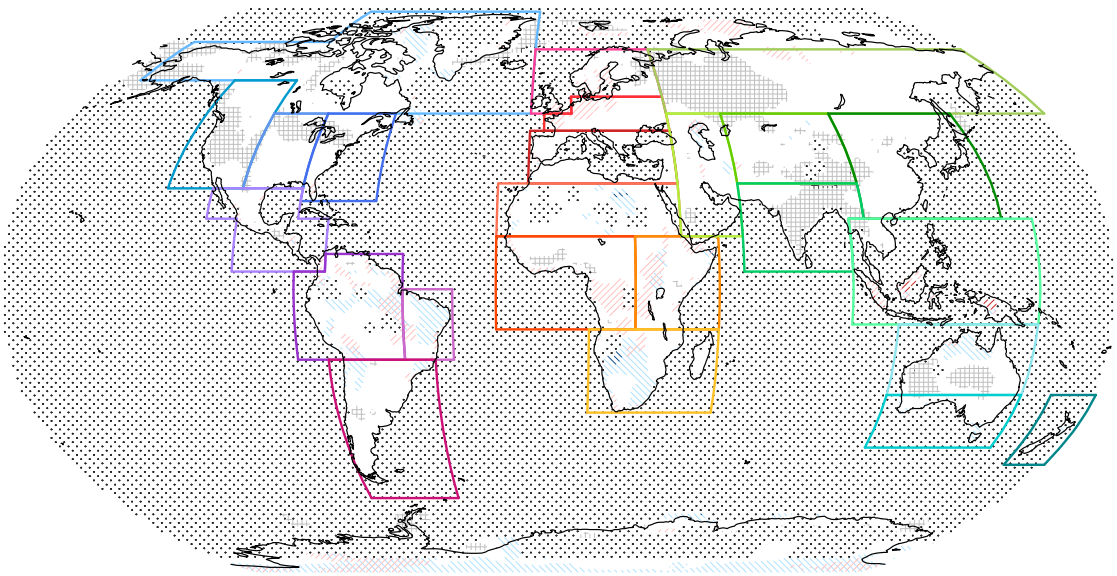
CanESM5 vs GISTEMPv4 TAS DJF



CanESM5 vs GISTEMPv4 TAS JJA



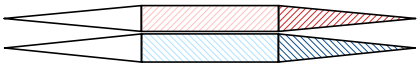
CESM-LE vs GISTEMPv4 White Area = 64.1 %



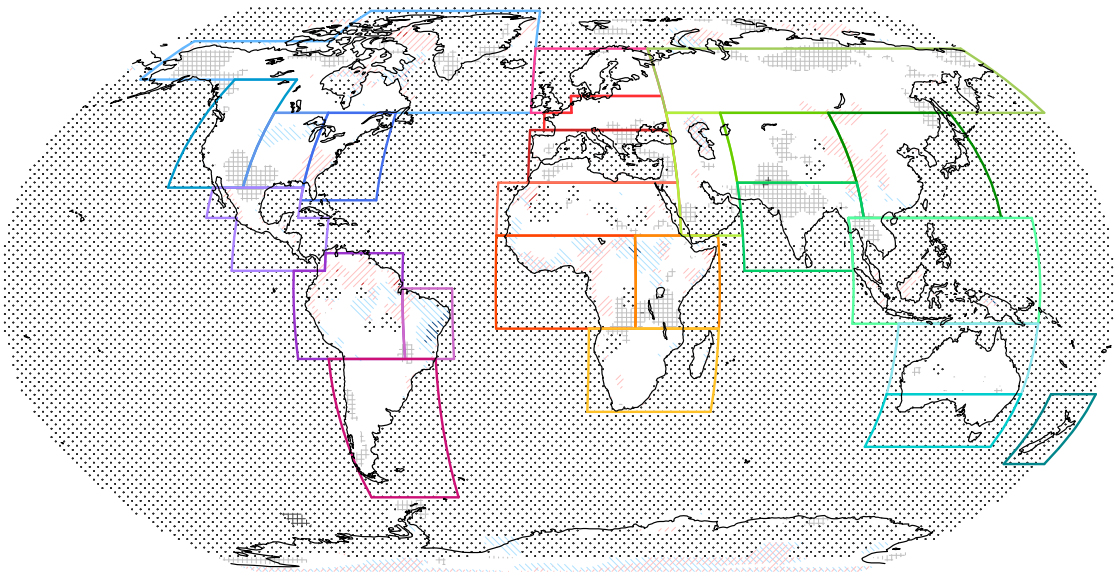
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



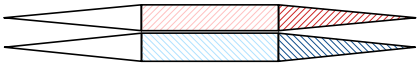
CESM-LE vs GISTEMPv4 White Area = 65.9 %



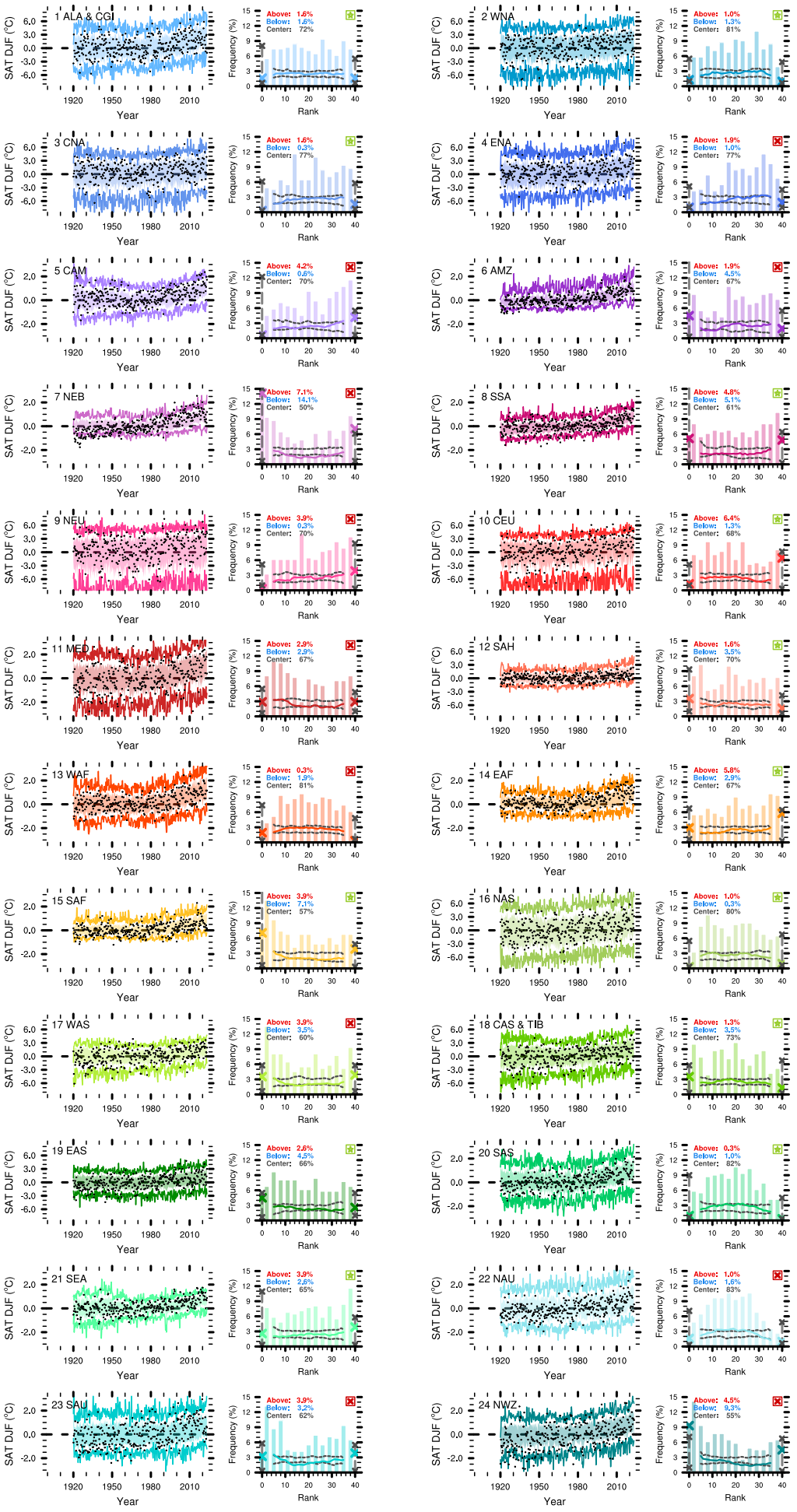
TAS JJA Obs inside central 75th percentile (%)



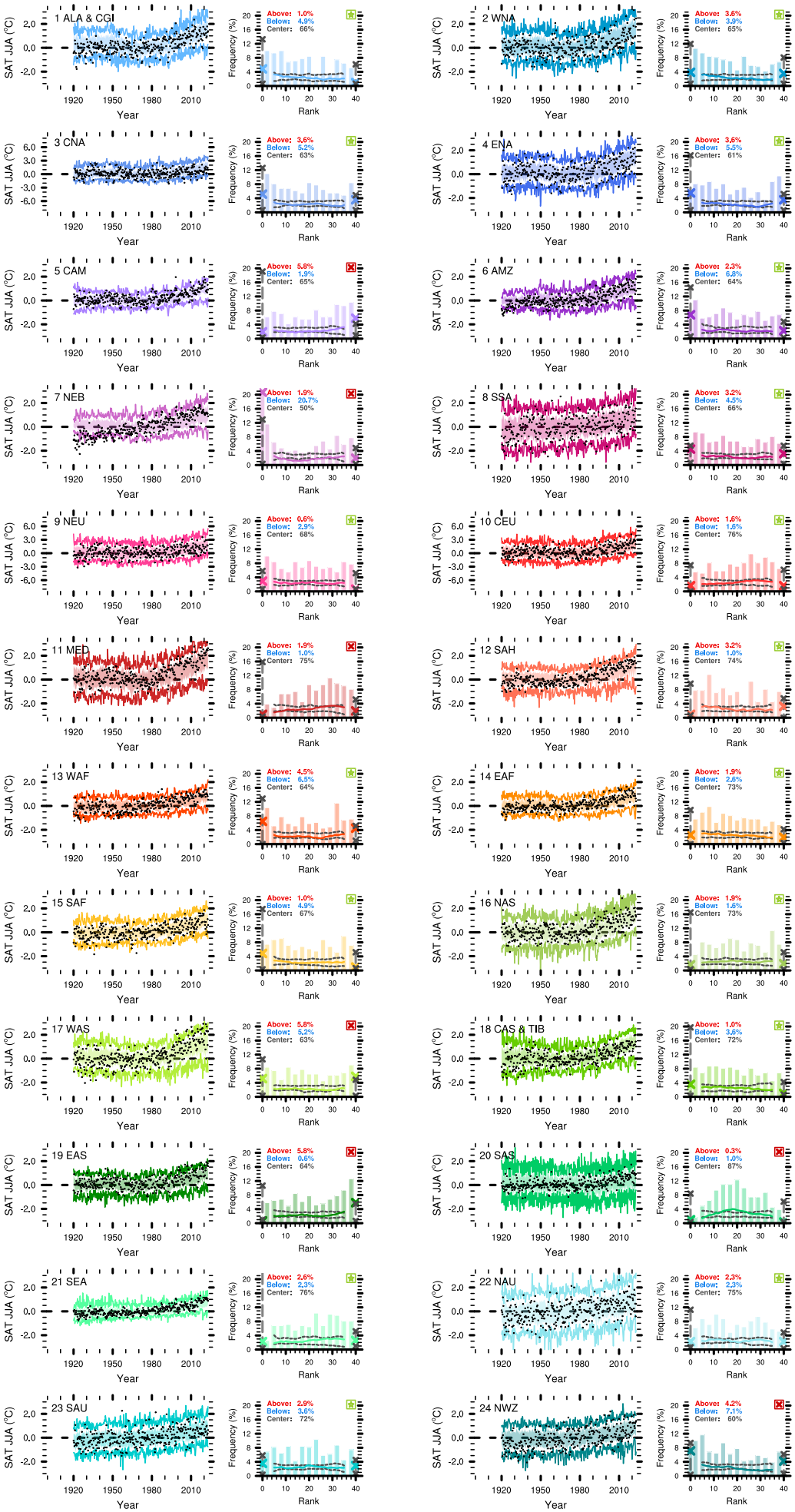
TAS JJA Obs outside ensemble spread (%)



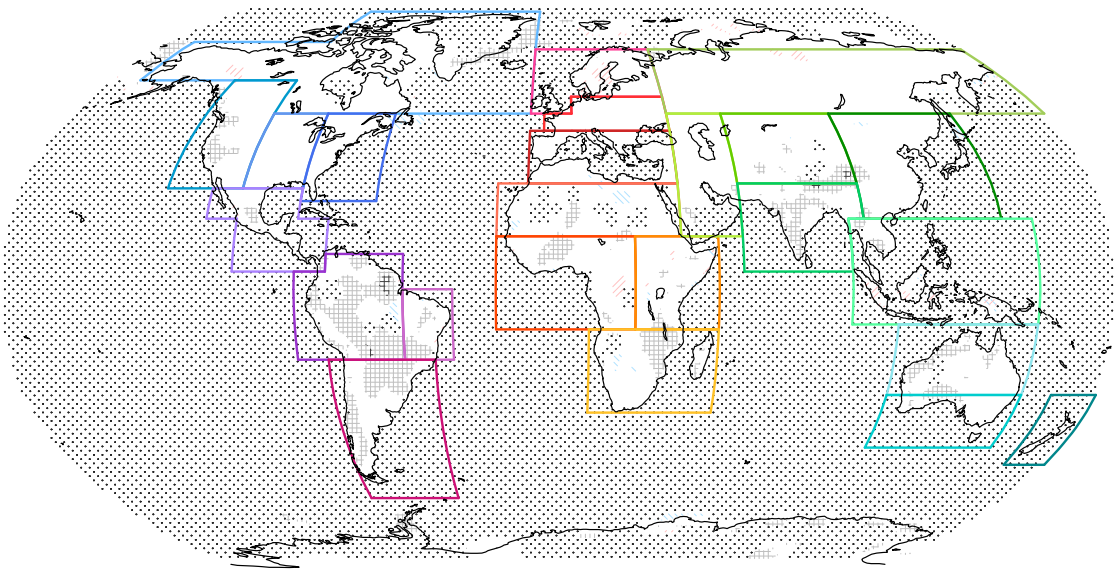
CESM-LE vs GISTEMPv4 TAS DJF



CESM-LE vs GISTEMPv4 TAS JJA



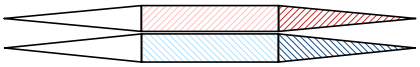
CESM2-LE vs GISTEMPv4 White Area = 81.4 %



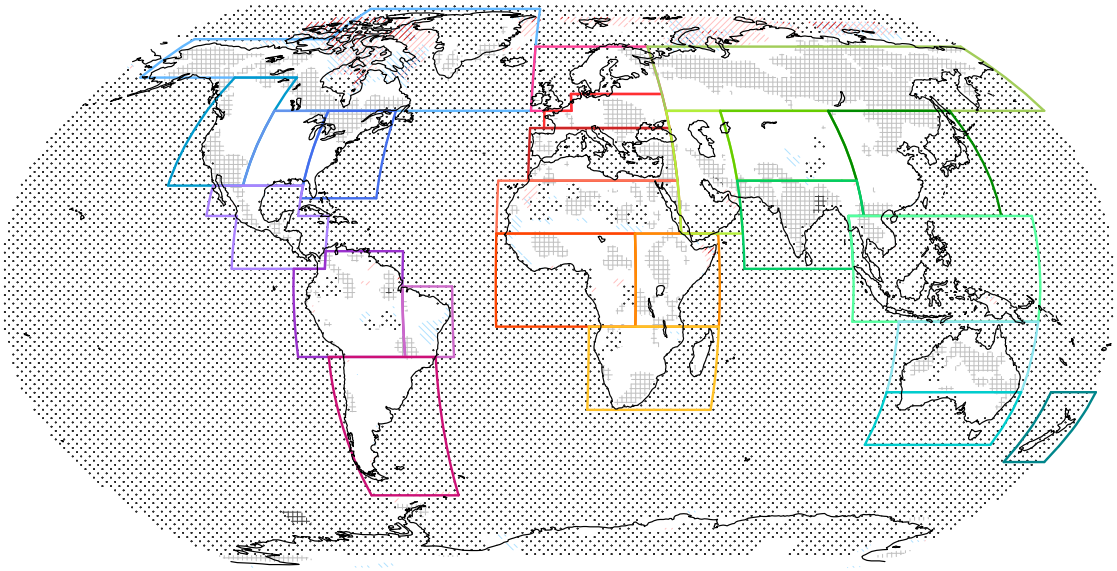
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



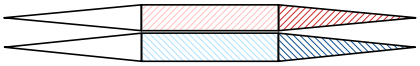
CESM2-LE vs GISTEMPv4 White Area = 63.5 %



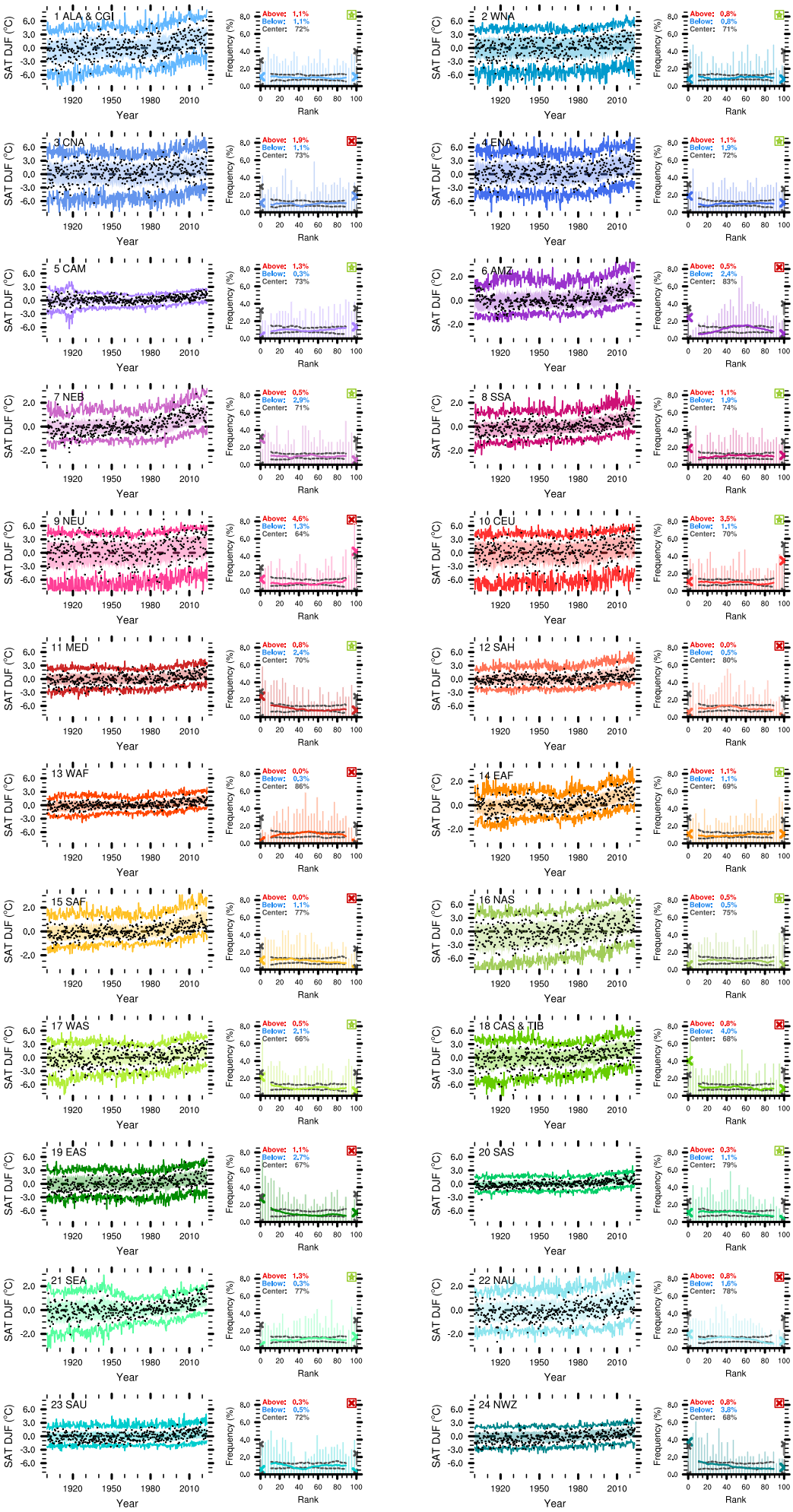
TAS JJA Obs inside central 75th percentile (%)



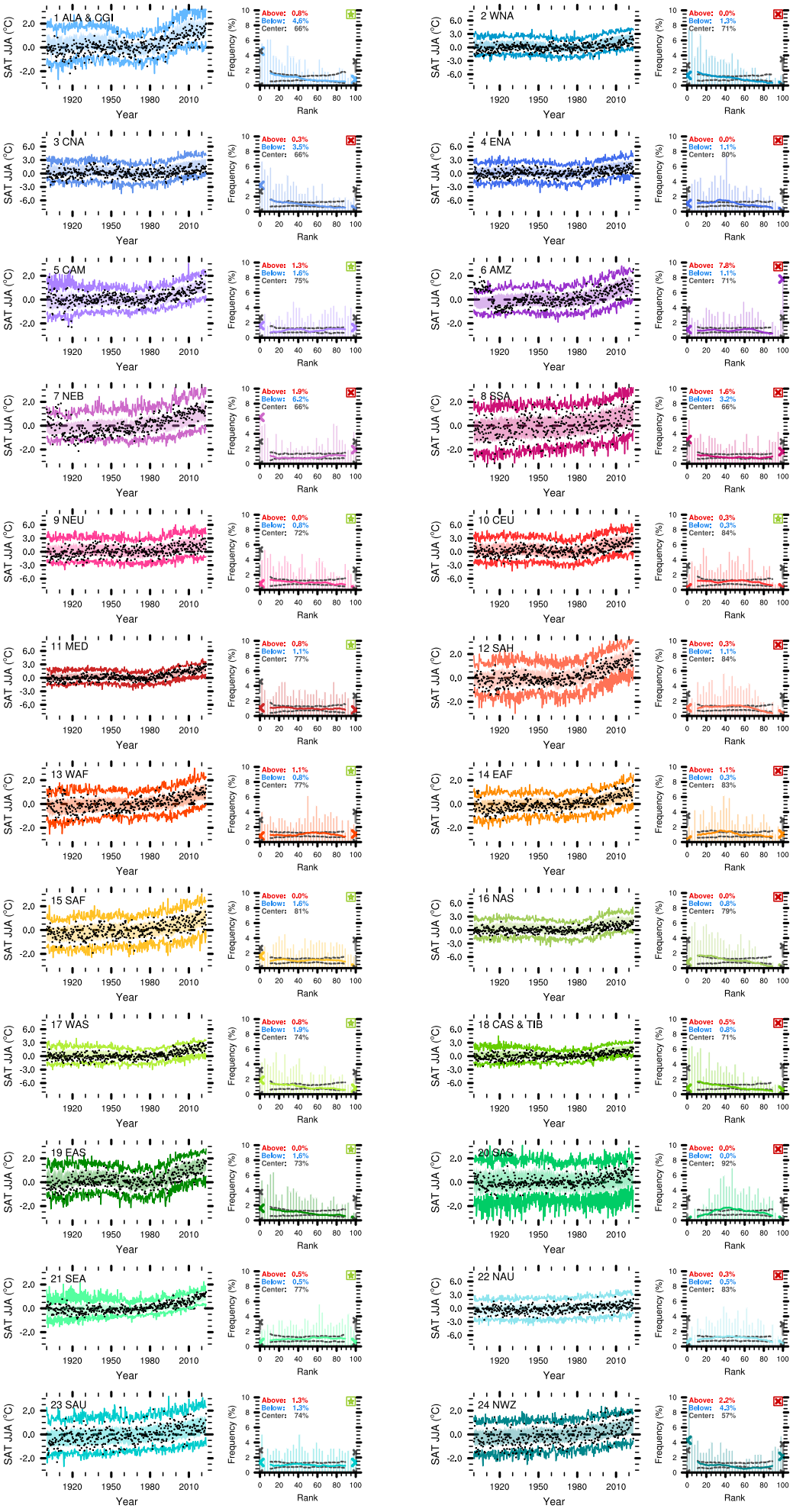
TAS JJA Obs outside ensemble spread (%)



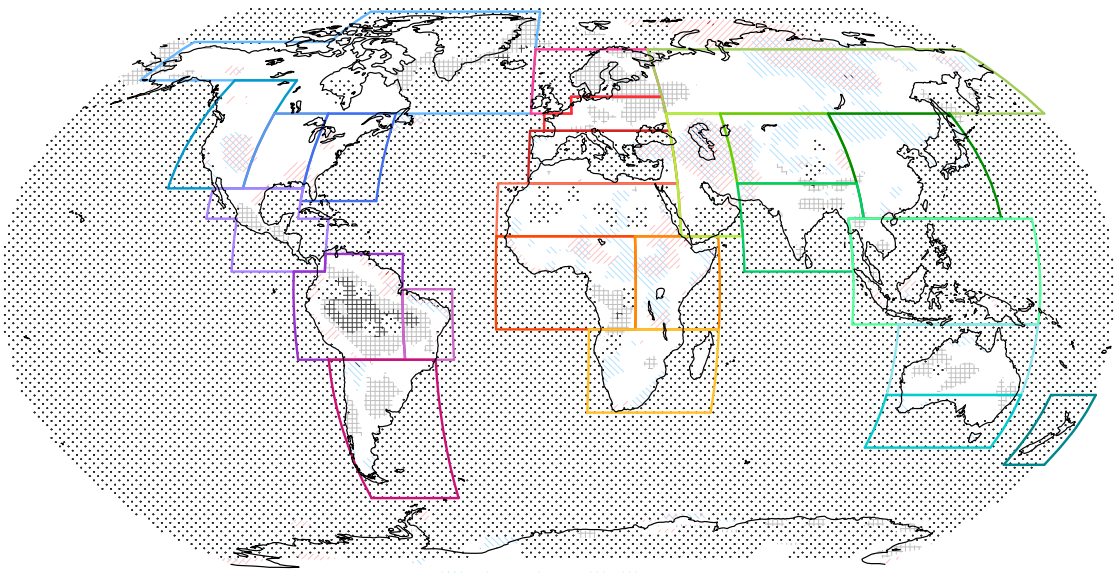
CESM2-LE vs GISTEMPv4 TAS DJF



CESM2-LE vs GISTEMPv4 TAS JJA



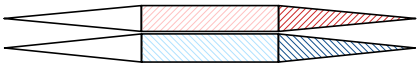
CSIRO-Mk360 vs GISTEMPv4 White Area = 57.4 %



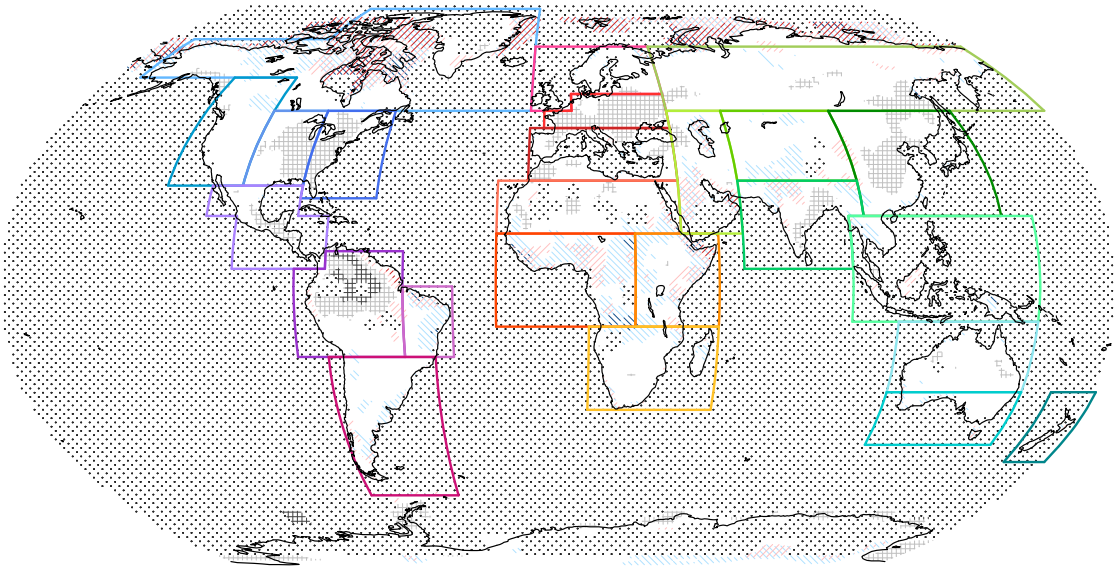
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



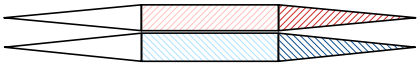
CSIRO-Mk360 vs GISTEMPv4 White Area = 54.9 %



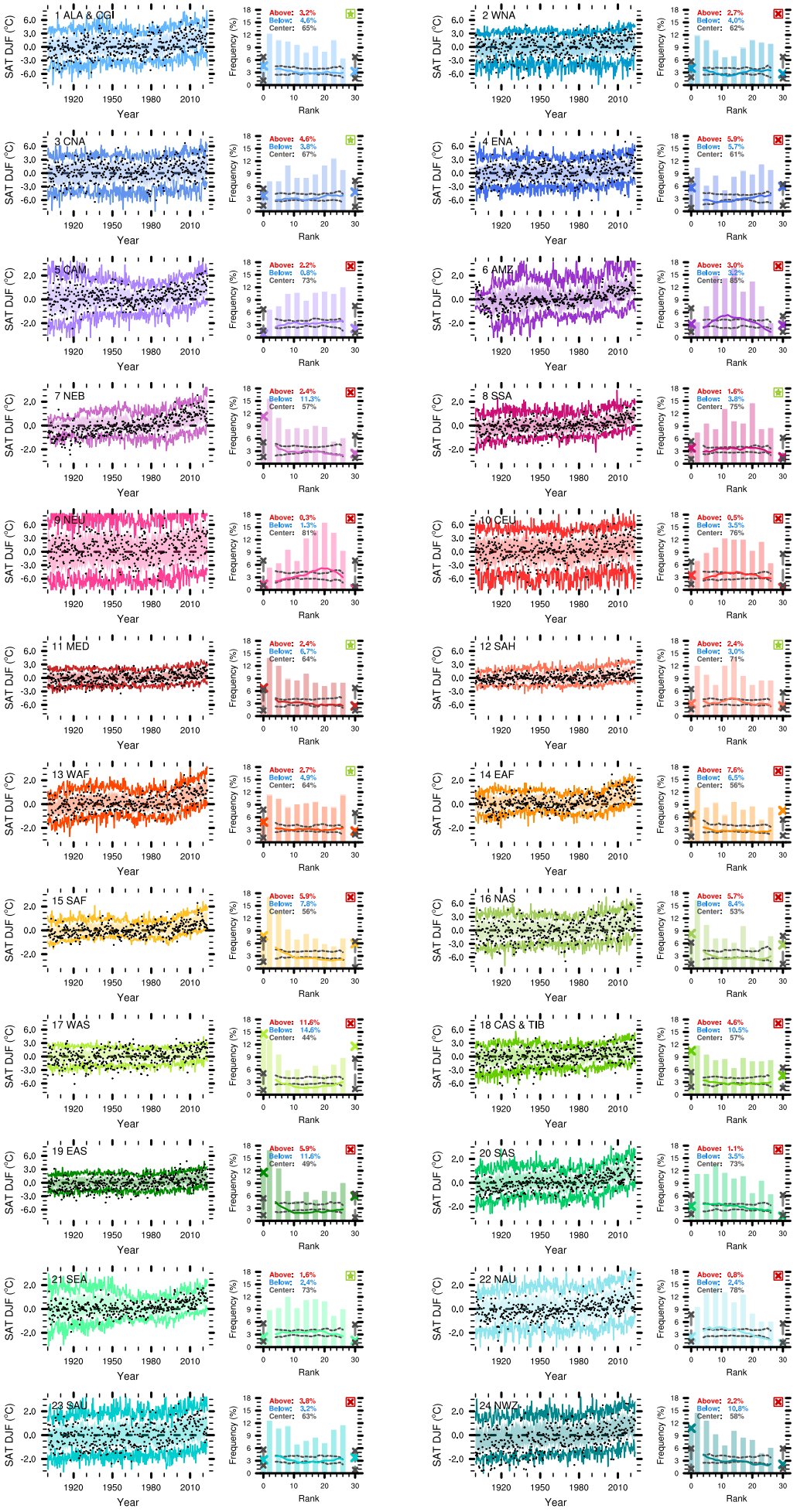
TAS JJA Obs inside central 75th percentile (%)



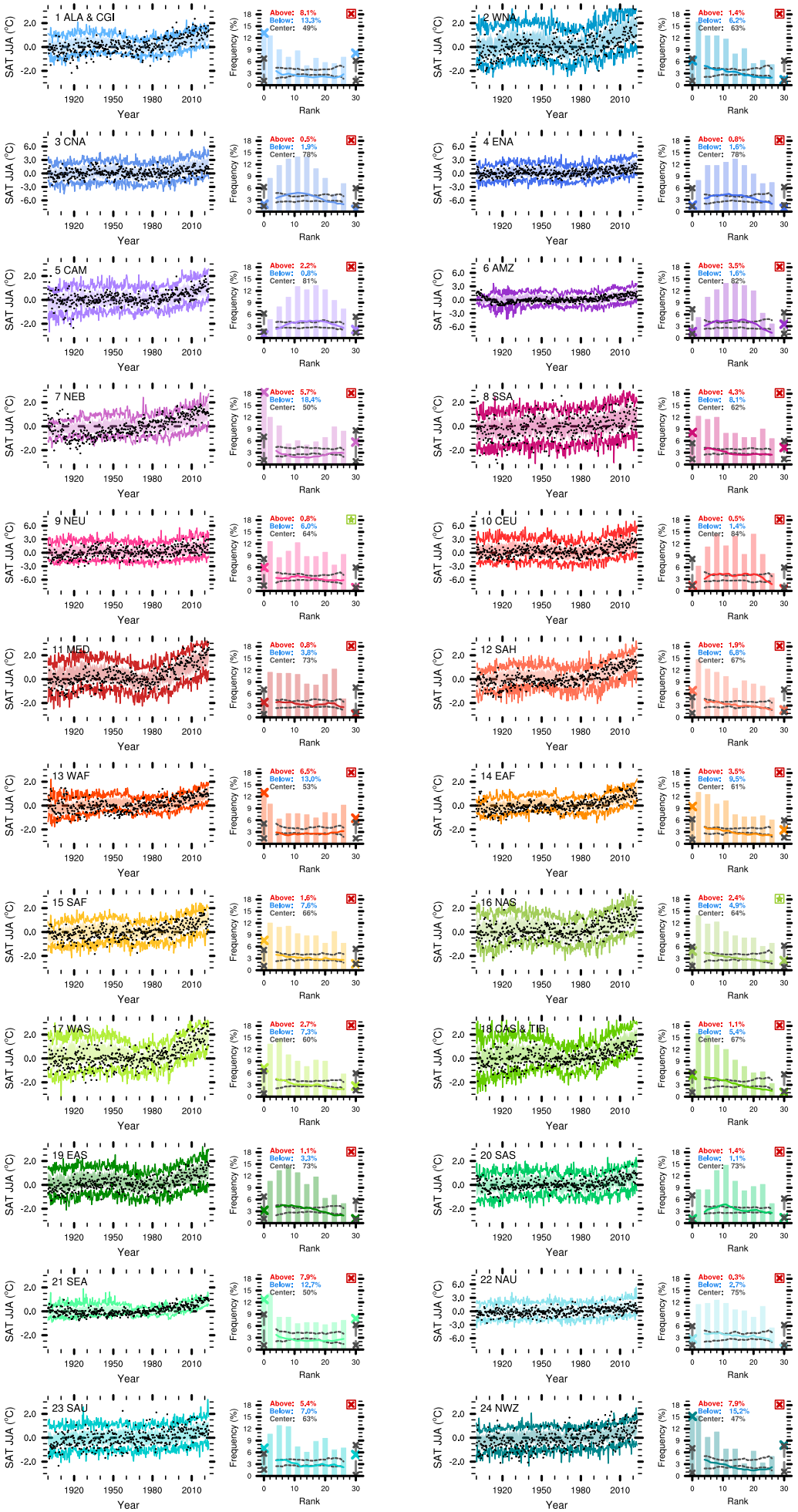
TAS JJA Obs outside ensemble spread (%)



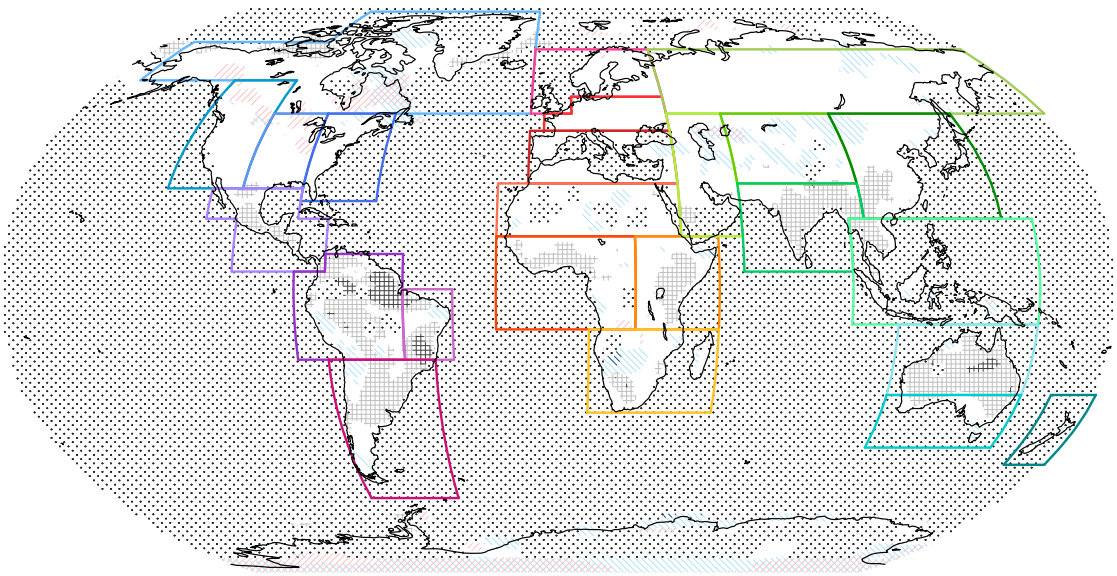
CSIRO-Mk360 vs GISTEMPv4 TAS DJF



CSIRO-Mk360 vs GISTEMPv4 TAS JJA



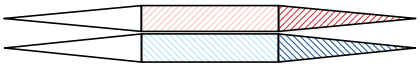
GFDL-ESM2M vs GISTEMPv4 White Area = 62.6 %



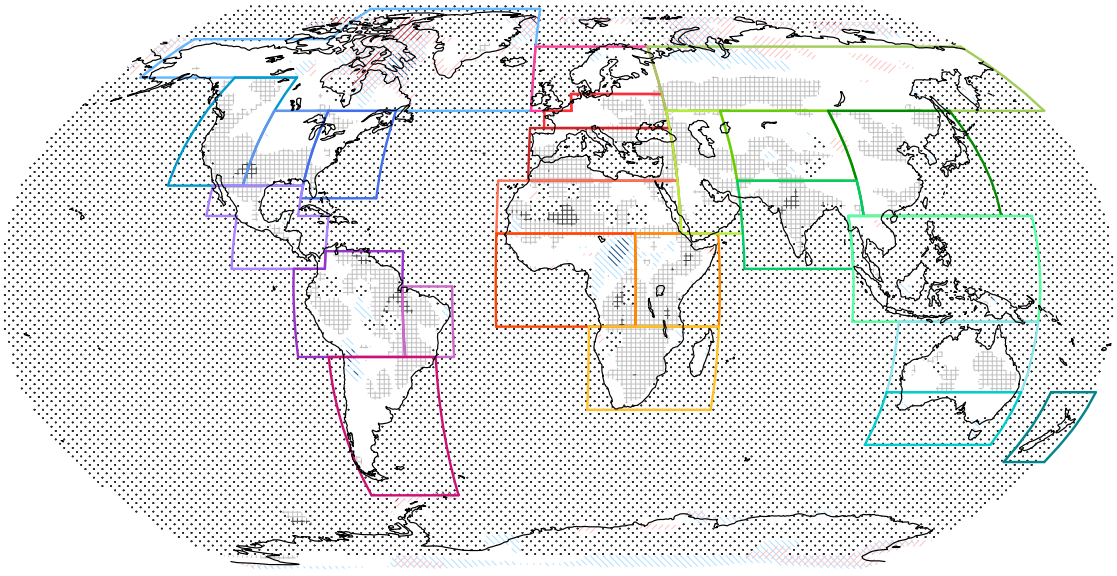
TAS DJF Obs inside central 75th percentile (%)



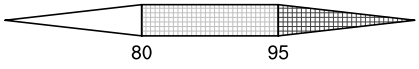
TAS DJF Obs outside ensemble spread (%)



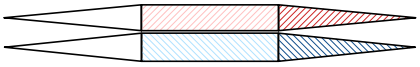
GFDL-ESM2M vs GISTEMPv4 White Area = 50.5 %



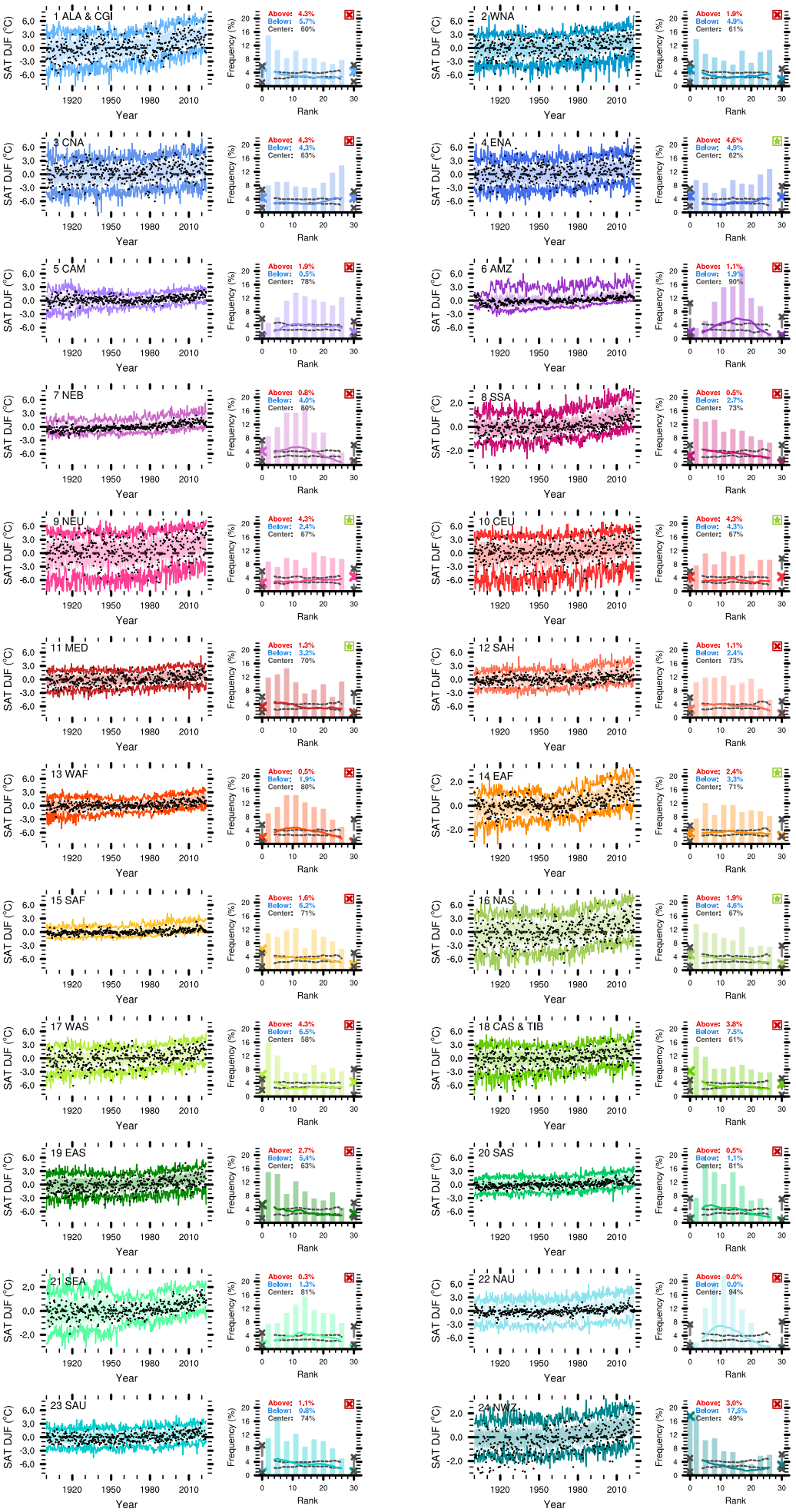
TAS JJA Obs inside central 75th percentile (%)



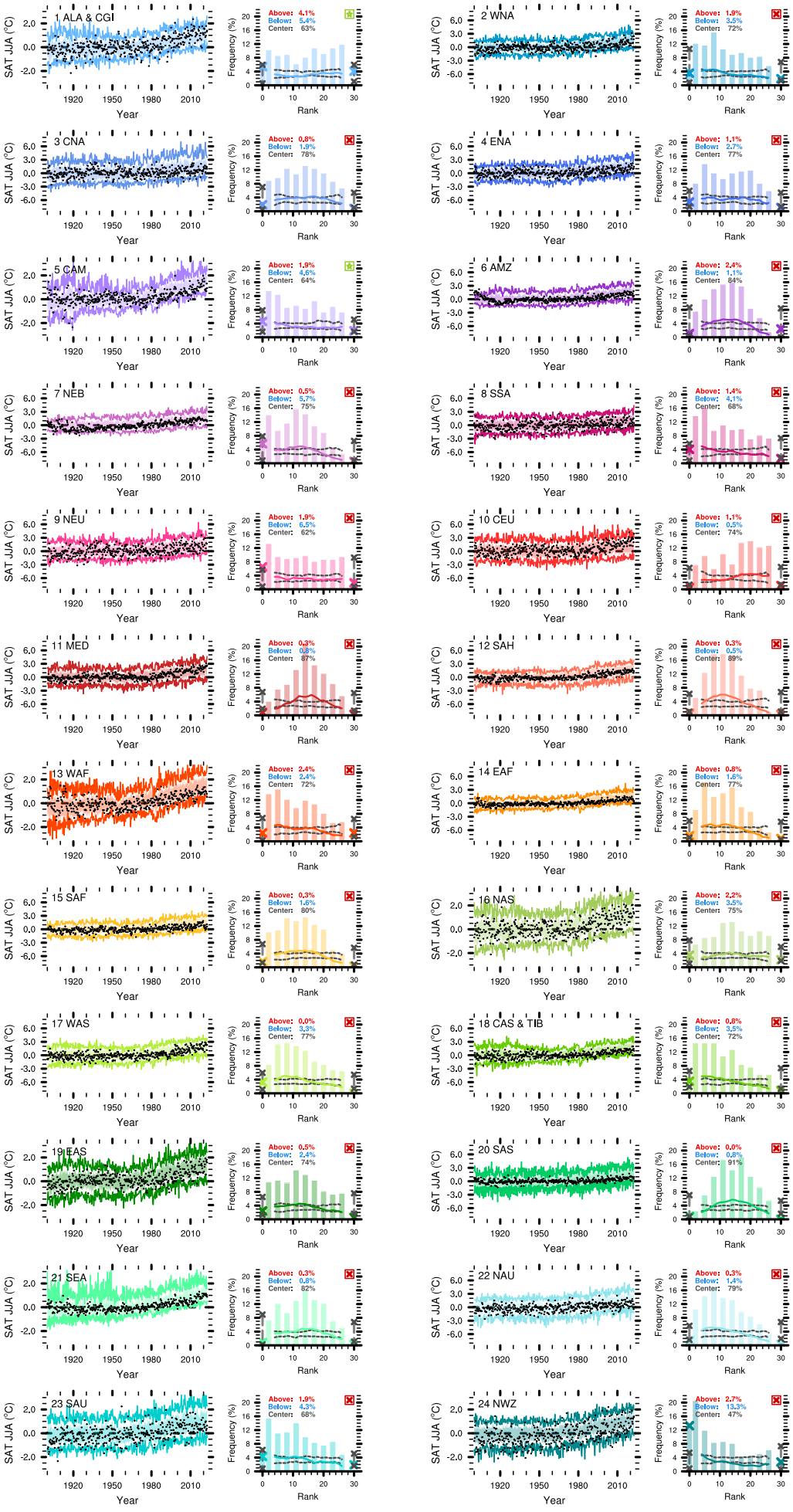
TAS JJA Obs outside ensemble spread (%)



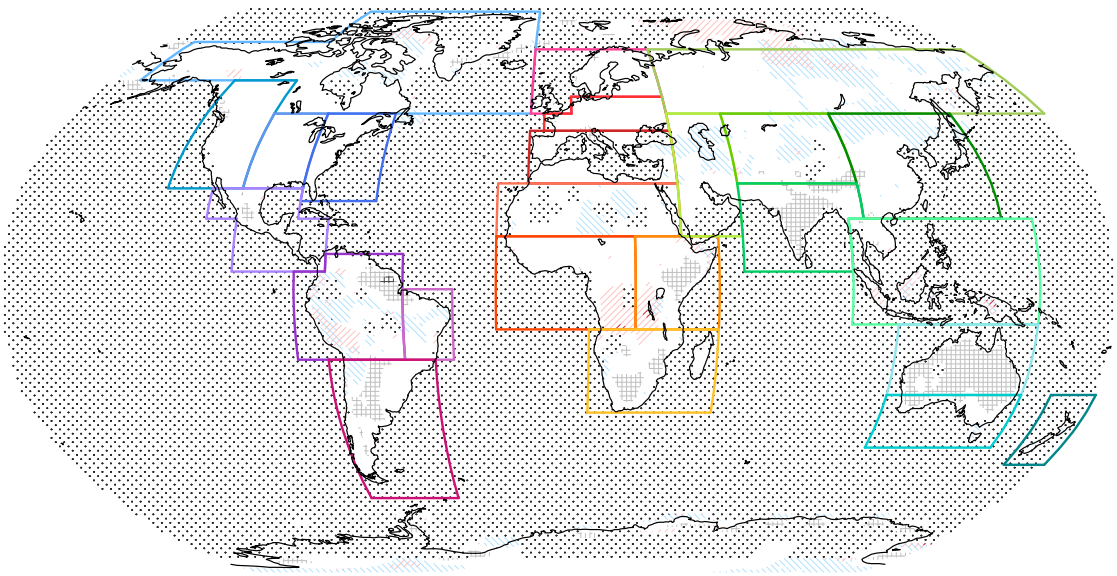
GFDL-ESM2M vs GISTEMPv4 TAS DJF



GFDL-ESM2M vs GISTEMPv4 TAS JJA



GFDL-SPEAR vs GISTEMPv4 White Area = 67.9 %



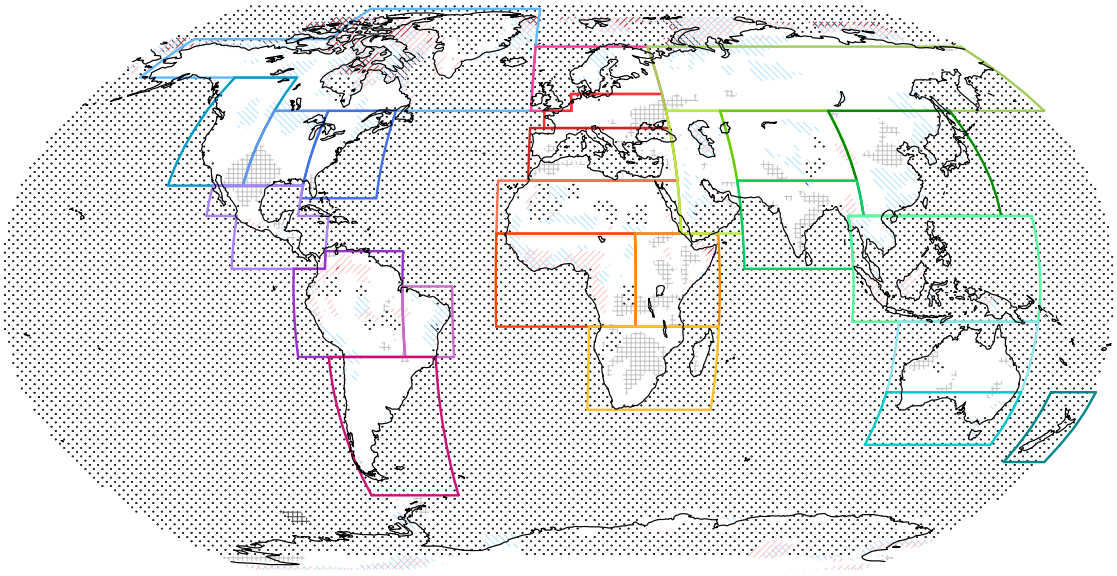
TAS DJF Obs inside central 75th percentile (%)

80 95

TAS DJF Obs outside ensemble spread (%)

8 20

GFDL-SPEAR vs GISTEMPv4 White Area = 64.0 %



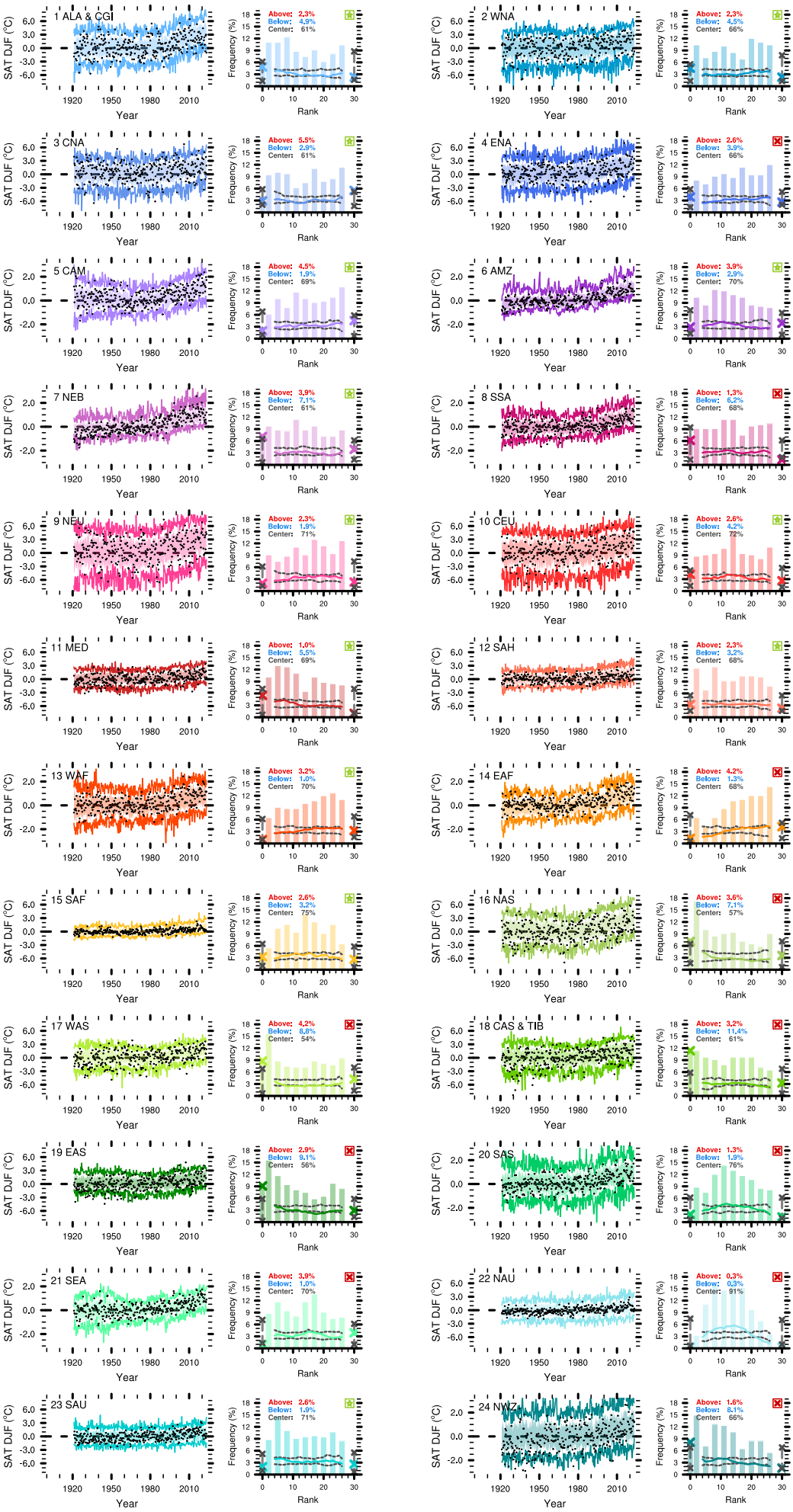
TAS JJA Obs inside central 75th percentile (%)

80 95

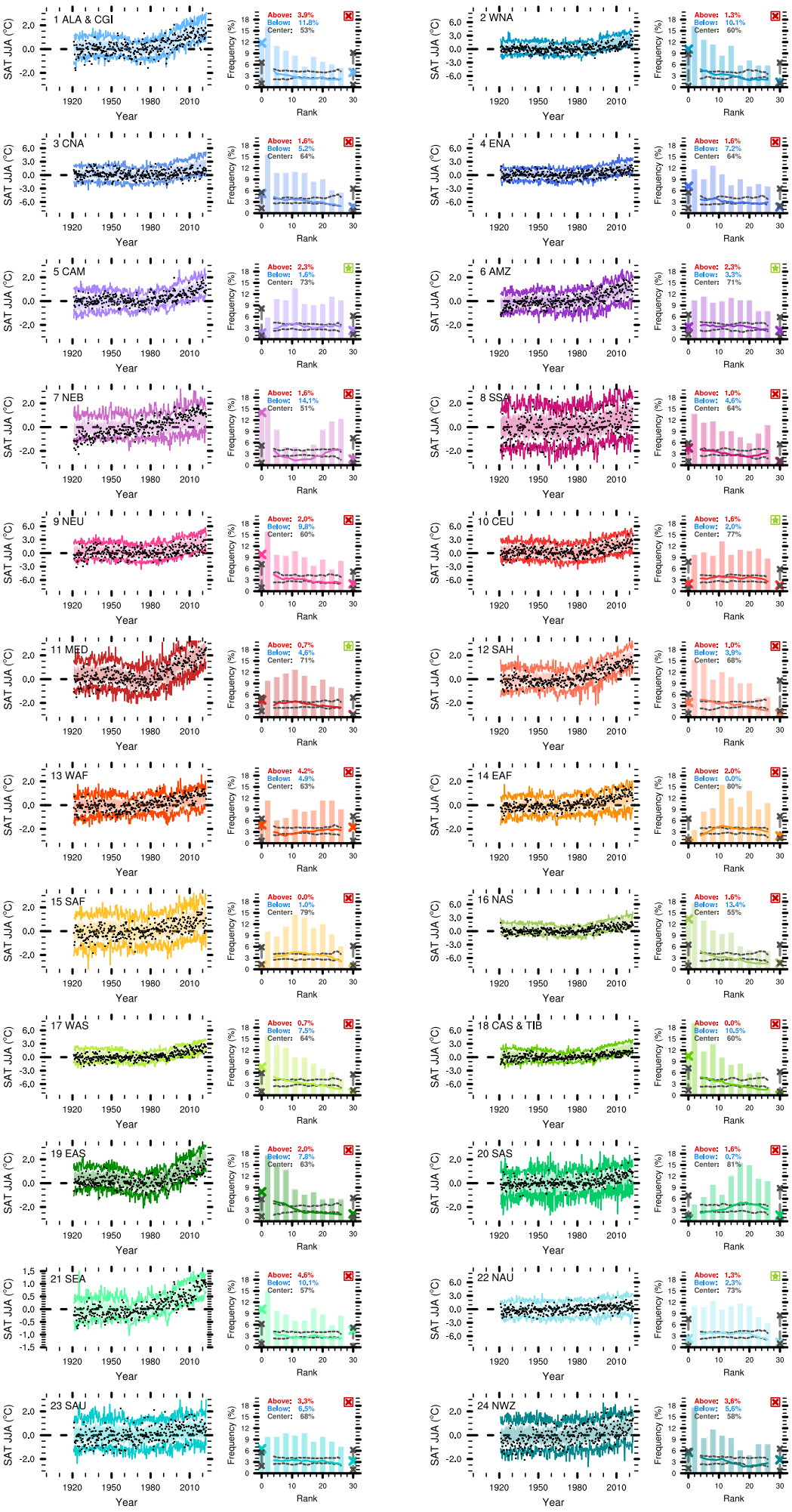
TAS JJA Obs outside ensemble spread (%)

8 20

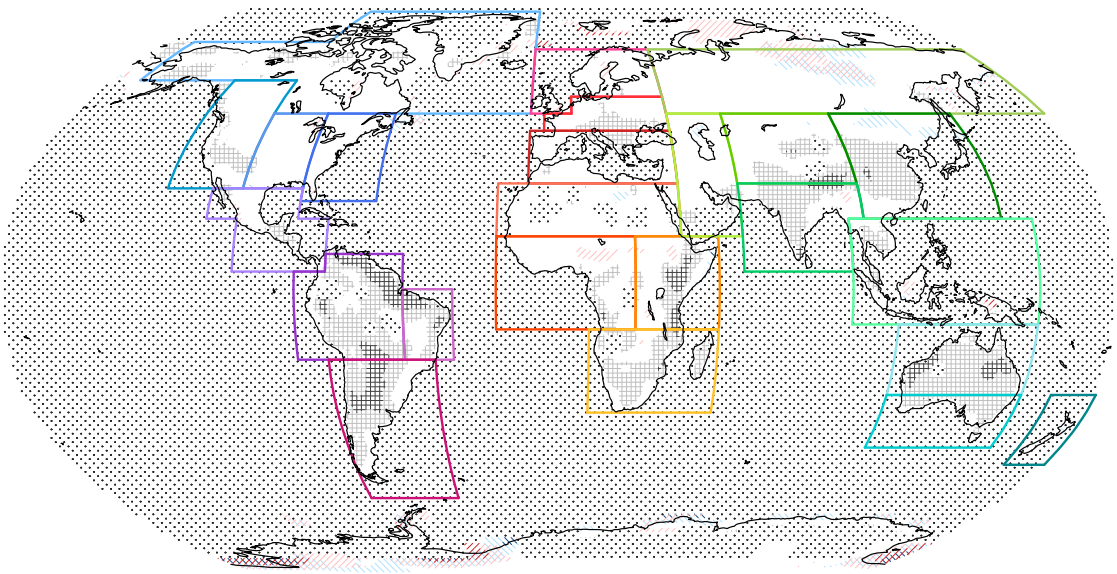
GFDL-SPEAR vs GISTEMPv4 TAS DJF



GFDL-SPEAR vs GISTEMPv4 TAS JJA



MIROC6 vs GISTEMPv4 White Area = 60.3 %



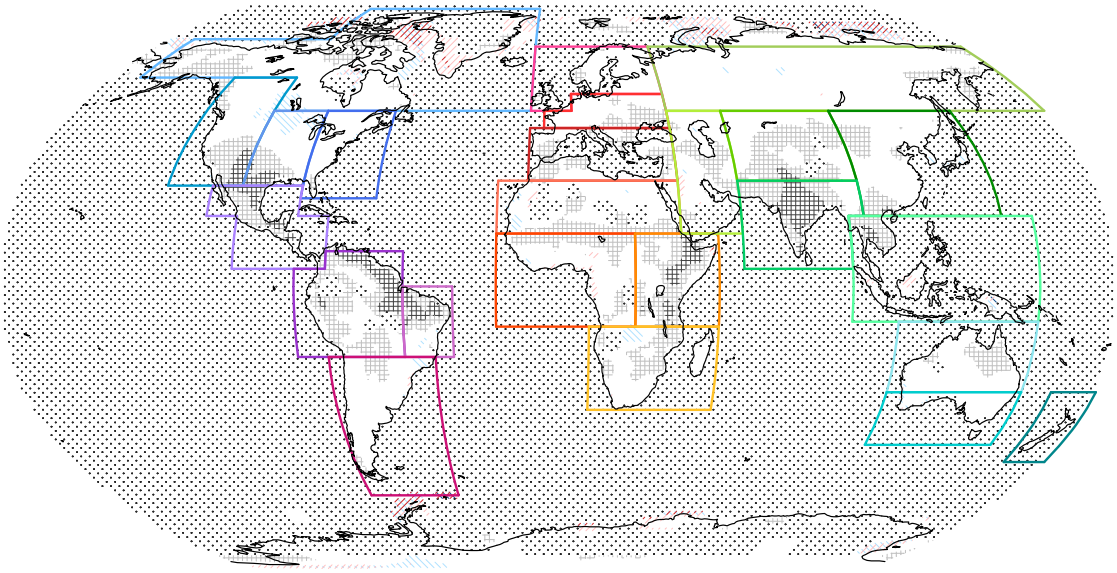
TAS DJF Obs inside central 75th percentile (%)

80 95

TAS DJF Obs outside ensemble spread (%)

8 20

MIROC6 vs GISTEMPv4 White Area = 59.5 %



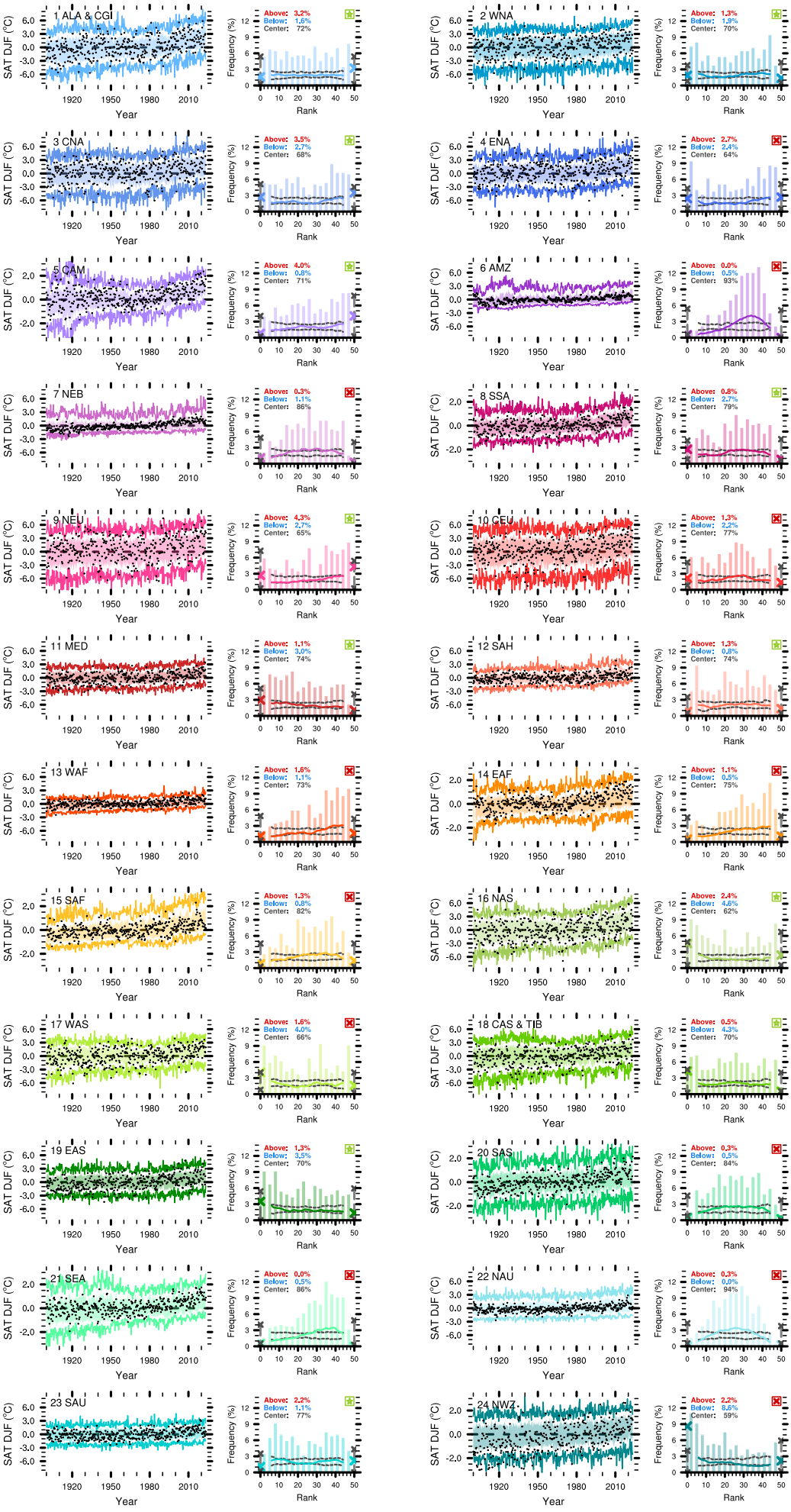
TAS JJA Obs inside central 75th percentile (%)

80 95

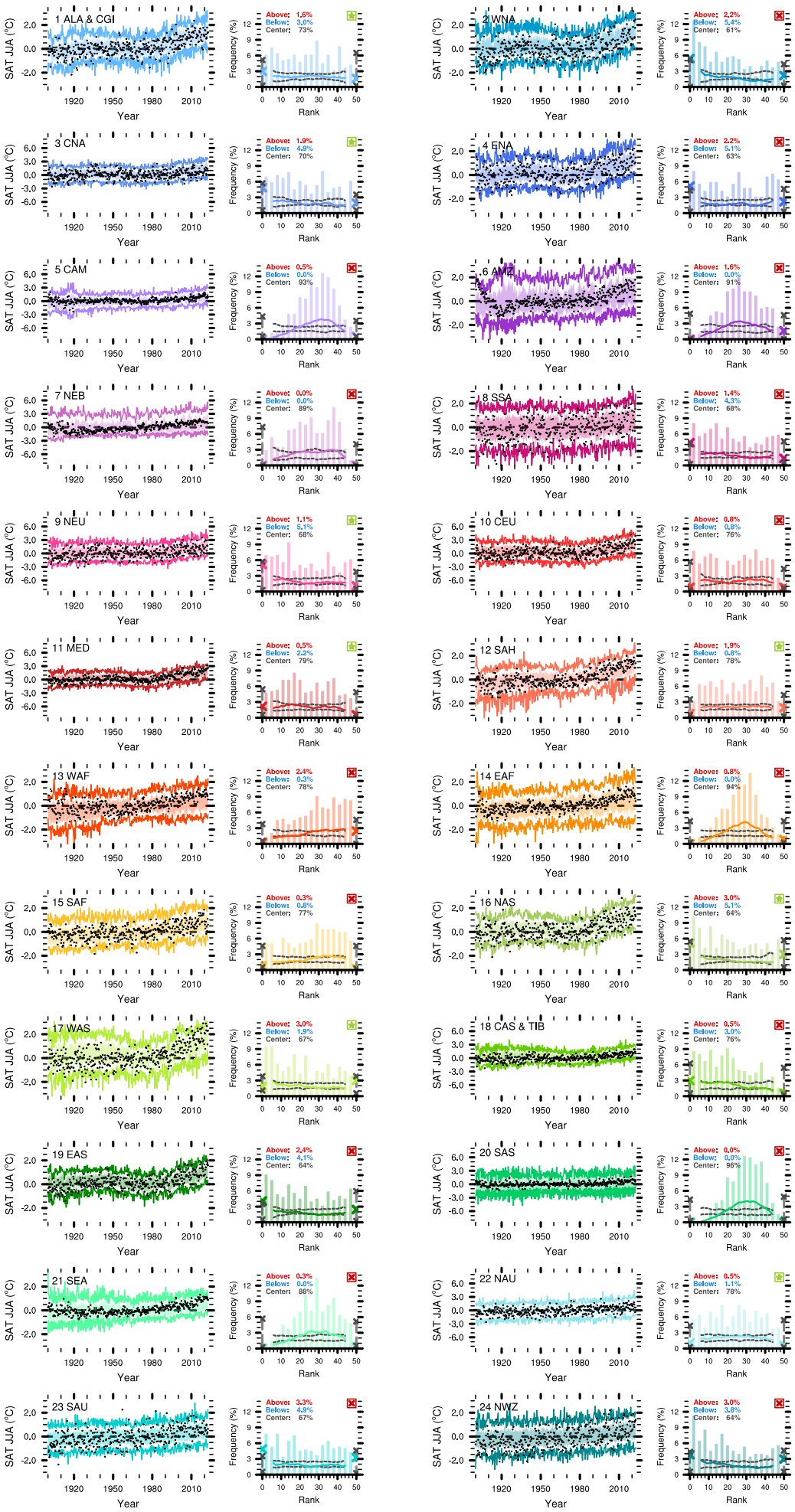
TAS JJA Obs outside ensemble spread (%)

8 20

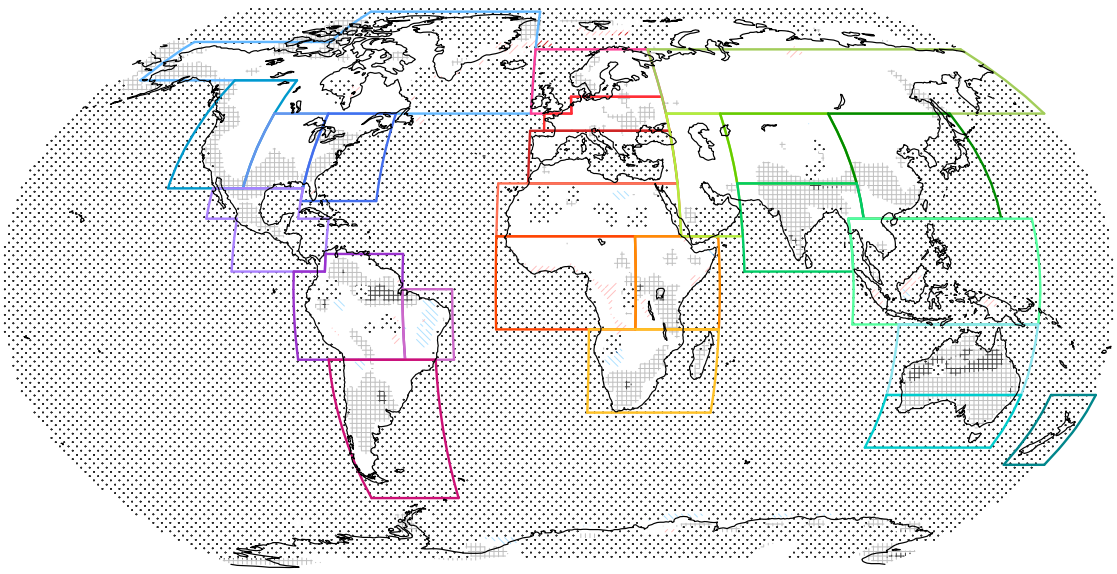
MIROC6 vs GISTEMPv4 TAS DJF



MIROC6 vs GISTEMPv4 TAS JJA



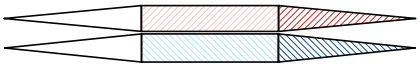
MPI-GE5 vs GISTEMPv4 White Area = 69.8 %



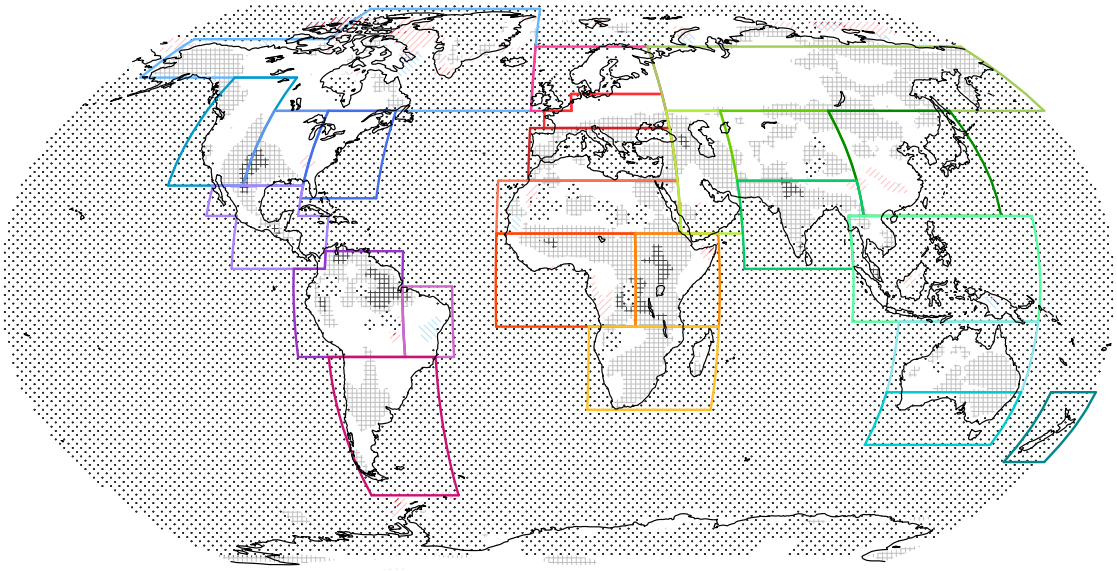
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



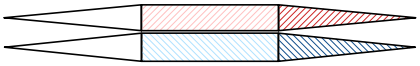
MPI-GE5 vs GISTEMPv4 White Area = 55.8 %



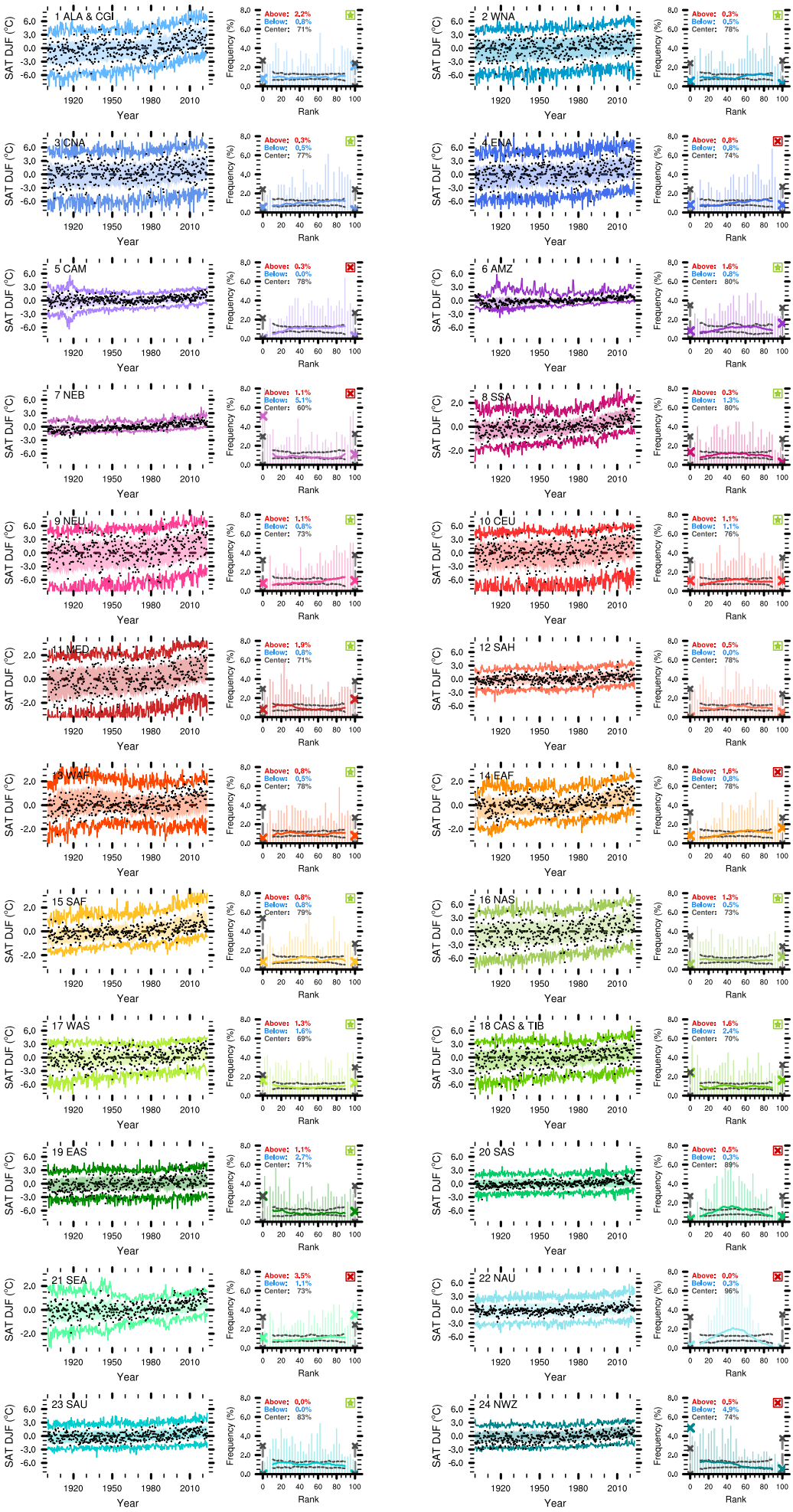
TAS JJA Obs inside central 75th percentile (%)



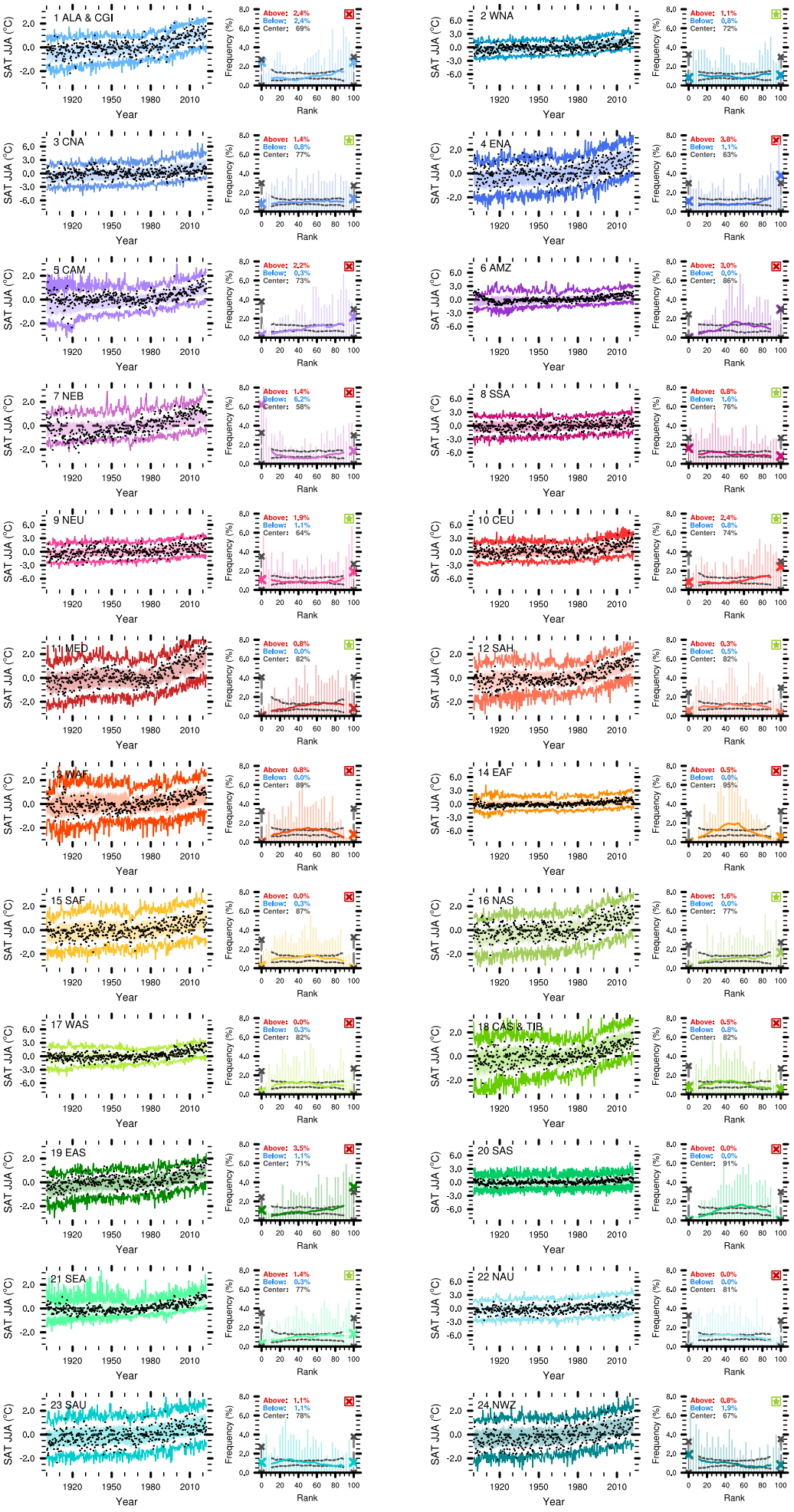
TAS JJA Obs outside ensemble spread (%)



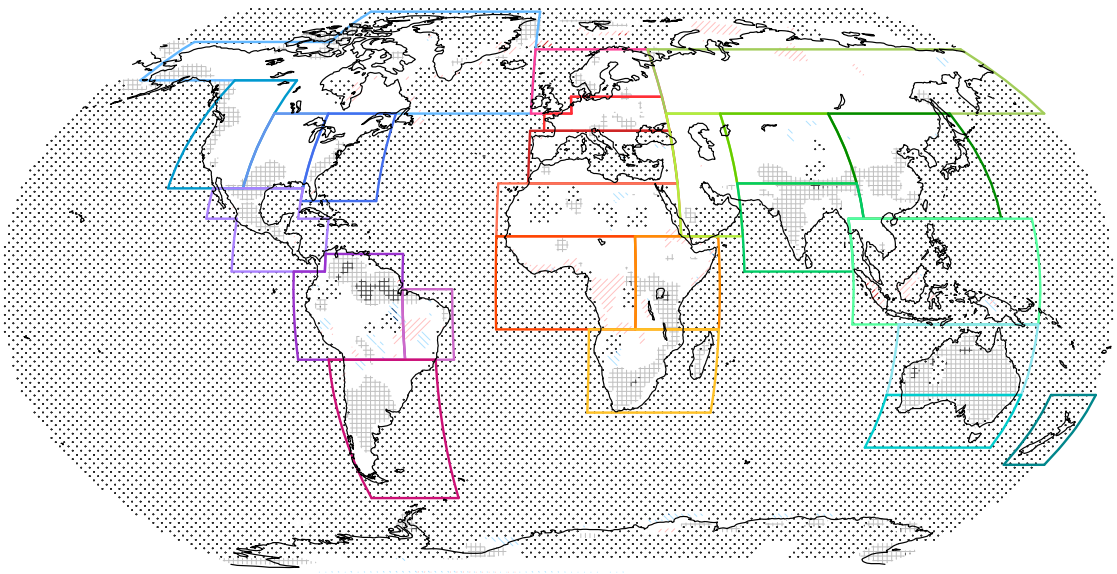
MPI-GE5 vs GISTEMPv4 TAS DJF



MPI-GE5 vs GISTEMPv4 TAS JJA



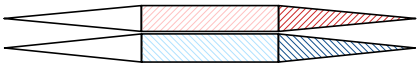
MPI-GE6 vs GISTEMPv4 White Area = 70.0 %



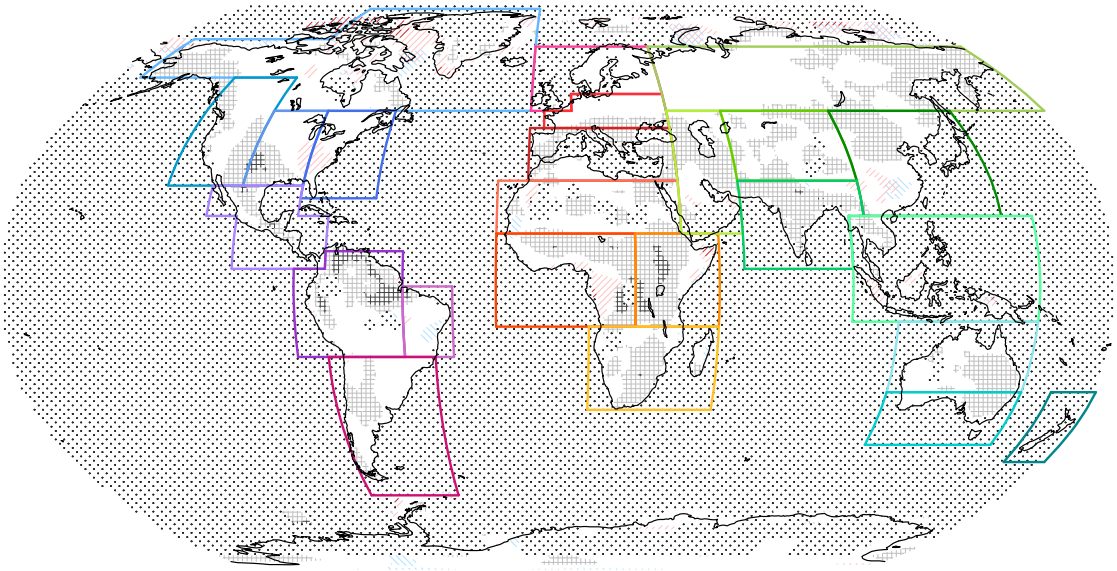
TAS DJF Obs inside central 75th percentile (%)



TAS DJF Obs outside ensemble spread (%)



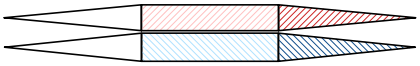
MPI-GE6 vs GISTEMPv4 White Area = 54.1 %



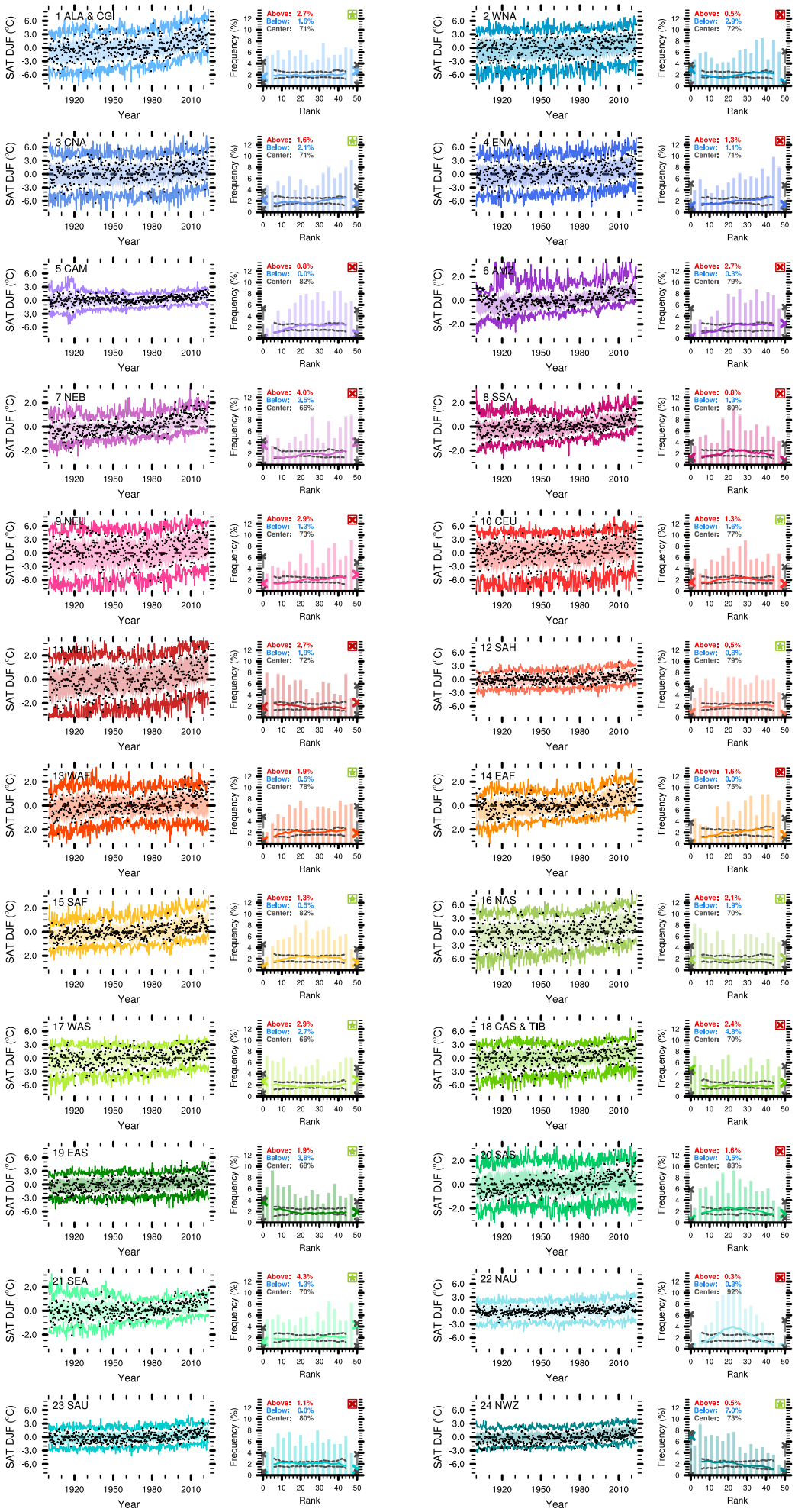
TAS JJA Obs inside central 75th percentile (%)



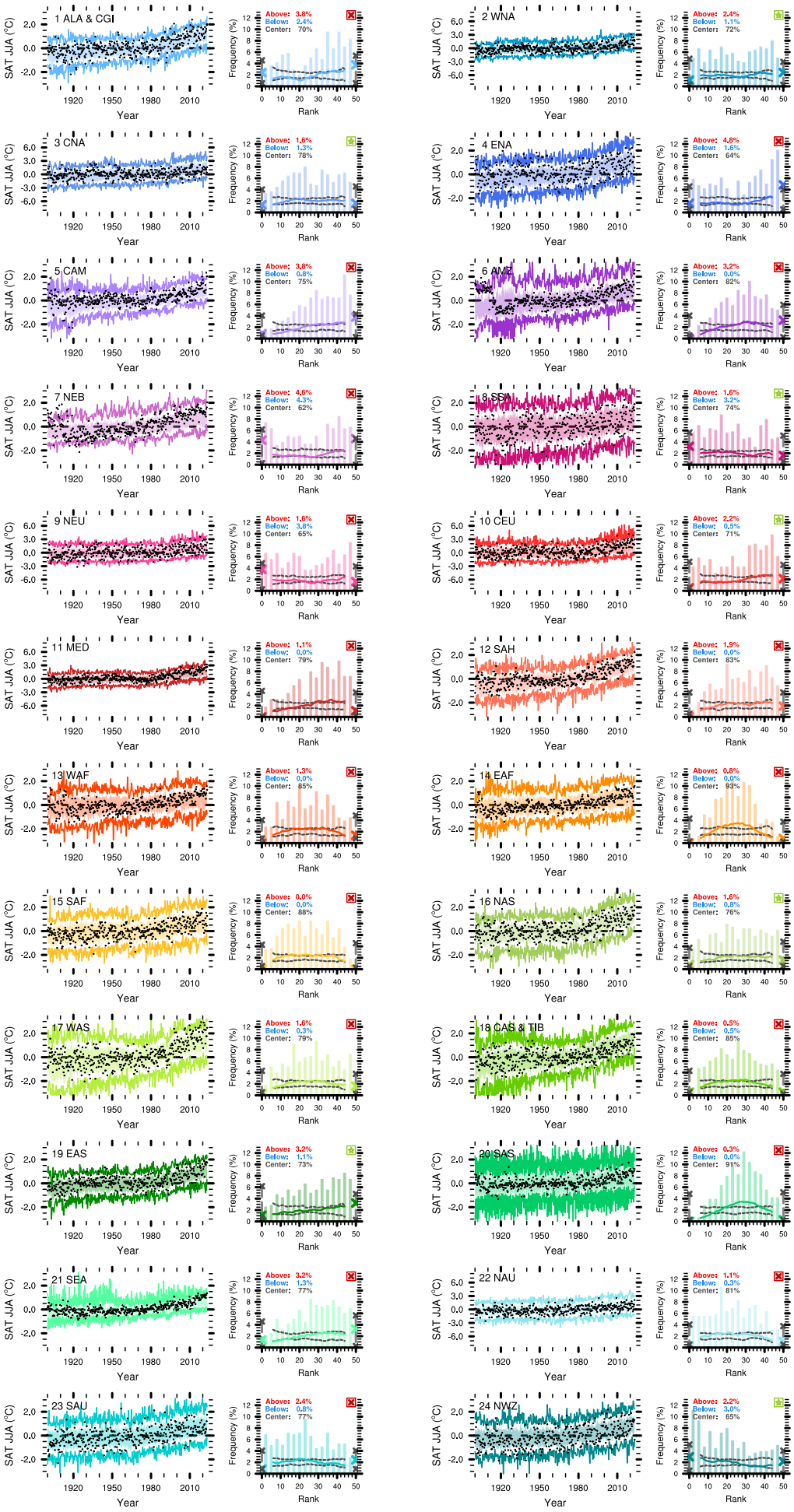
TAS JJA Obs outside ensemble spread (%)



MPI-GE6 vs GISTEMPv4 TAS DJF



MPI-GE6 vs GISTEMPv4 TAS JJA



Detrended Ocean Surface Temperatures

Rank-frequency variability evaluation framework for detrended sea surface temperature (SST) anomalies over ocean grid cells.

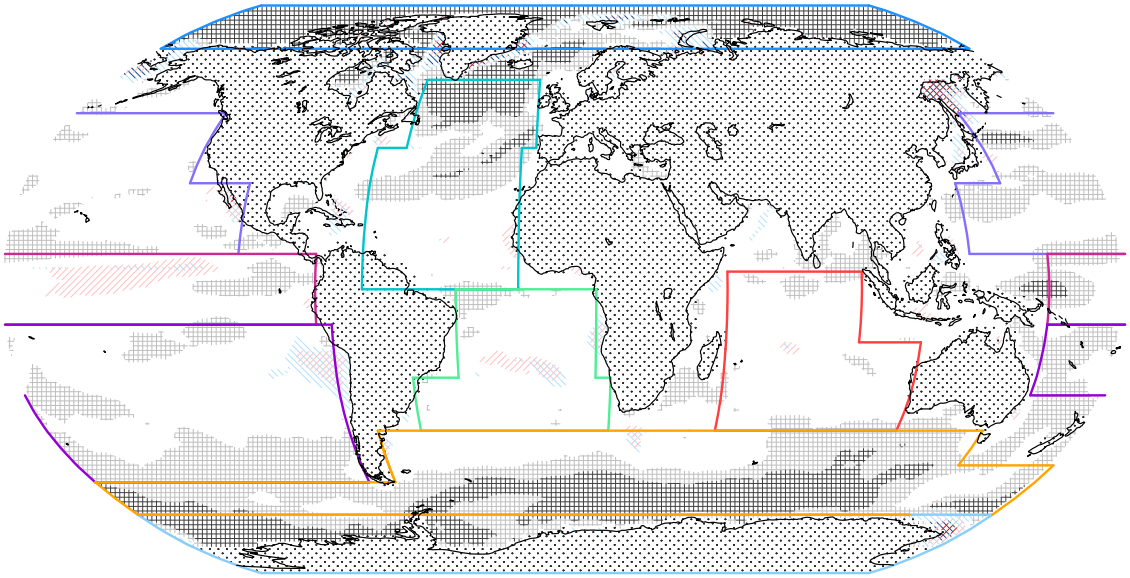
Maps show grid-cell evaluation of the simulated DJF and JJA monthly mean temperature anomalies for the 11 SMILEs included against ERSSTv5 observations globally. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensemble (12.5th to 87.5th percentiles) for more than 80% of months (light grey) or for more than 95% of months (dark grey). Red and blue shading represents where observations are larger than the ensemble maximum (red) or smaller than the ensemble minimum (blue), respectively, for more than 8% of the months (light red and blue) or for more than 20% of the months (dark red and blue). Dotted areas represent land areas or grid cells where observations are missing and are therefore excluded from this analysis. Colored boxes demark the boundaries of each ocean region assessed. The percentage of assessed grid-cells that present none of these biases, unbiased grid-cells, in given at the top (white area). For each assessed region, models are considered to provide an adequate grid-cell representation when they exhibit more than 50% of unbiased grid cells in the region, fulfilling Criteria 2 of our evaluation framework.

Time series and rank frequency histograms show spatially aggregated DJF and JJA SST for each ocean region for all 11 SMILEs. Time series show the ensemble maximum and minimum (color lines) and central 75th percentile ensemble spread (shading) are shown against observations (black dots).

Rank histograms represent the frequency of each place that observations would take in a list of ensemble members ordered by ascending temperature anomaly values. Rank histograms show the observations rank frequency accumulated for 3-rank bins (bars), the running mean rank frequency over a centered n/5 rank window (lines; for 1 to n-1 ranks) and the absolute frequencies of rank 0 and n (crosses), with n the number of ensemble members, for observations (color) and perfect model rank range (gray).

To illustrate how internal variability sampling may affect rank frequencies given the non-infinite record length considered, we include a comparison to the perfect-model rank range, which shows the range of rank frequencies that each ensemble member would take if it were observations (see Methods in main article for further details). If the rank exhibited by observations (colors) is within this perfect-model range (gray; allowing a maximum 10% deviation outside of this range), the rank frequency evaluation shows an adequate model performance, and any deviations from a perfectly frank rank histogram can be assumed to be within the range of deviations that could be caused by internal variability. In the case the observations rank frequency is within the perfect-model range for all rank windows, Criteria 1 of our evaluation framework is met and this is highlighted by a green star at the top right; if not, by a red cross. Percentages at the top left show the frequency of regionally averaged observations occurring above (red) or below (blue) ensemble limits, or clustering within the central 75th percentile range (gray), analogous to the grid-cell evaluation in Criteria 2.

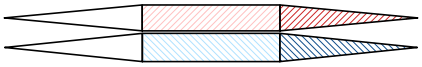
ACCESS vs ERSSTv5 White Area = 59.0 %



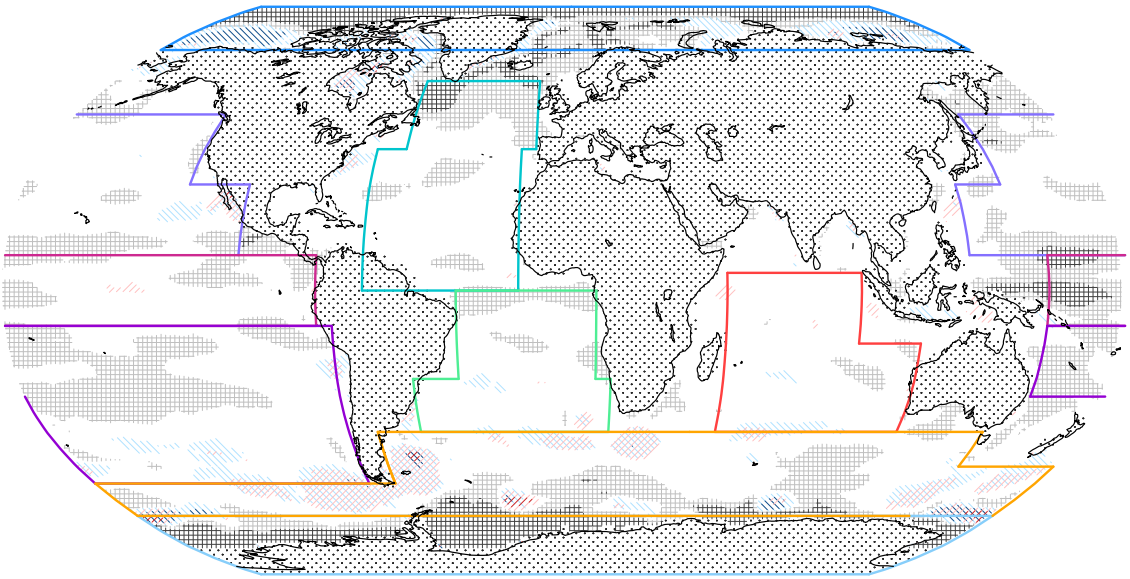
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



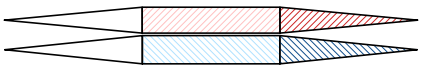
ACCESS vs ERSSTv5 White Area = 62.7 %



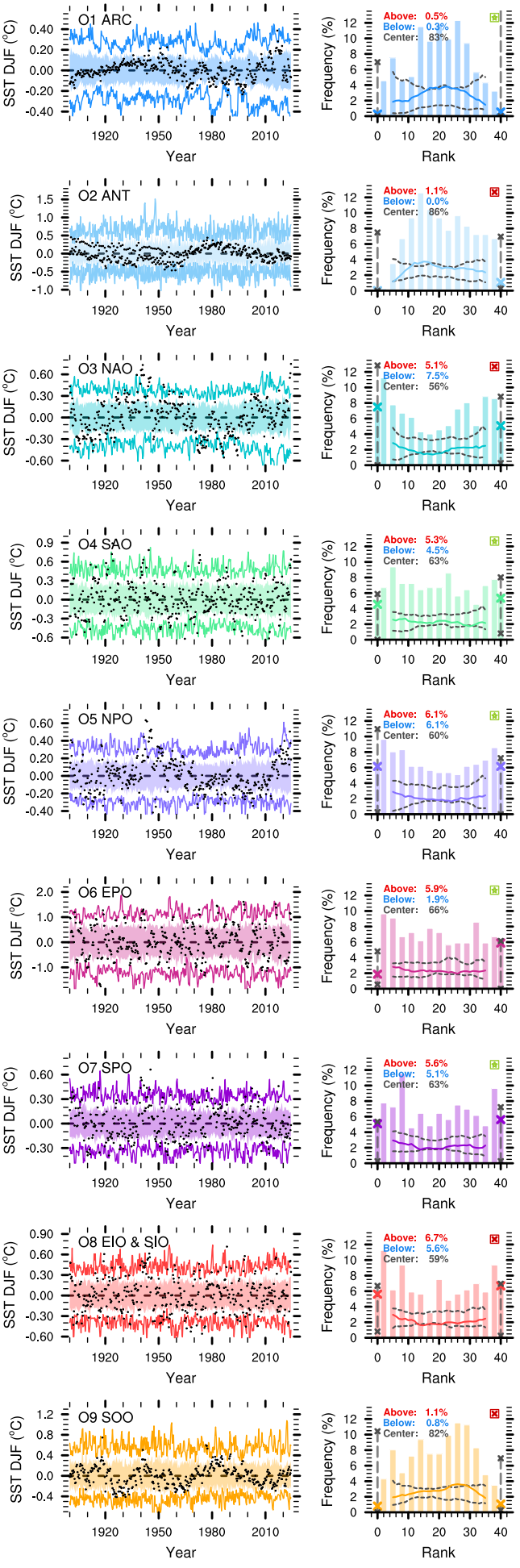
SST JJA Obs inside central 75th percentile (%)



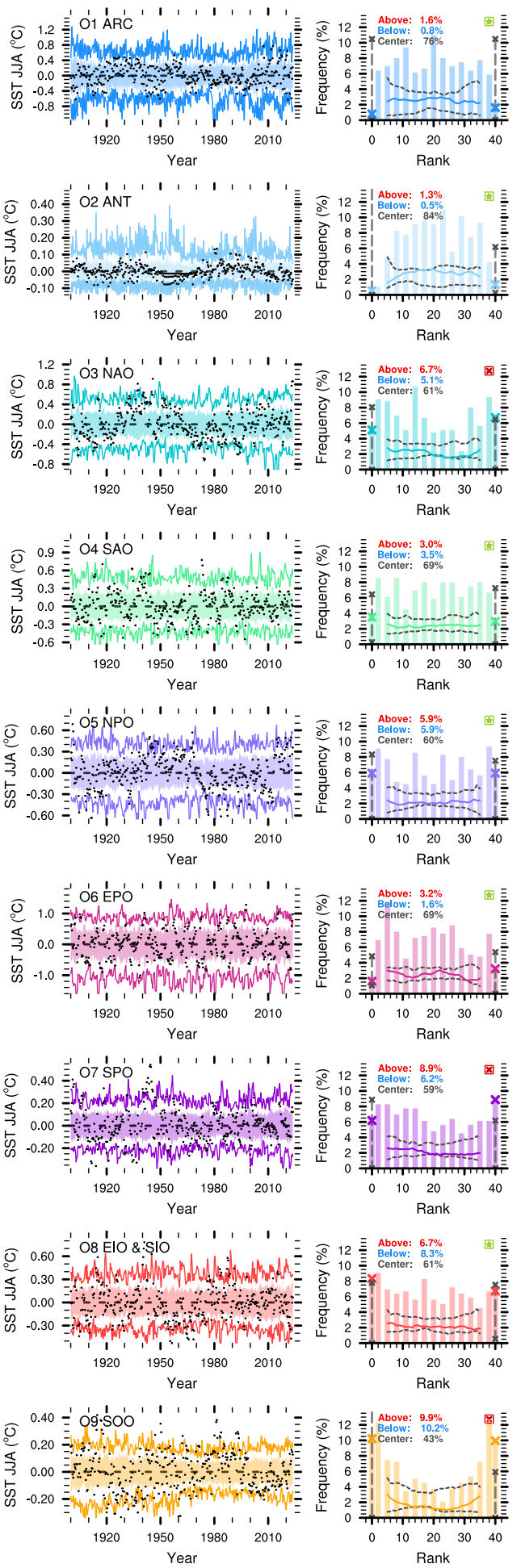
SST JJA Obs outside ensemble spread (%)



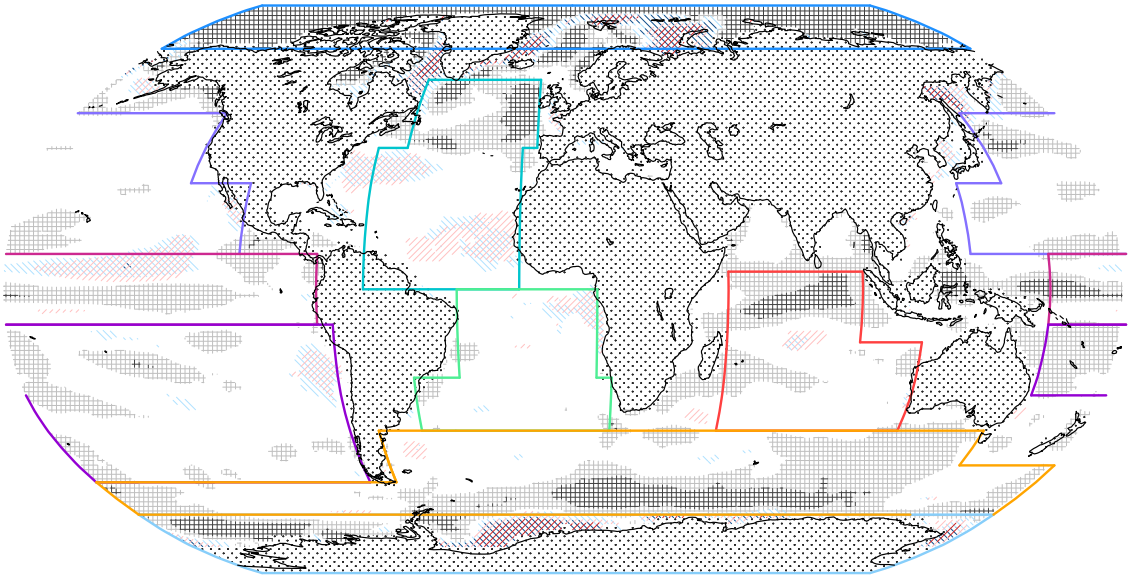
ACCESS vs ERSSTv5 SST DJF



ACCESS vs ERSSTv5 SST JJA



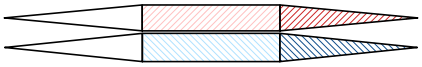
CanESM2 vs ERSSTv5 White Area = 57.0 %



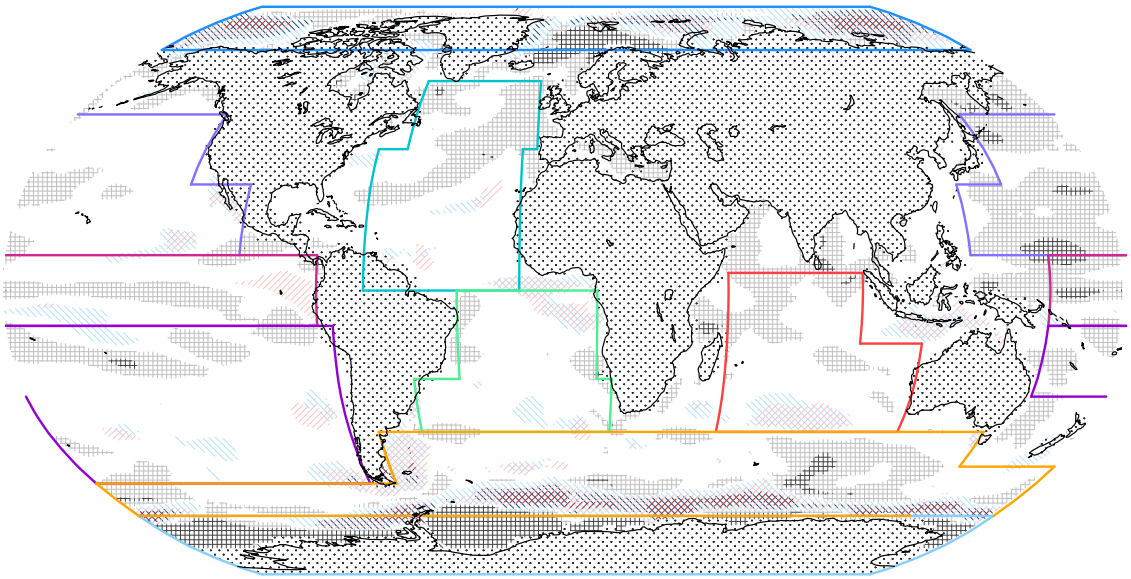
SST DJF Obs inside central 75th percentile (%)



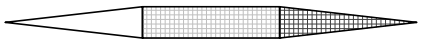
SST DJF Obs outside ensemble spread (%)



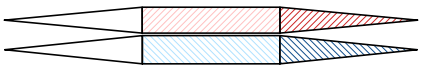
CanESM2 vs ERSSTv5 White Area = 56.2 %



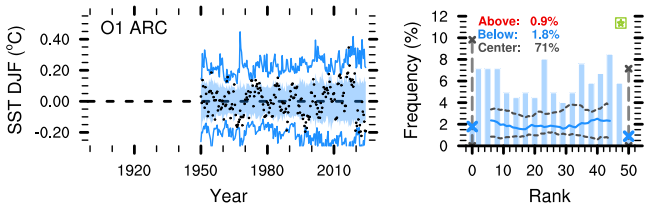
SST JJA Obs inside central 75th percentile (%)

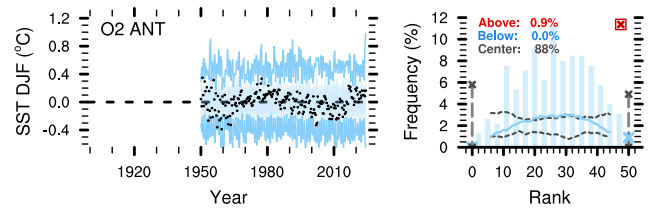


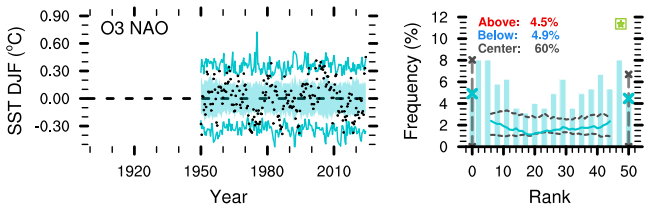
SST JJA Obs outside ensemble spread (%)

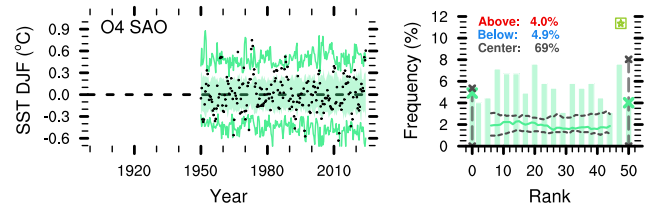


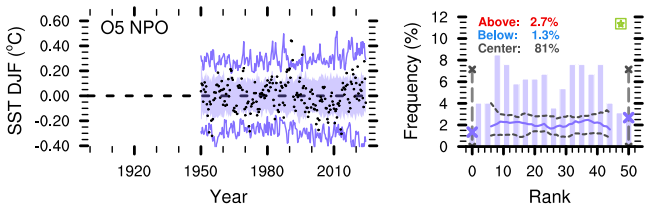
CanESM2 vs ERSSTv5 SST DJF

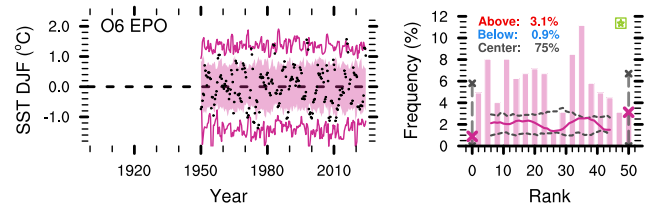


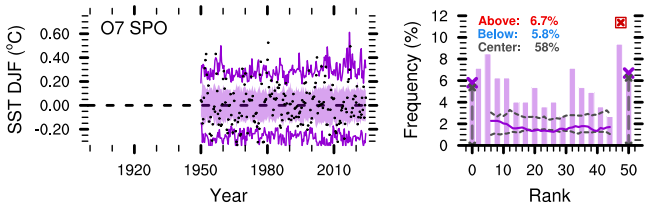


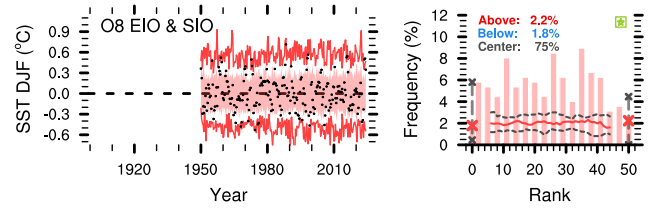


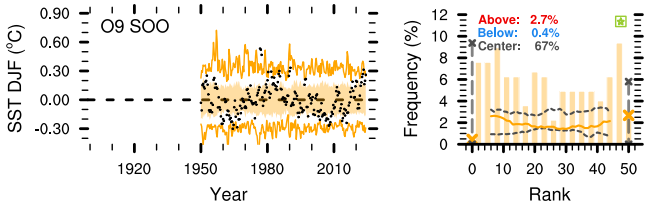




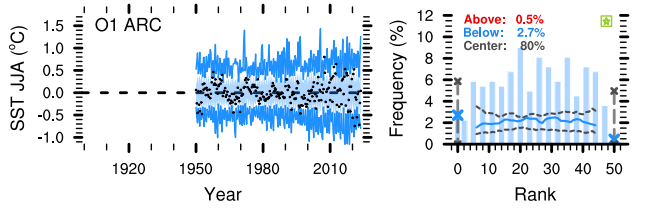


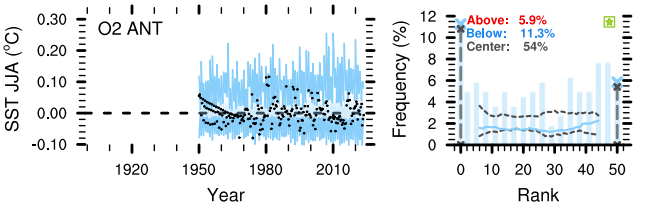


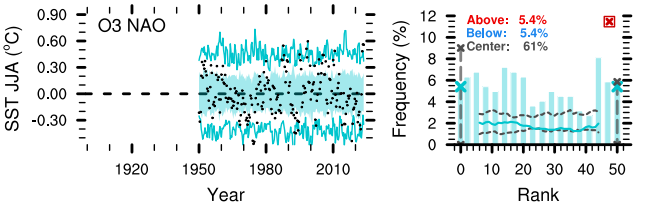


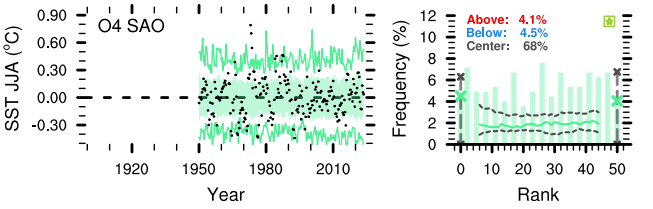


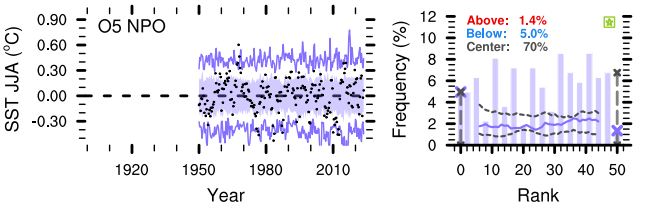
CanESM2 vs ERSSTv5 SST JJA

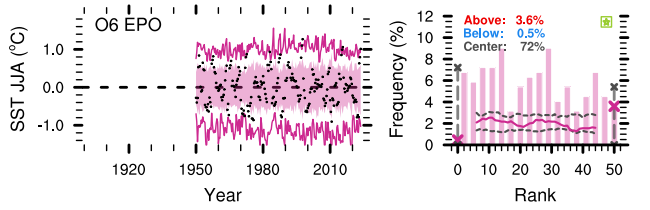


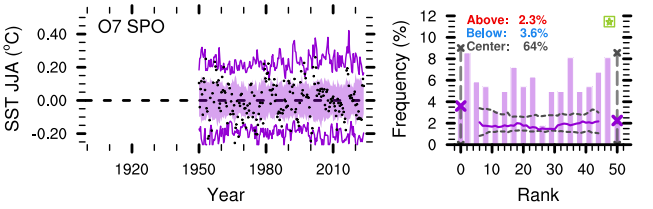


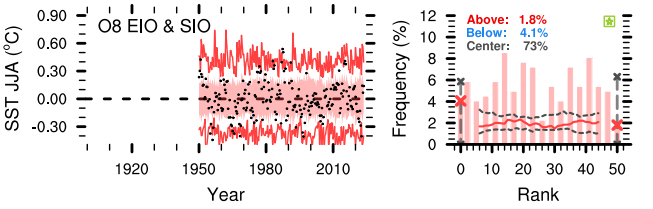


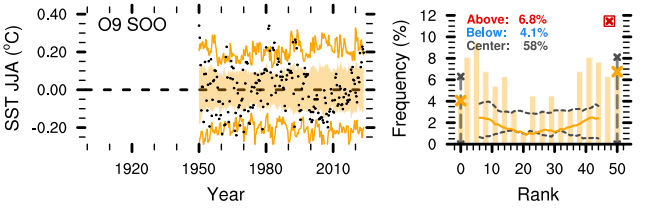




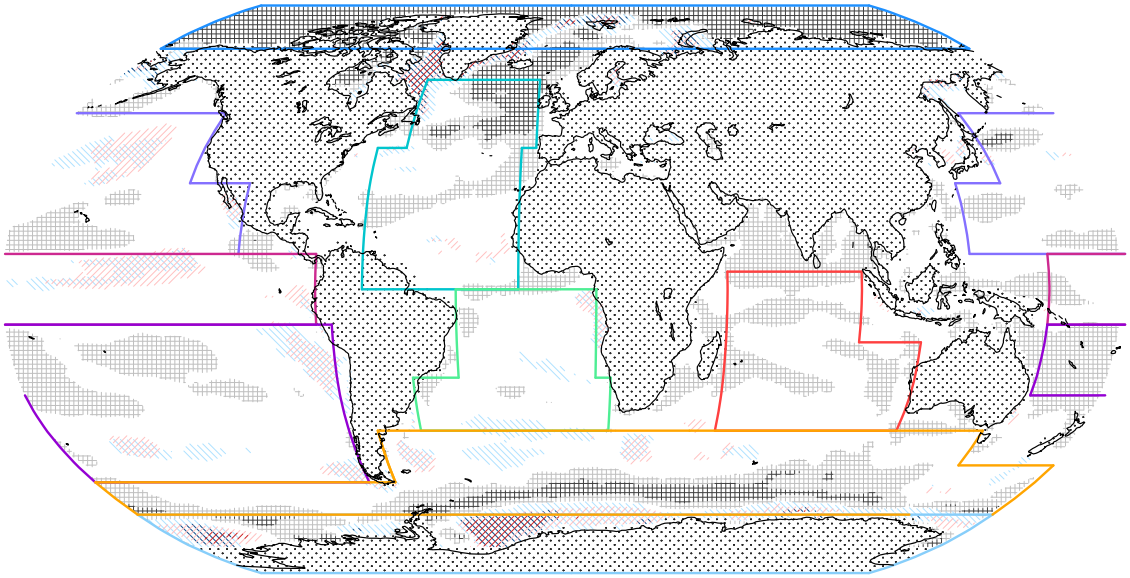








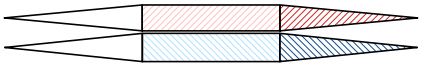
CanESM5 vs ERSSTv5 White Area = 58.4 %



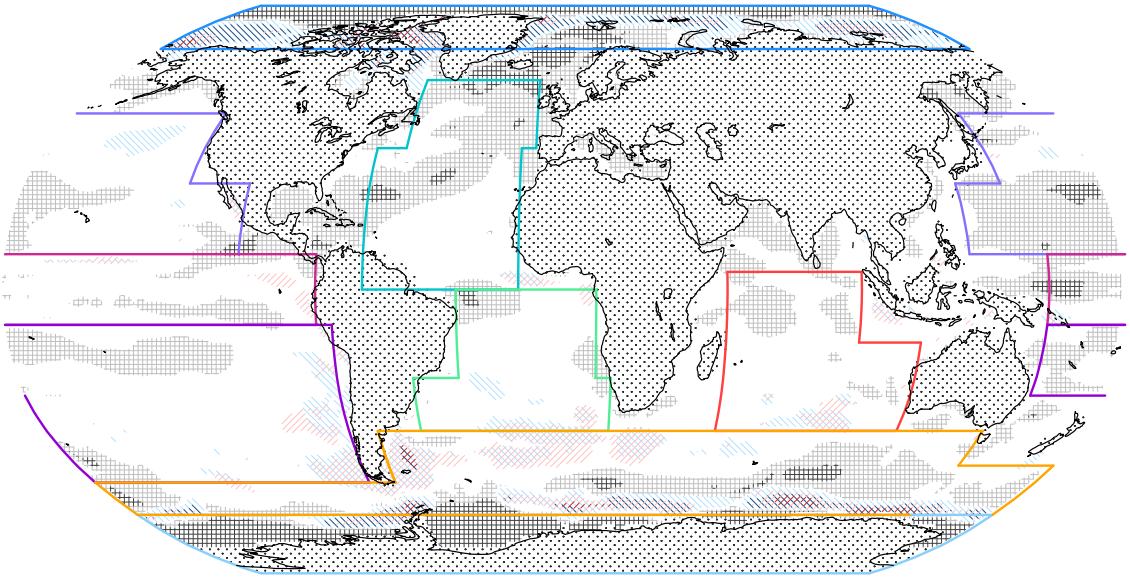
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



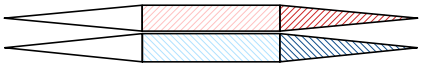
CanESM5 vs ERSSTv5 White Area = 56.2 %



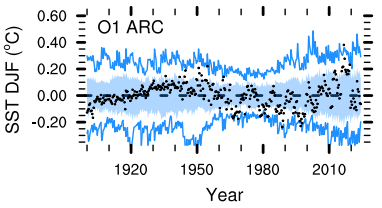
SST JJA Obs inside central 75th percentile (%)

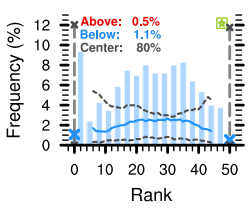


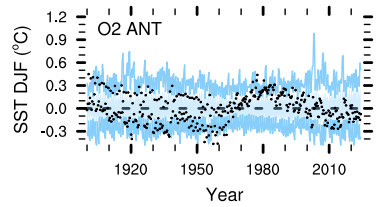
SST JJA Obs outside ensemble spread (%)

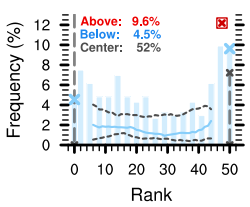


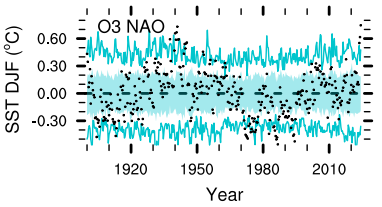
CanESM5 vs ERSSTv5 SST DJF

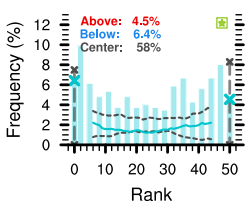


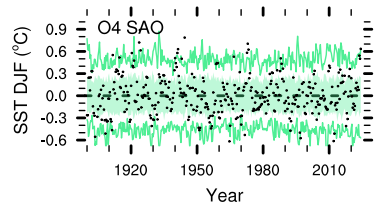


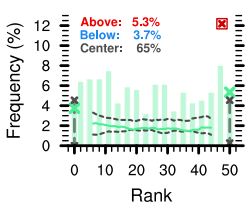


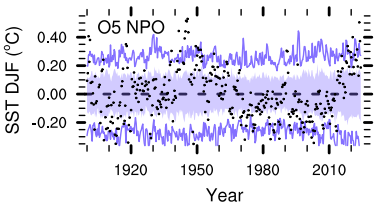


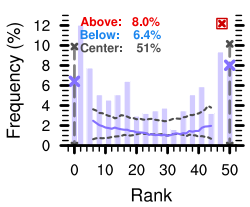


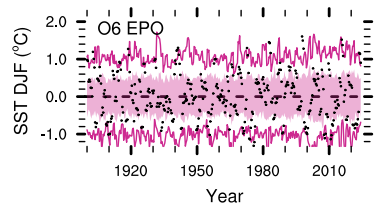


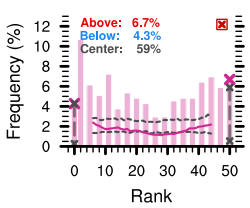


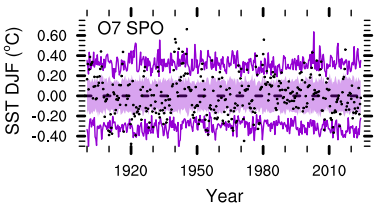


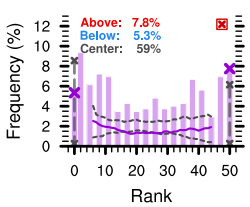


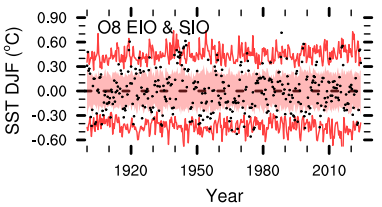


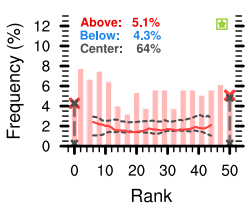


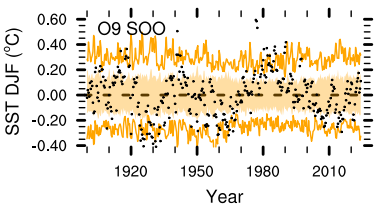


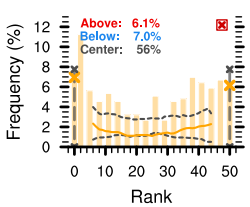




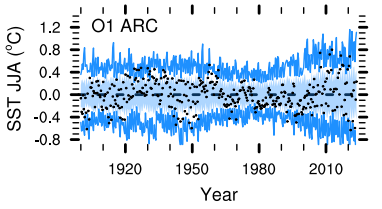


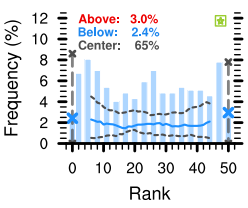


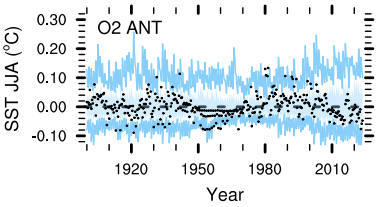


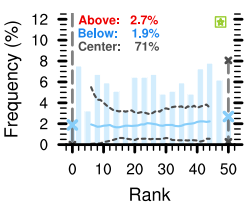


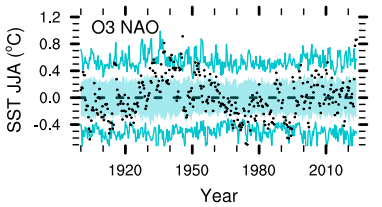
CanESM5 vs ERSSTv5 SST JJA

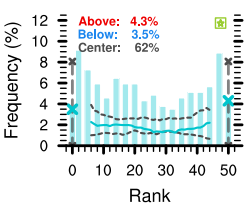


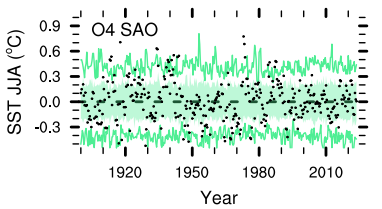


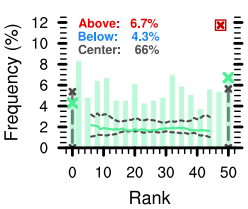


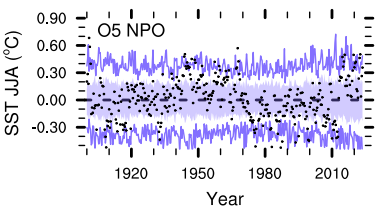


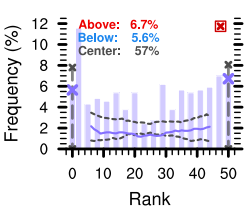


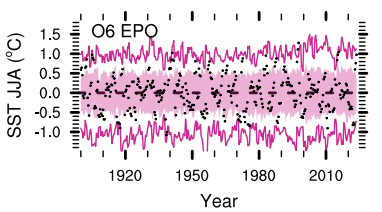


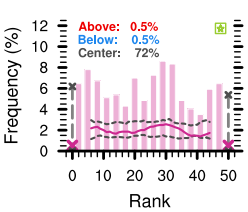


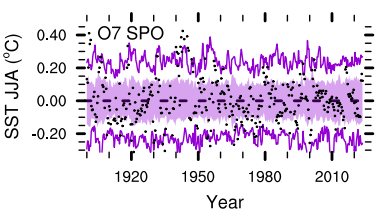


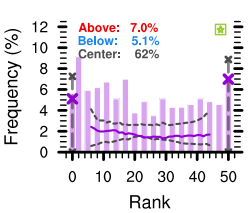


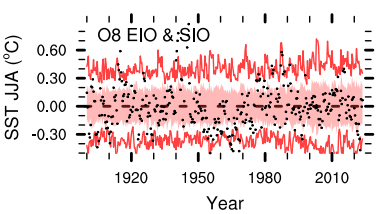


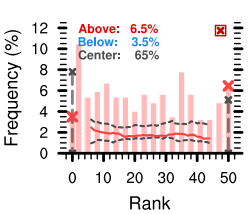


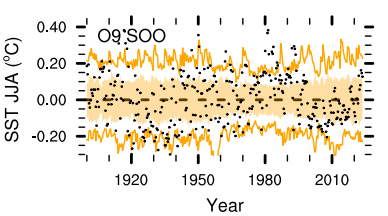


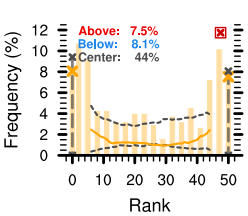




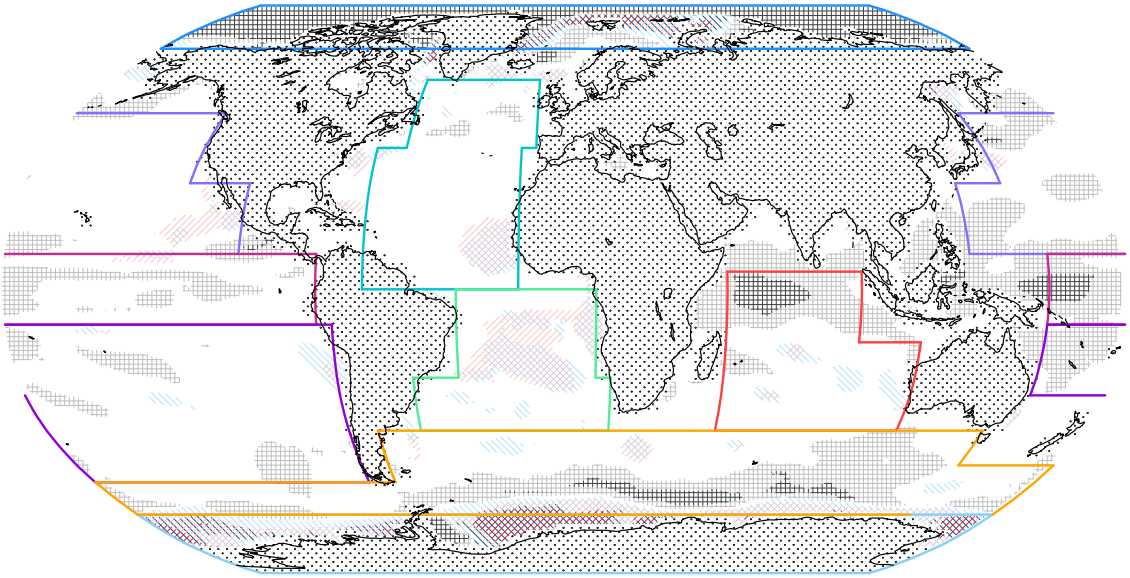








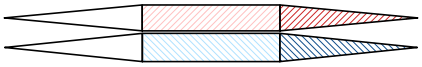
CESM-LE vs ERSSTv5 White Area = 62.8 %



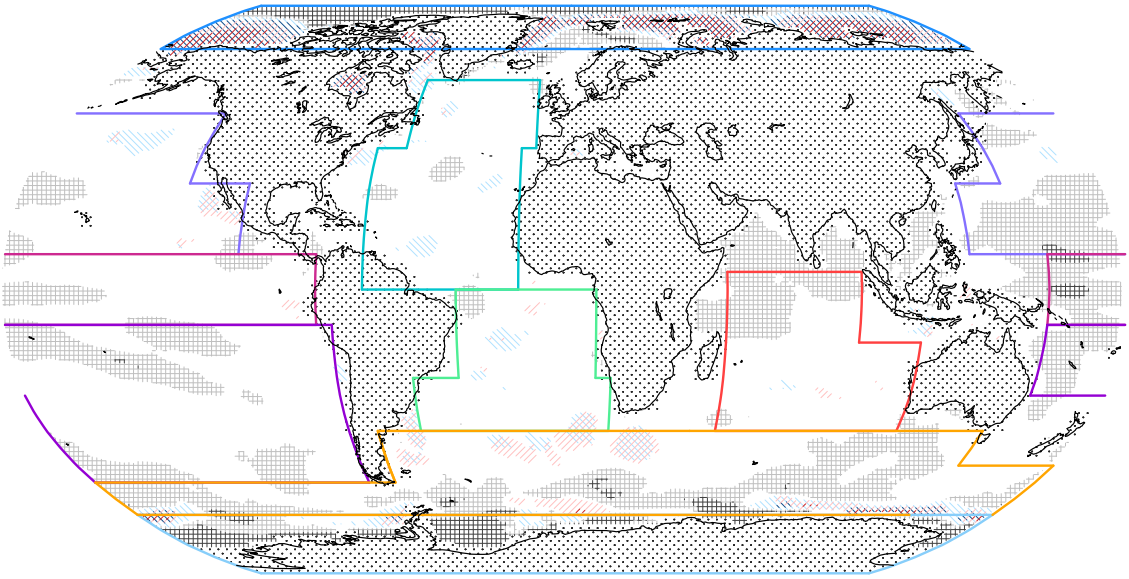
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



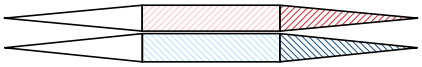
CESM-LE vs ERSSTv5 White Area = 65.9 %



SST JJA Obs inside central 75th percentile (%)

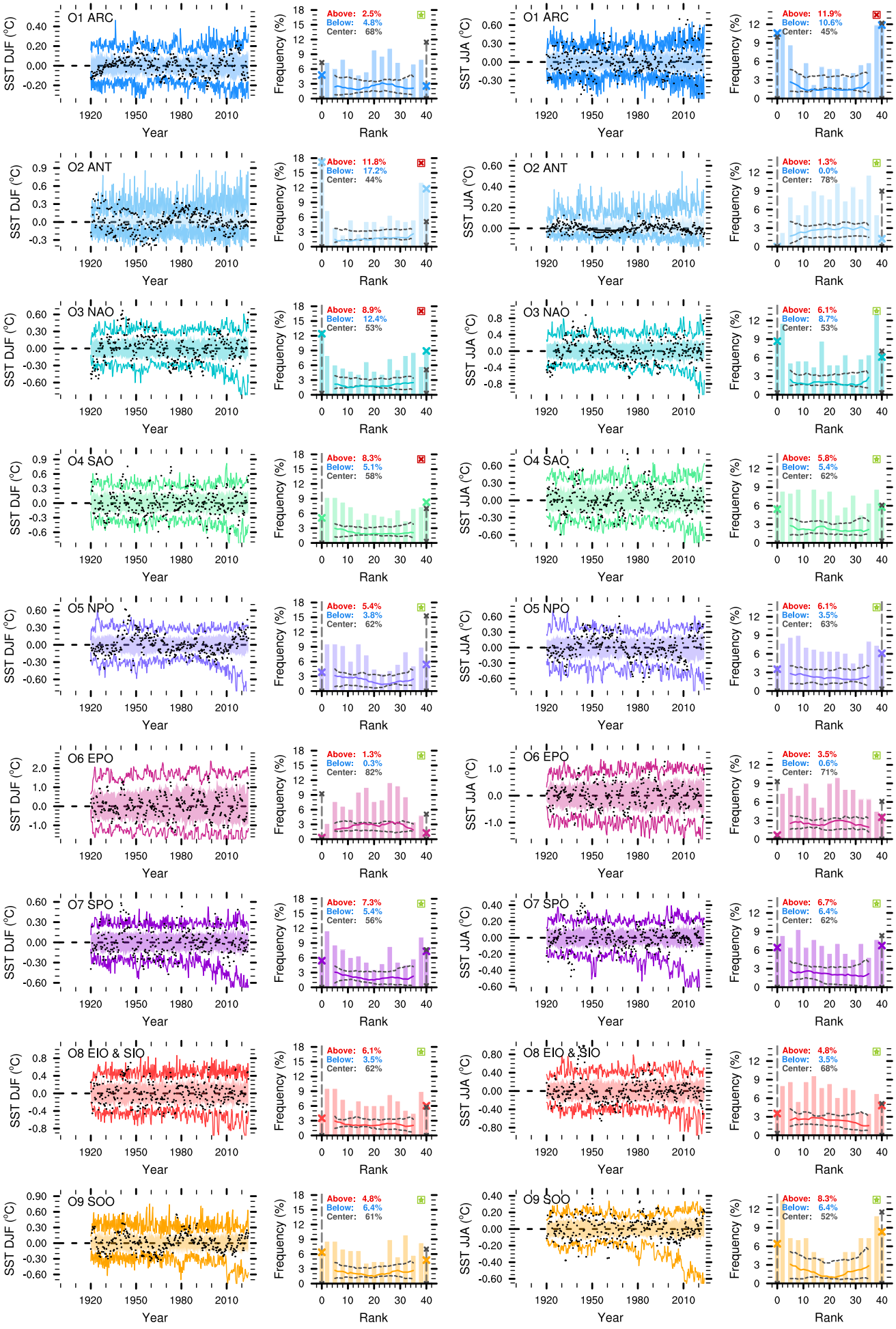


SST JJA Obs outside ensemble spread (%)

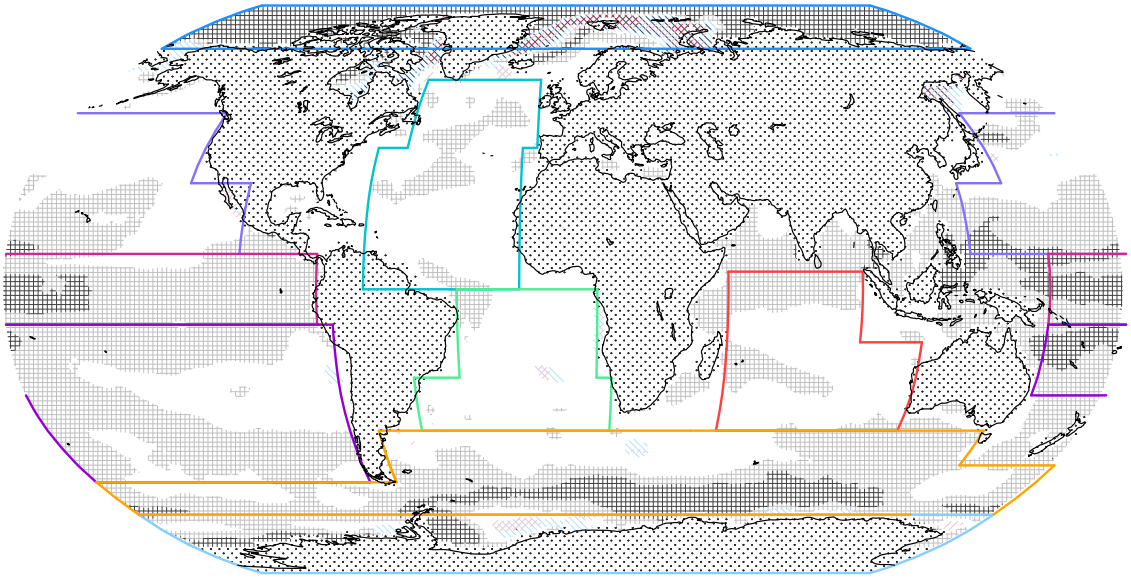


CESM-LE vs ERSSTv5 SST DJF

CESM-LE vs ERSSTv5 SST JJA



CESM2-LE vs ERSSTv5 White Area = 47.0 %



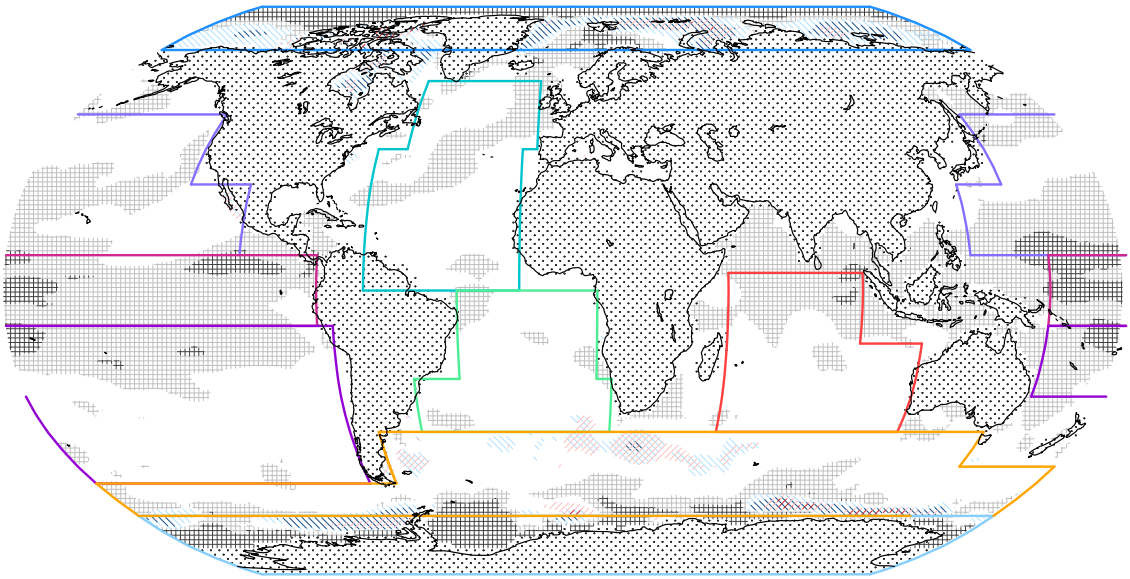
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

CESM2-LE vs ERSSTv5 White Area = 53.2 %



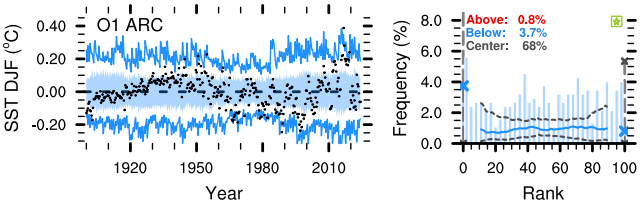
SST JJA Obs inside central 75th percentile (%)

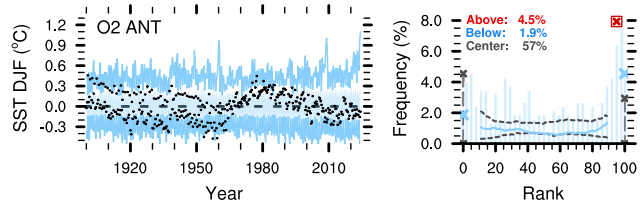
80 95

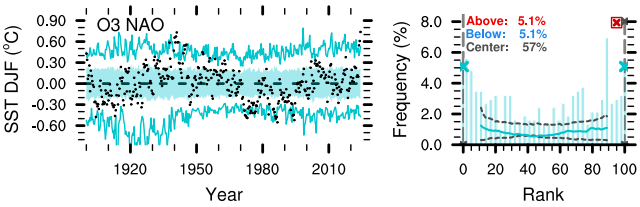
SST JJA Obs outside ensemble spread (%)

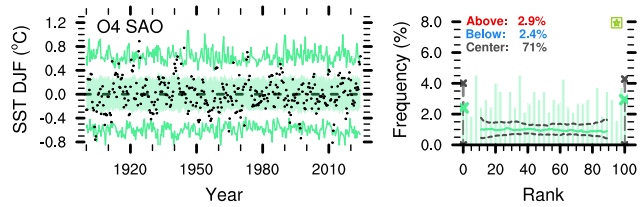
8 20

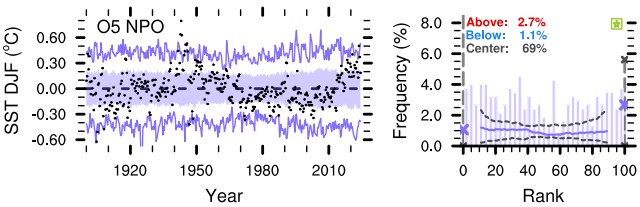
CESM2-LE vs ERSSTv5 SST DJF

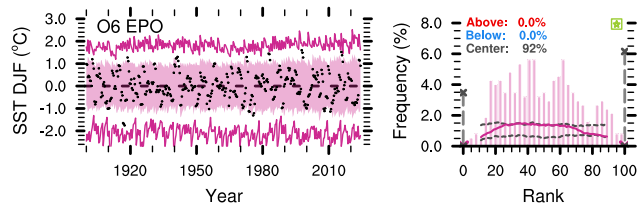


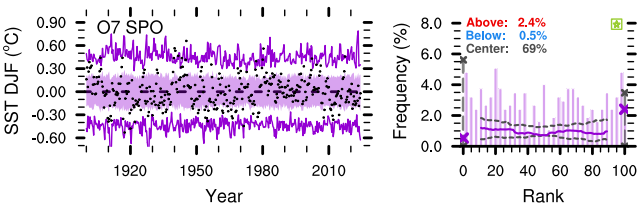


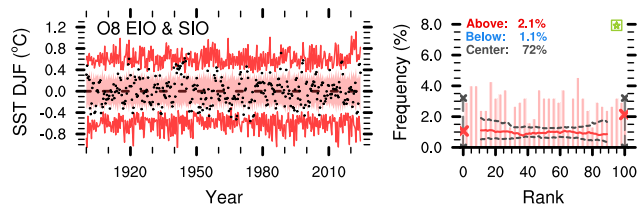


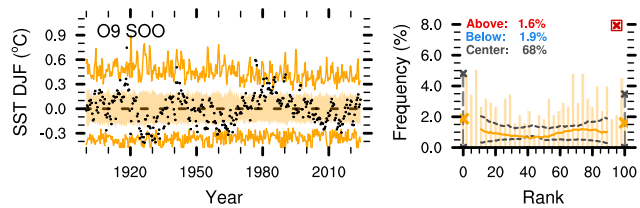




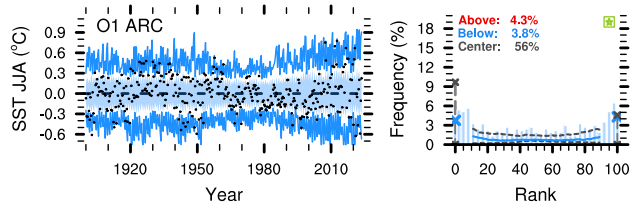


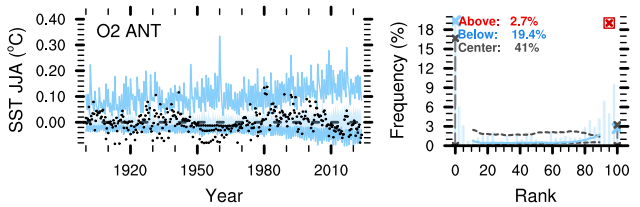


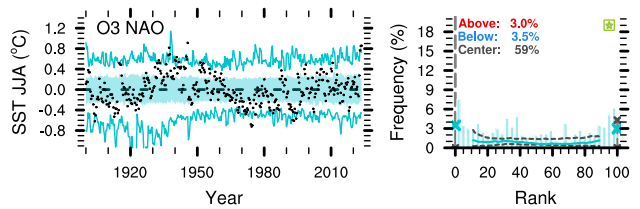


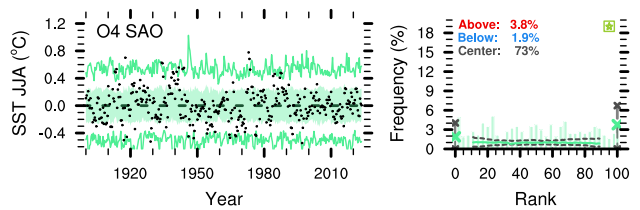


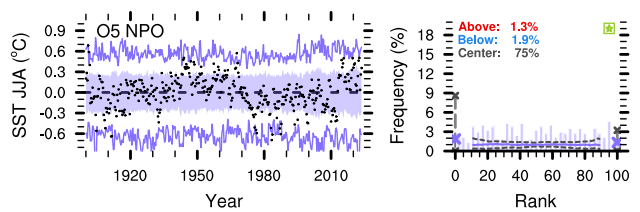
CESM2-LE vs ERSSTv5 SST JJA

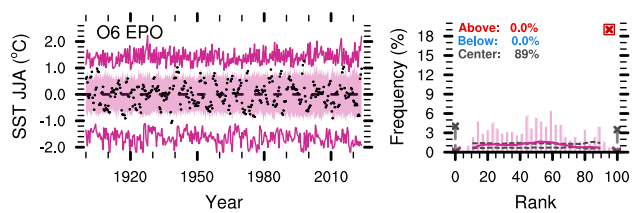


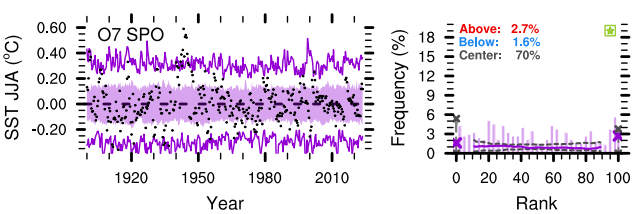


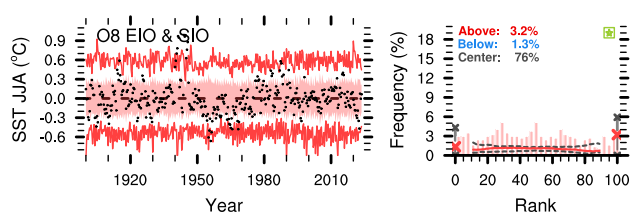


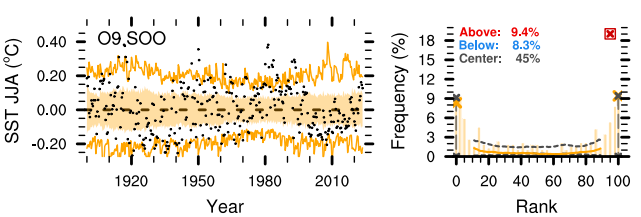




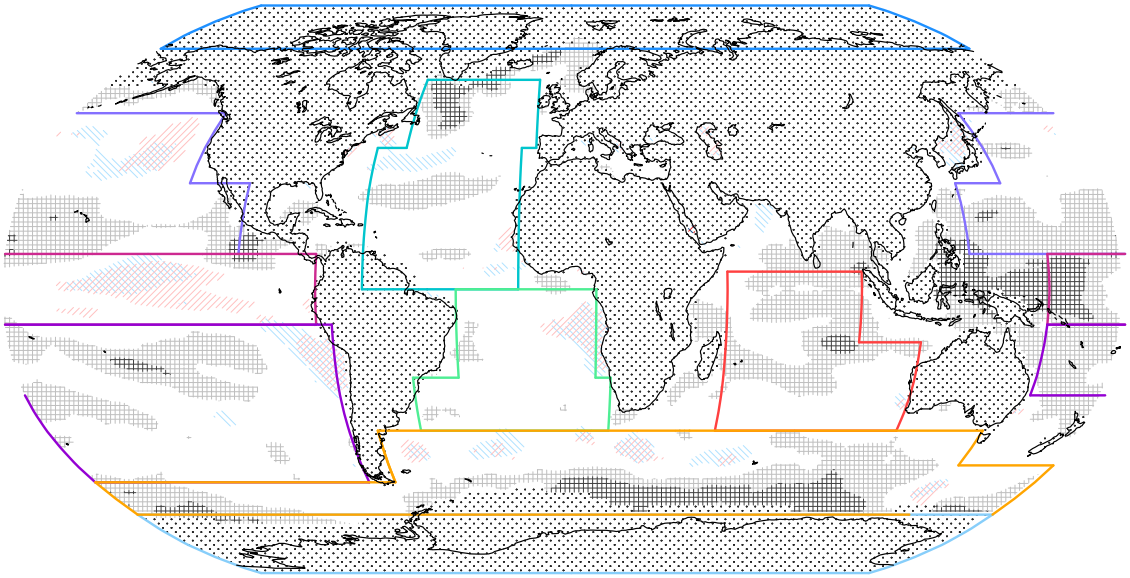








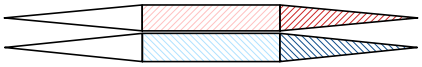
CSIRO-Mk360 vs ERSSTv5 White Area = 57.8 %



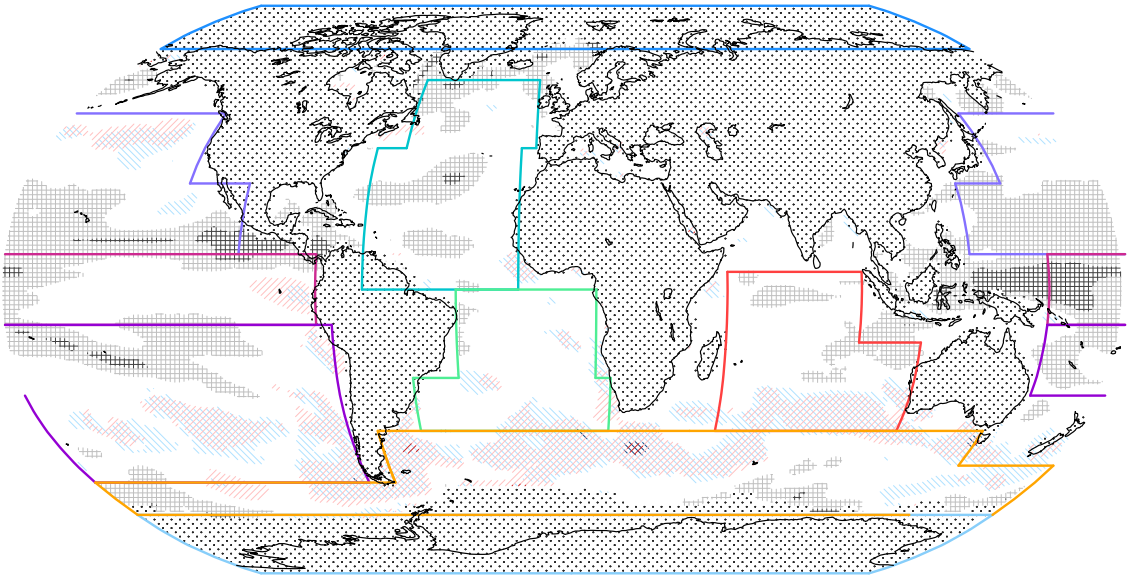
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



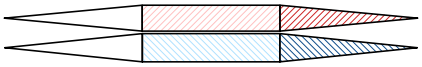
CSIRO-Mk360 vs ERSSTv5 White Area = 55.7 %



SST JJA Obs inside central 75th percentile (%)

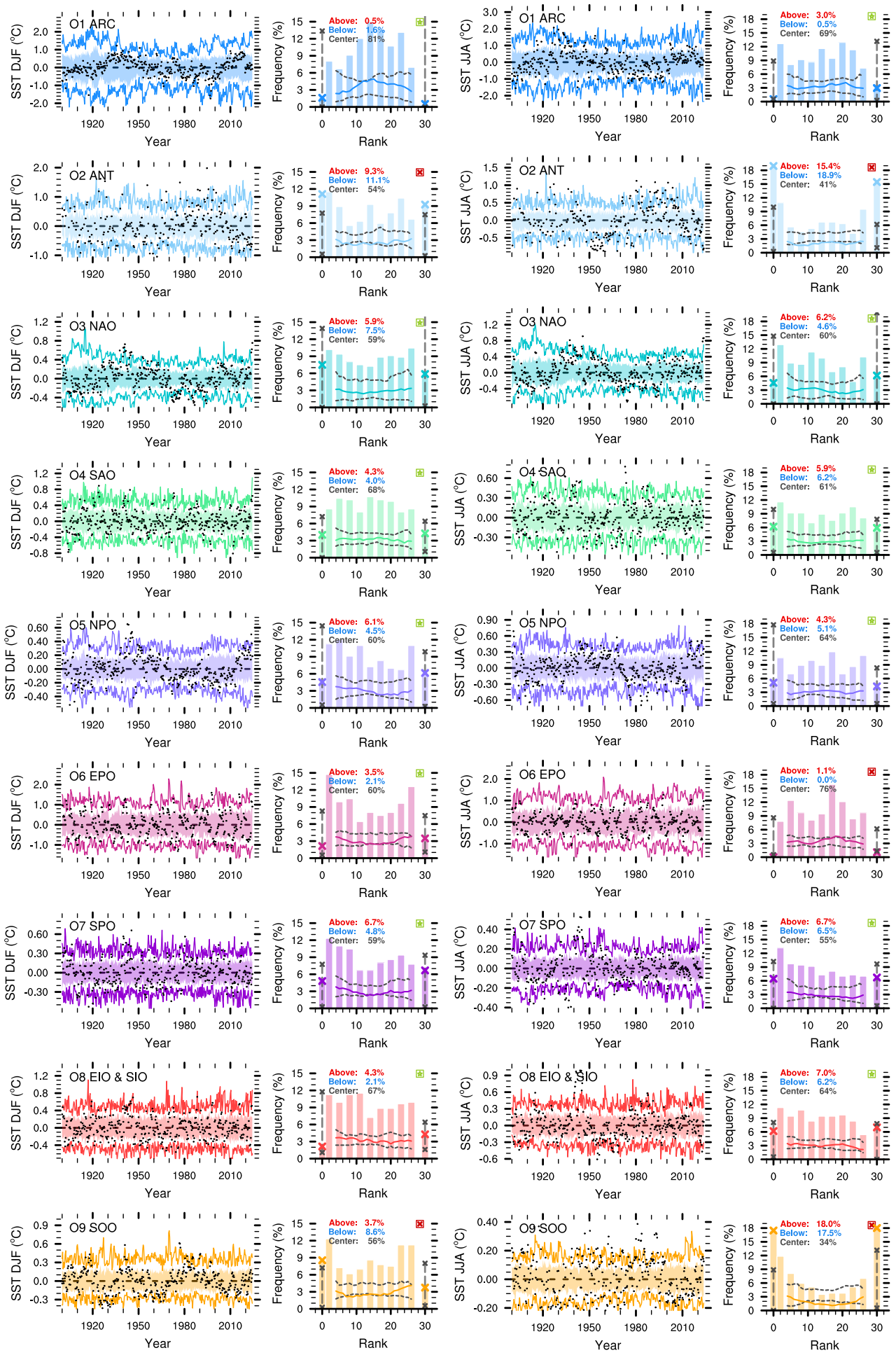


SST JJA Obs outside ensemble spread (%)

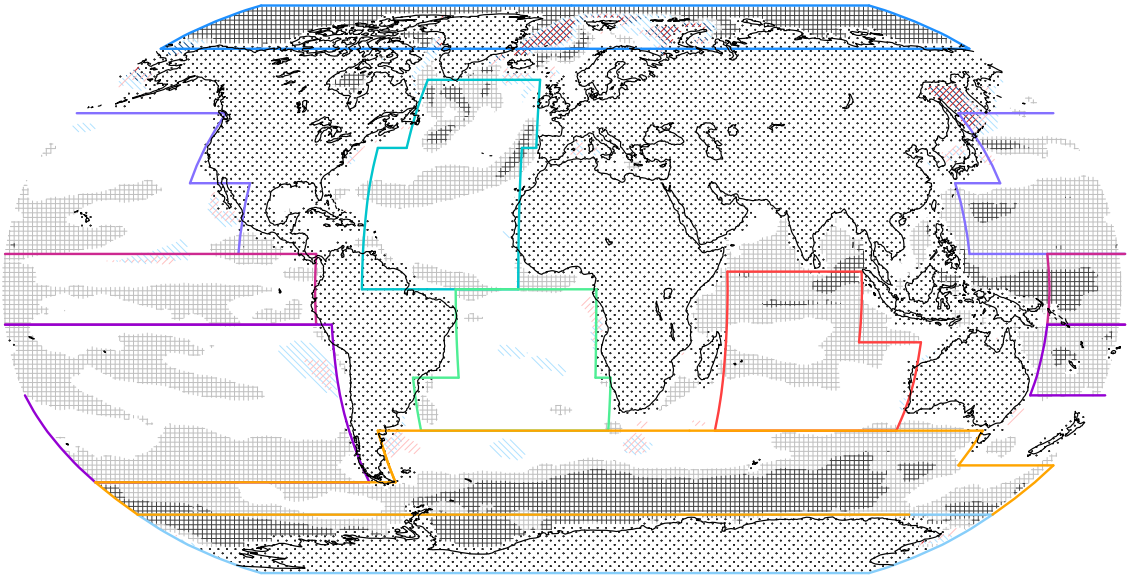


CSIRO-Mk360 vs ERSSTv5 SST DJF

CSIRO-Mk360 vs ERSSTv5 SST JJA



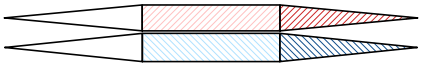
GFDL-ESM2M vs ERSSTv5 White Area = 49.3 %



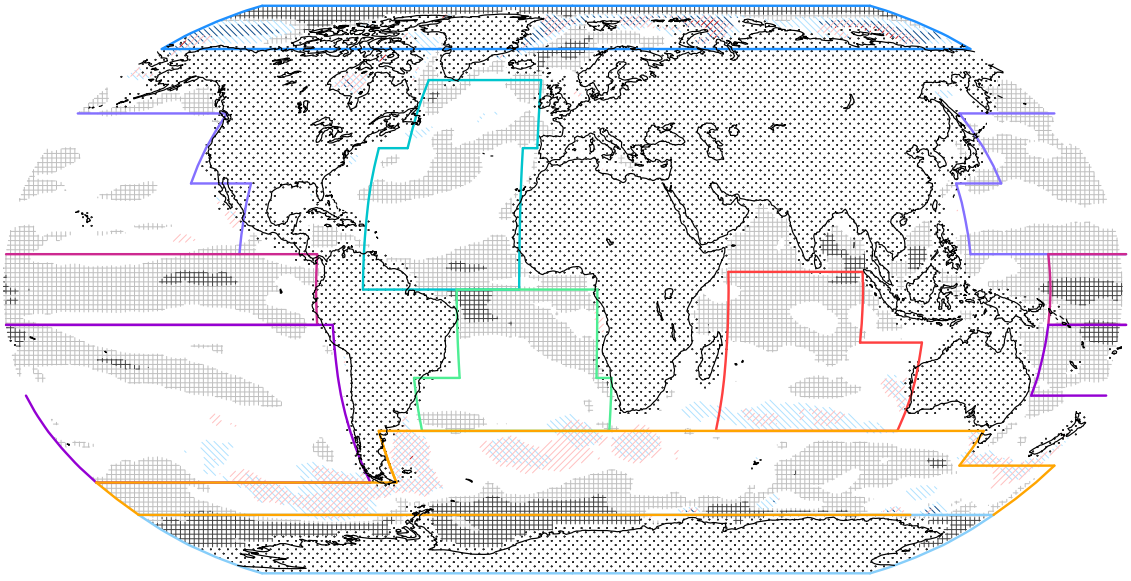
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



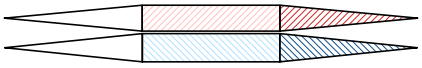
GFDL-ESM2M vs ERSSTv5 White Area = 49.6 %



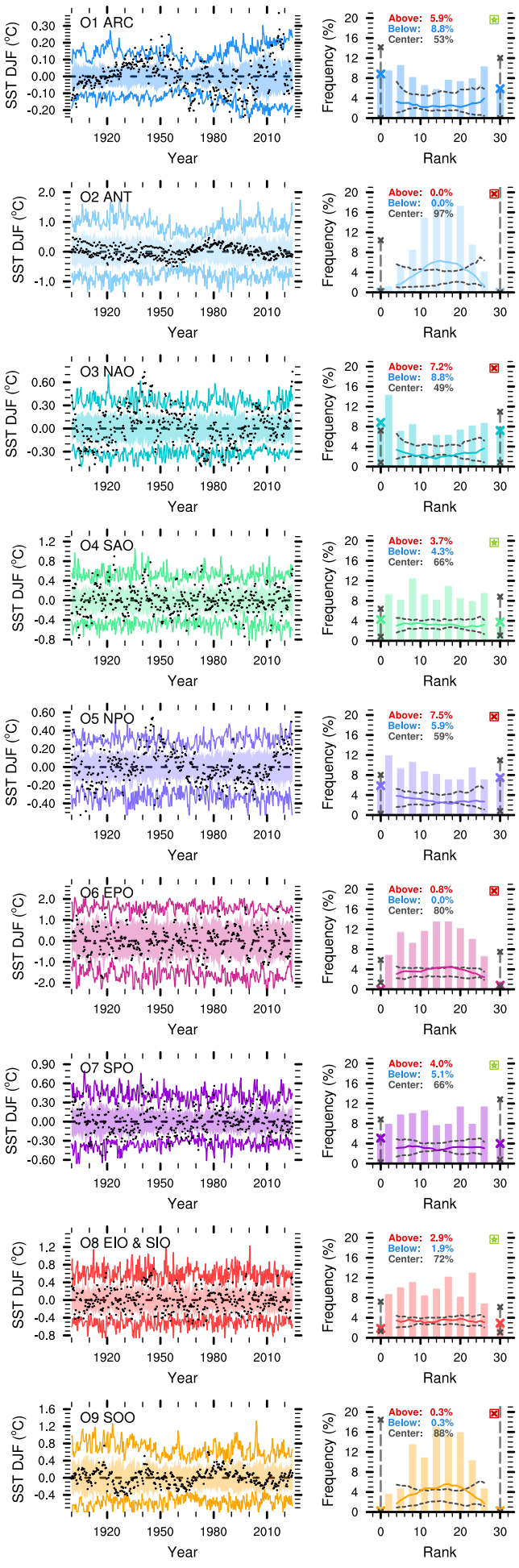
SST JJA Obs inside central 75th percentile (%)



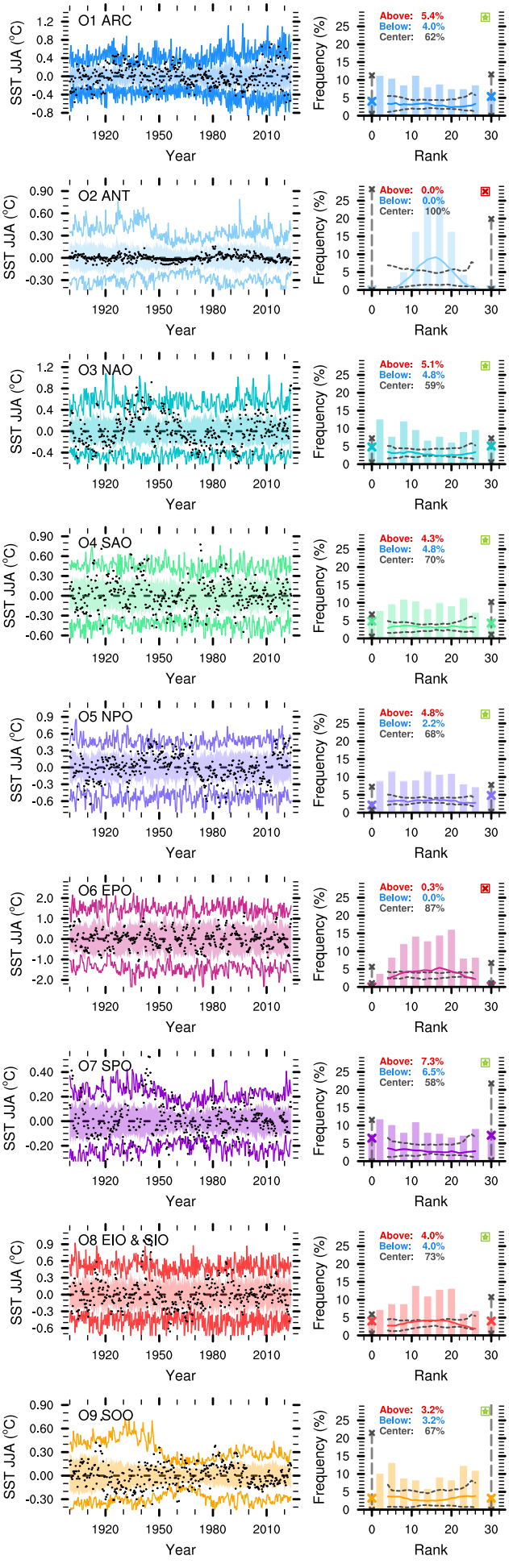
SST JJA Obs outside ensemble spread (%)



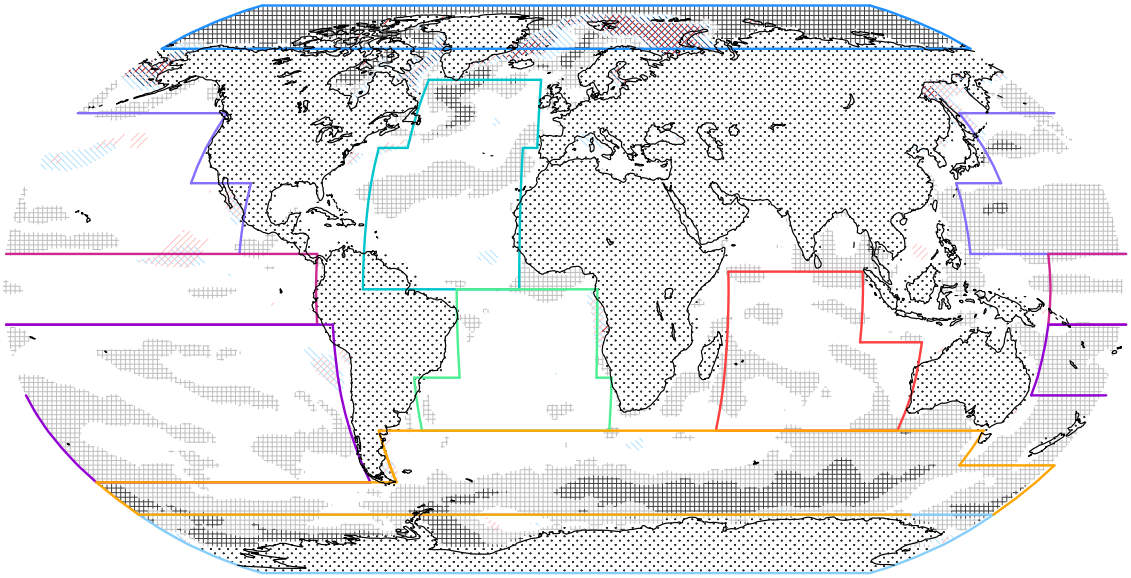
GFDL-ESM2M vs ERSSTv5 SST DJF



GFDL-ESM2M vs ERSSTv5 SST JJA



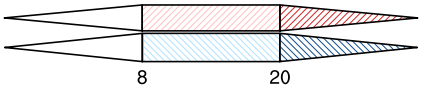
GFDL-SPEAR vs ERSSTv5 White Area = 58.9 %



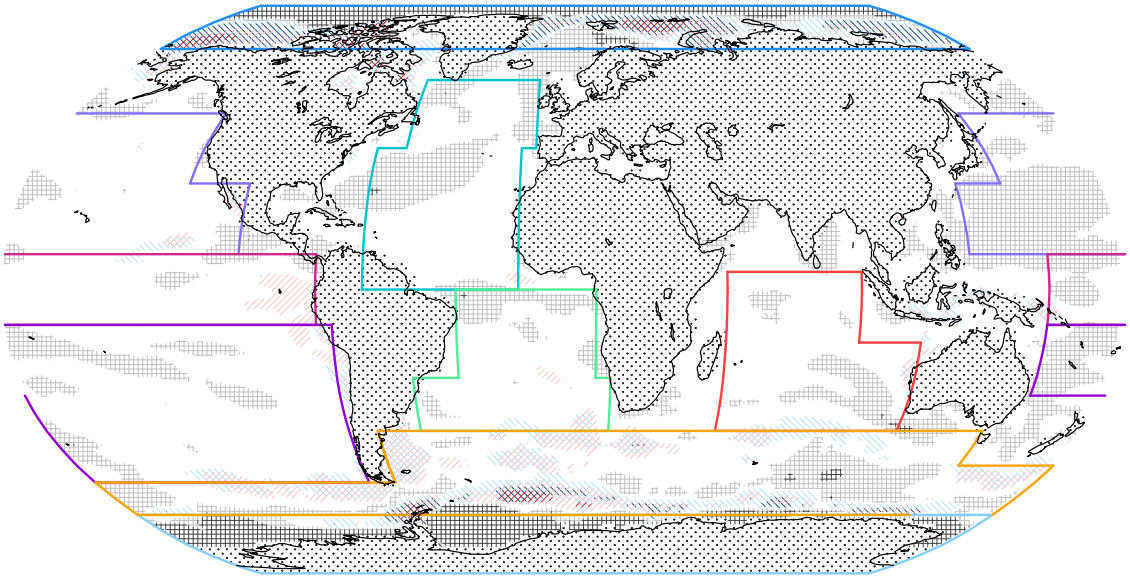
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



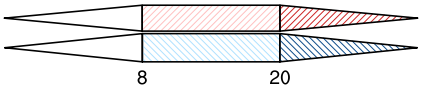
GFDL-SPEAR vs ERSSTv5 White Area = 62.0 %



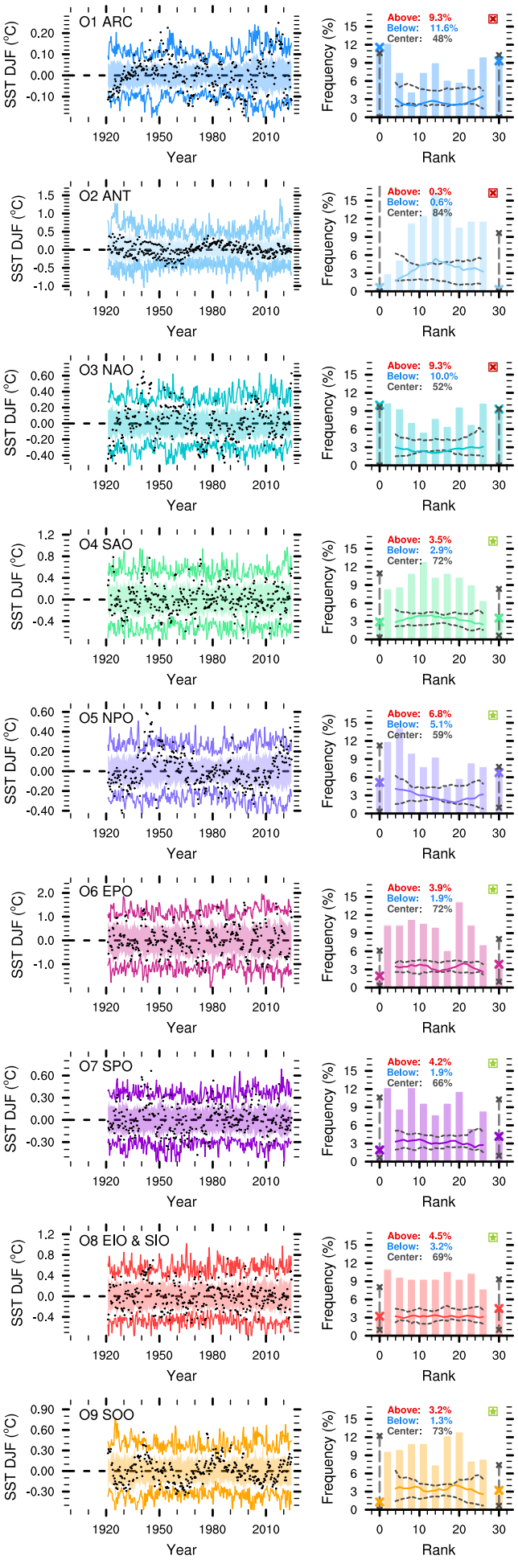
SST JJA Obs inside central 75th percentile (%)



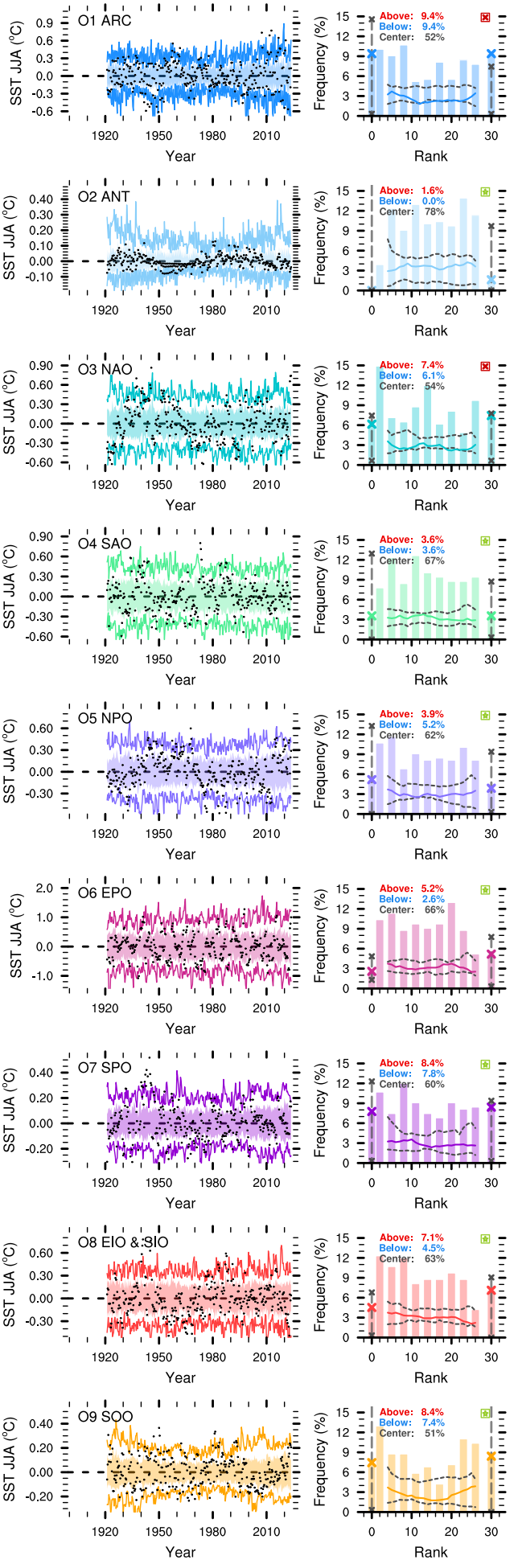
SST JJA Obs outside ensemble spread (%)



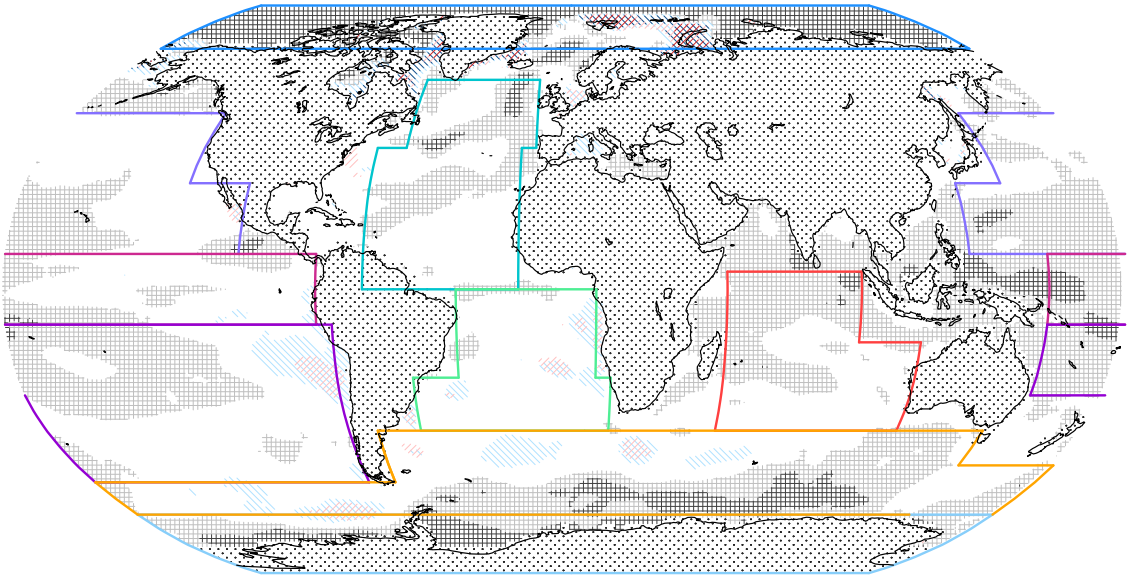
GFDL-SPEAR vs ERSSTv5 SST DJF



GFDL-SPEAR vs ERSSTv5 SST JJA



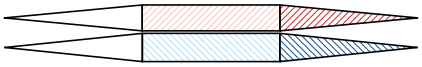
MIROC6 vs ERSSTv5 White Area = 49.7 %



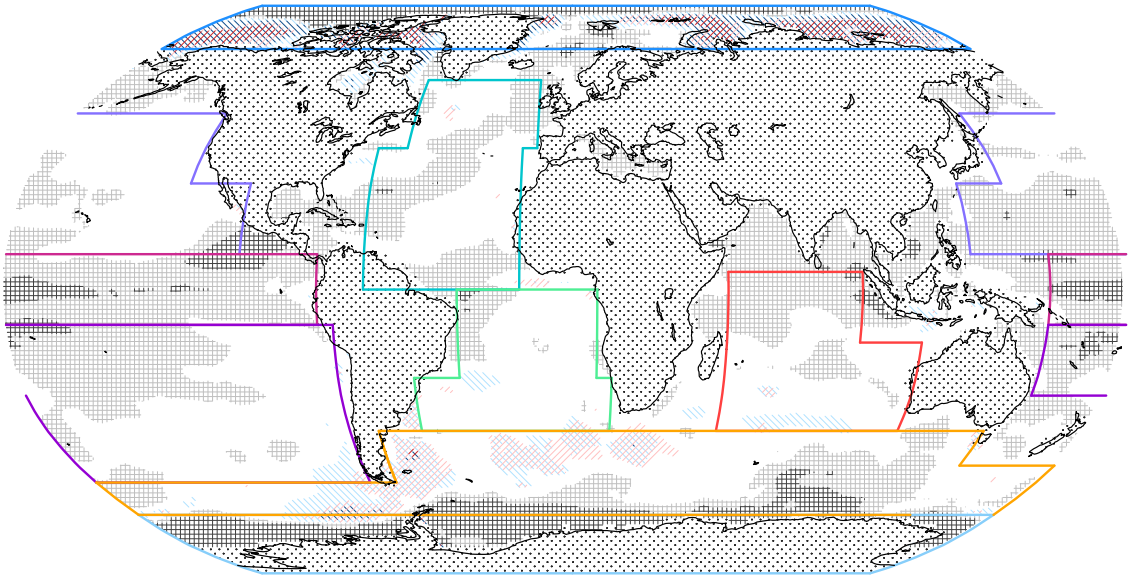
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



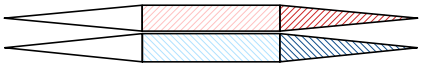
MIROC6 vs ERSSTv5 White Area = 46.4 %



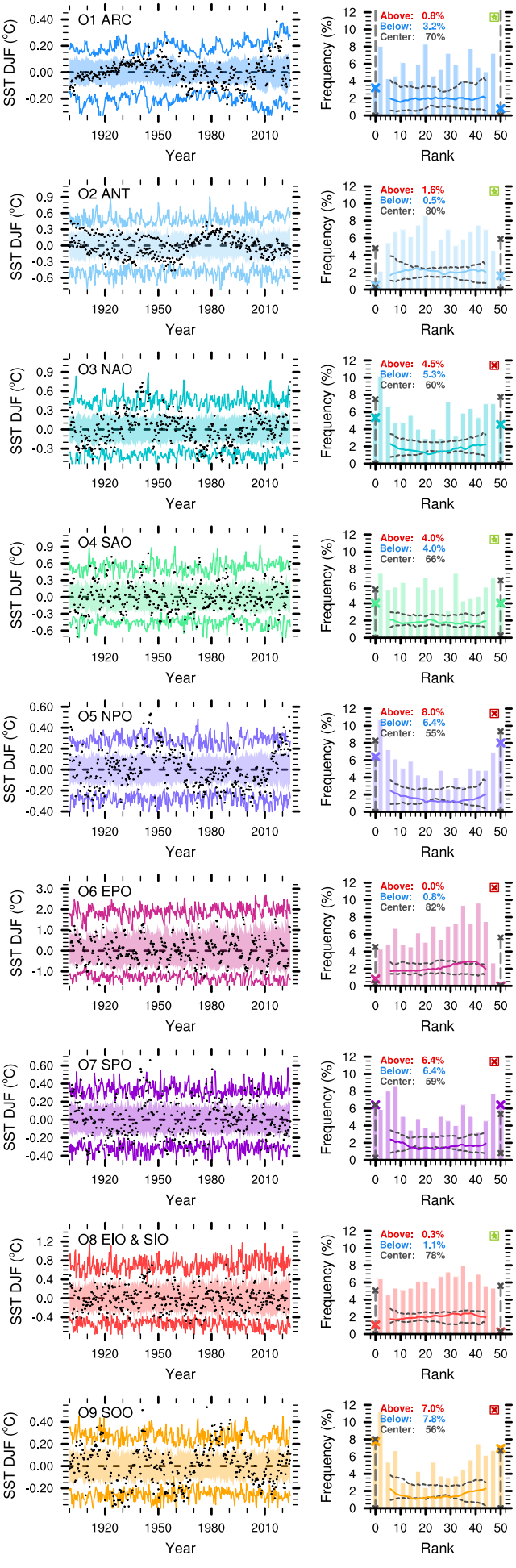
SST JJA Obs inside central 75th percentile (%)



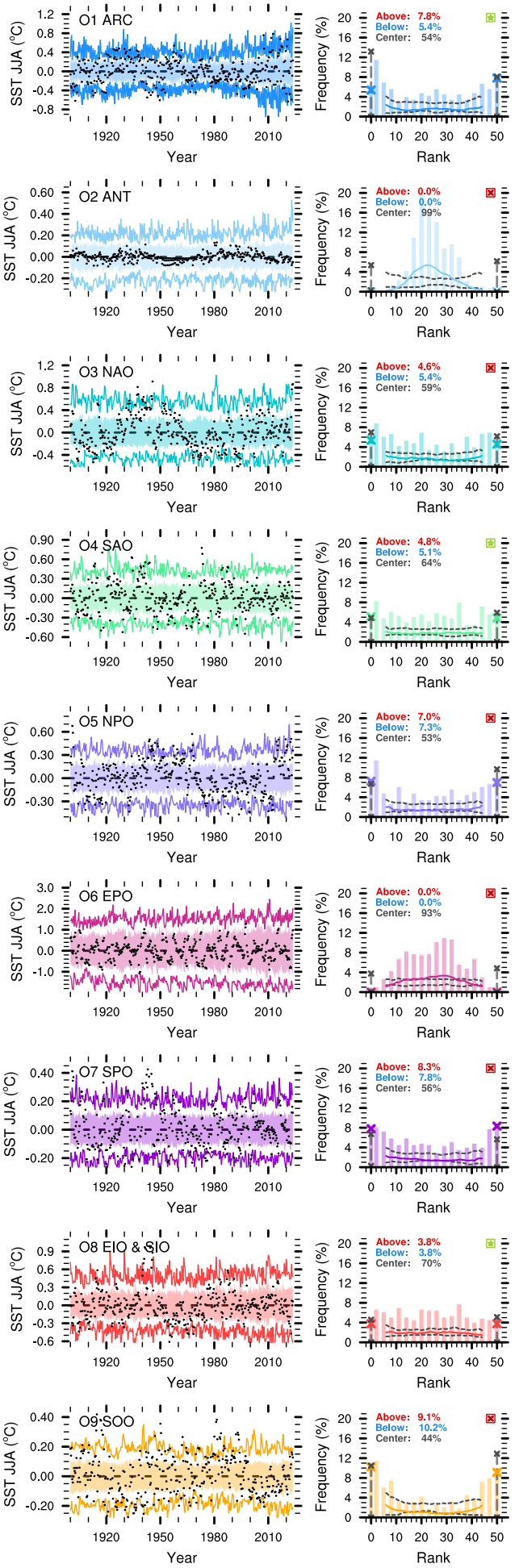
SST JJA Obs outside ensemble spread (%)



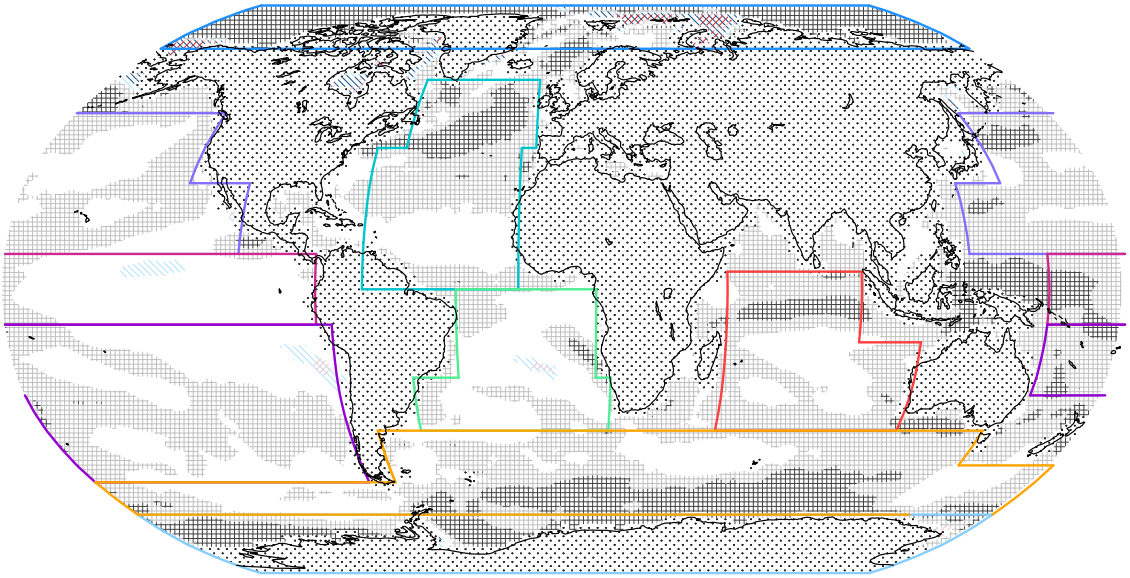
MIROC6 vs ERSSTv5 SST DJF



MIROC6 vs ERSSTv5 SST JJA



MPI-GE5 vs ERSSTv5 White Area = 37.6 %



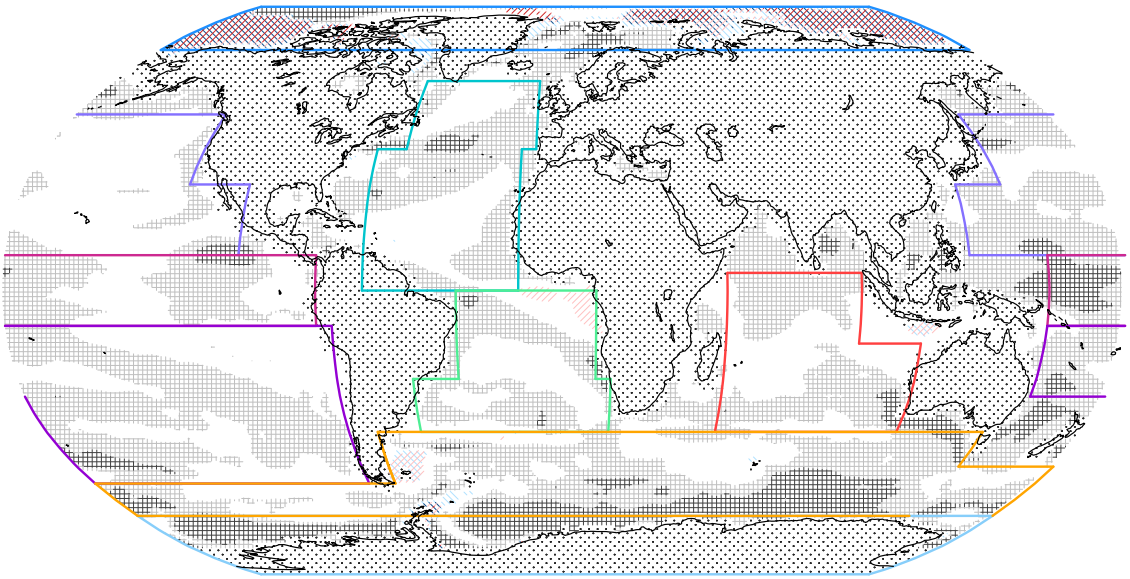
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

MPI-GE5 vs ERSSTv5 White Area = 40.2 %



SST JJA Obs inside central 75th percentile (%)

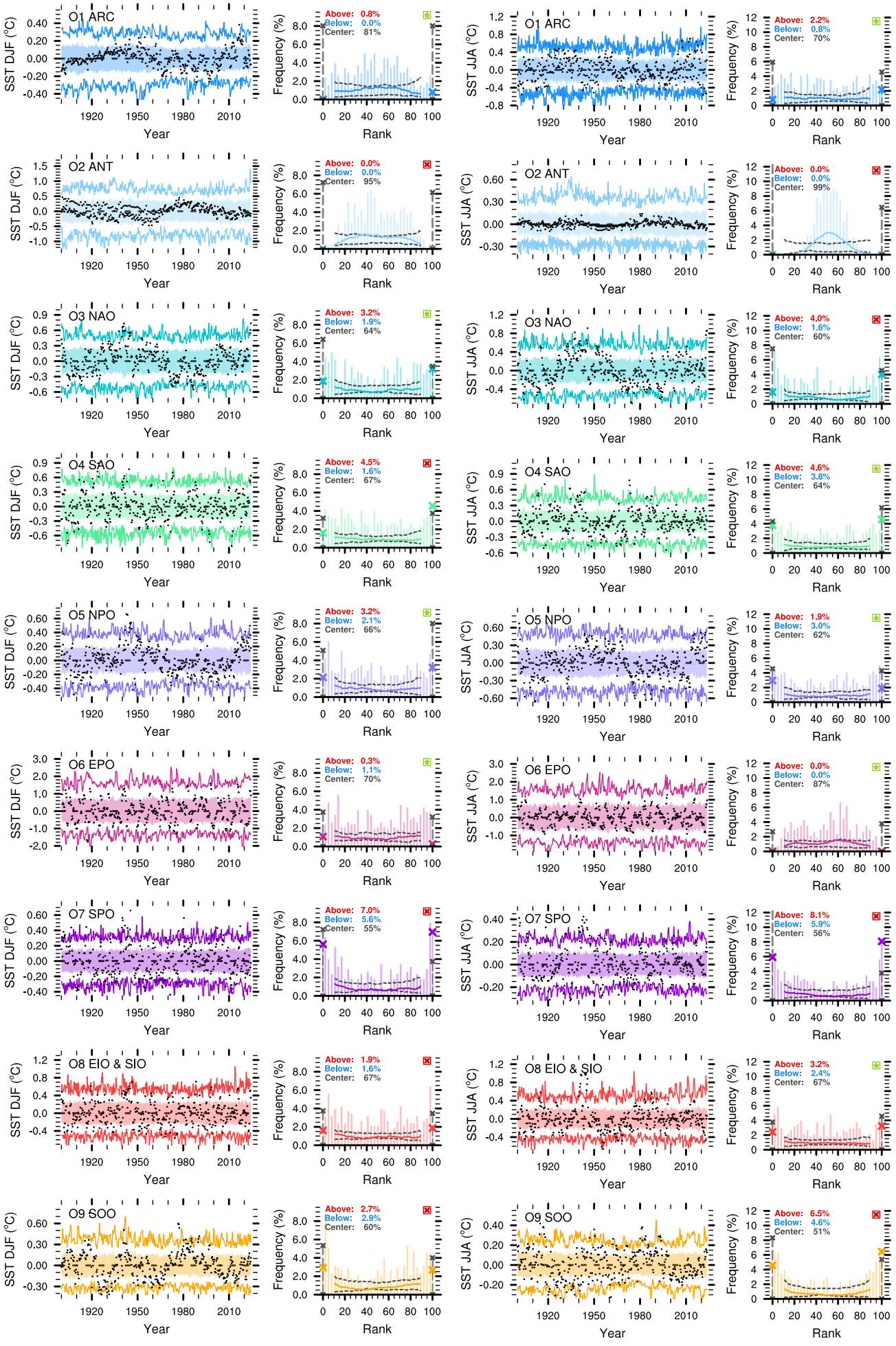
80 95

SST JJA Obs outside ensemble spread (%)

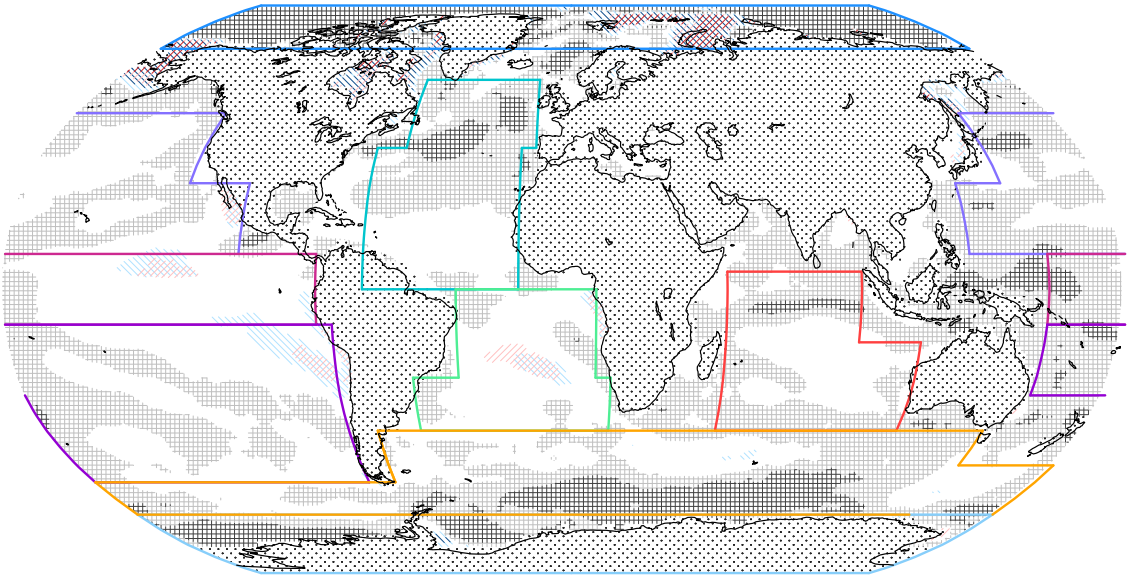
8 20

MPI-GE5 vs ERSSTv5 SST DJF

MPI-GE5 vs ERSSTv5 SST JJA



MPI-GE6 vs ERSSTv5 White Area = 41.9 %



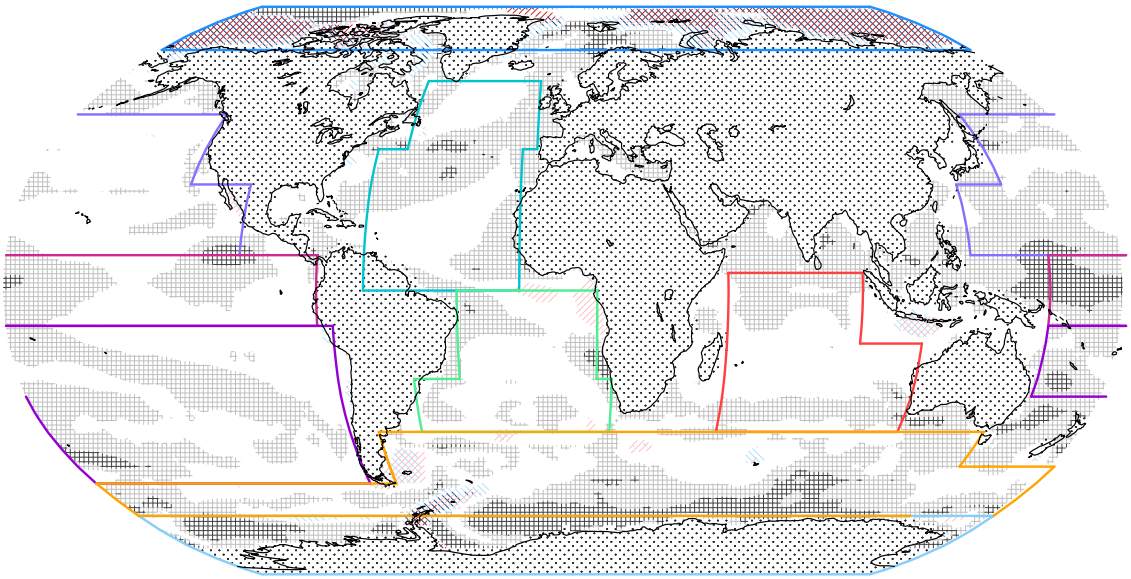
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

MPI-GE6 vs ERSSTv5 White Area = 45.9 %



SST JJA Obs inside central 75th percentile (%)

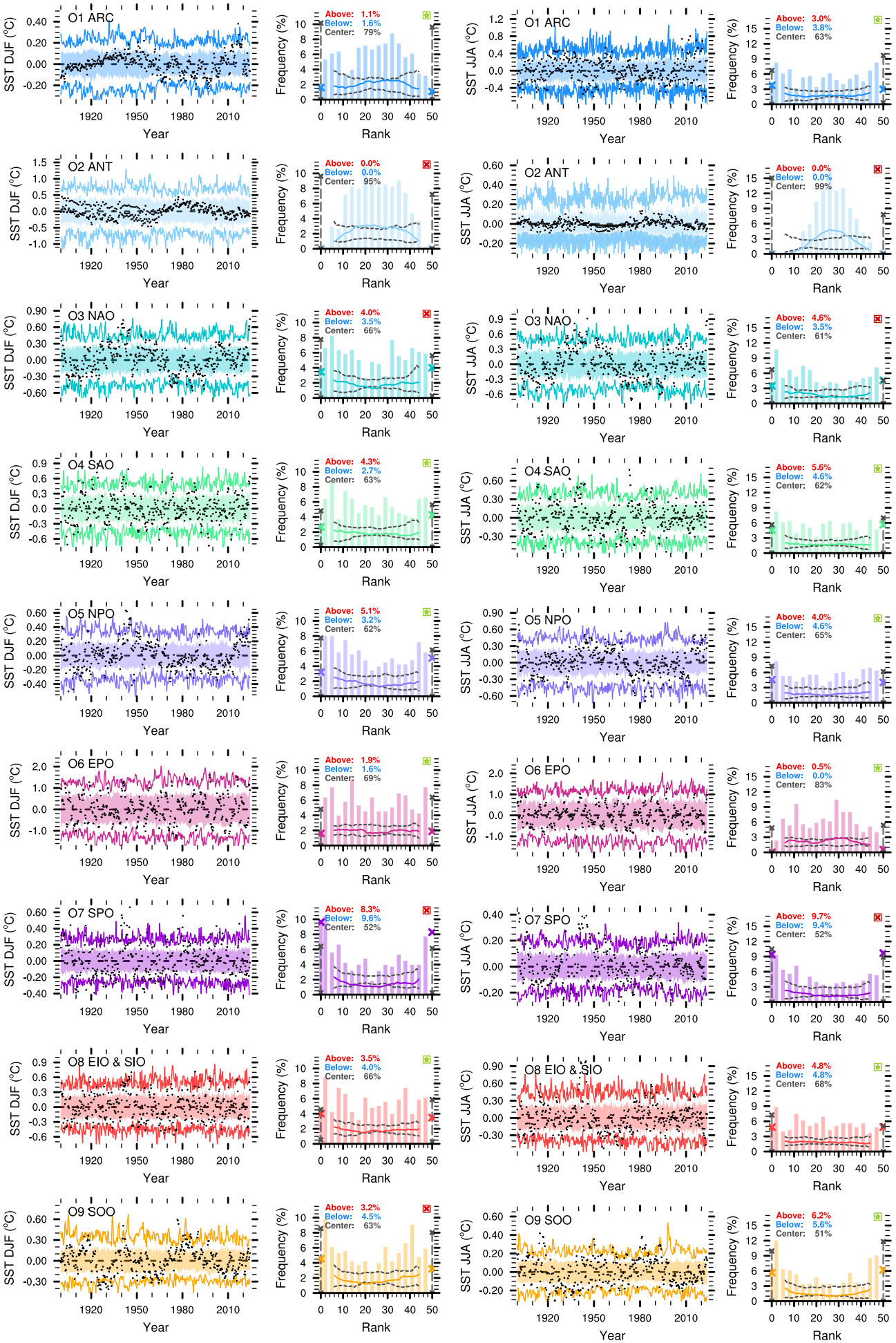
80 95

SST JJA Obs outside ensemble spread (%)

8 20

MPI-GE6 vs ERSSTv5 SST DJF

MPI-GE6 vs ERSSTv5 SST JJA



Non-Detrended Ocean Surface Temperatures

Rank-frequency variability evaluation framework for non-detrended sea surface temperature (SST) anomalies over ocean grid cells.

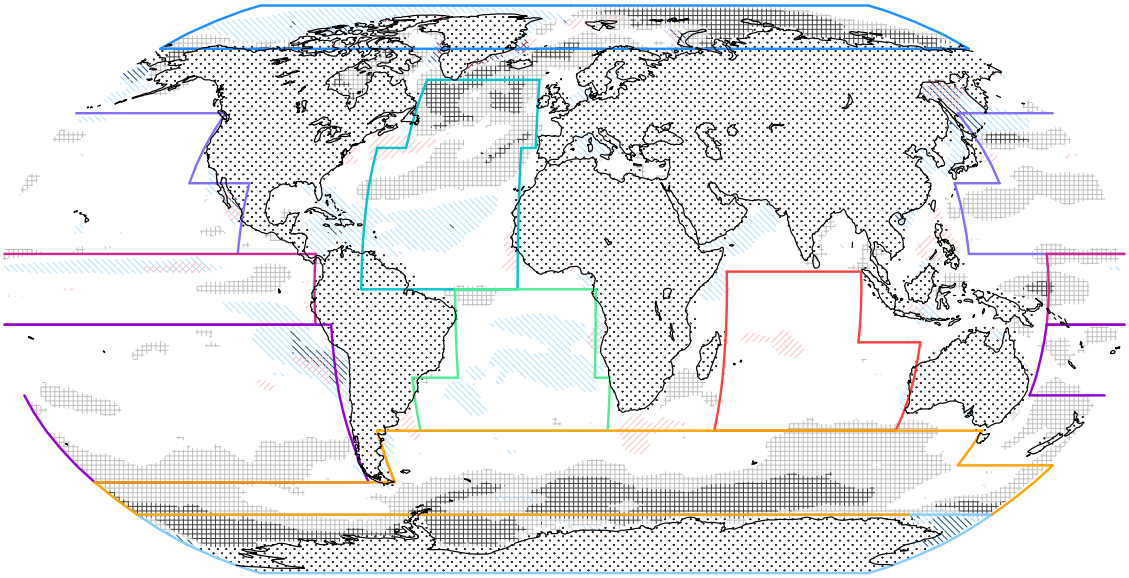
Maps show grid-cell evaluation of the simulated DJF and JJA monthly mean temperature anomalies for the 11 SMILEs included against ERSSTv5 observations globally. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensemble (12.5th to 87.5th percentiles) for more than 80% of months (light grey) or for more than 95% of months (dark grey). Red and blue shading represents where observations are larger than the ensemble maximum (red) or smaller than the ensemble minimum (blue), respectively, for more than 8% of the months (light red and blue) or for more than 20% of the months (dark red and blue). Dotted areas represent land areas or grid cells where observations are missing and are therefore excluded from this analysis. Colored boxes demark the boundaries of each ocean region assessed. The percentage of assessed grid-cells that present none of these biases, unbiased grid-cells, in given at the top (white area). For each assessed region, models are considered to provide an adequate grid-cell representation when they exhibit more than 50% of unbiased grid cells in the region, fulfilling Criteria 2 of our evaluation framework.

Time series and rank frequency histograms show spatially aggregated DJF and JJA SST for each ocean region for all 11 SMILEs. Time series show the ensemble maximum and minimum (color lines) and central 75th percentile ensemble spread (shading) are shown against observations (black dots).

Rank histograms represent the frequency of each place that observations would take in a list of ensemble members ordered by ascending temperature anomaly values. Rank histograms show the observations rank frequency accumulated for 3-rank bins (bars), the running mean rank frequency over a centered n/5 rank window (lines; for 1 to n-1 ranks) and the absolute frequencies of rank 0 and n (crosses), with n the number of ensemble members, for observations (color) and perfect model rank range (gray).

To illustrate how internal variability sampling may affect rank frequencies given the non-infinite record length considered, we include a comparison to the perfect-model rank range, which shows the range of rank frequencies that each ensemble member would take if it were observations (see Methods in main article for further details). If the rank exhibited by observations (colors) is within this perfect-model range (gray; allowing a maximum 10% deviation outside of this range), the rank frequency evaluation shows an adequate model performance, and any deviations from a perfectly frank rank histogram can be assumed to be within the range of deviations that could be caused by internal variability. In the case the observations rank frequency is within the perfect-model range for all rank windows, Criteria 1 of our evaluation framework is met and this is highlighted by a green star at the top right; if not, by a red cross. Percentages at the top left show the frequency of regionally averaged observations occurring above (red) or below (blue) ensemble limits, or clustering within the central 75th percentile range (gray), analogous to the grid-cell evaluation in Criteria 2.

ACCESS vs ERSSTv5 White Area = 60.5 %



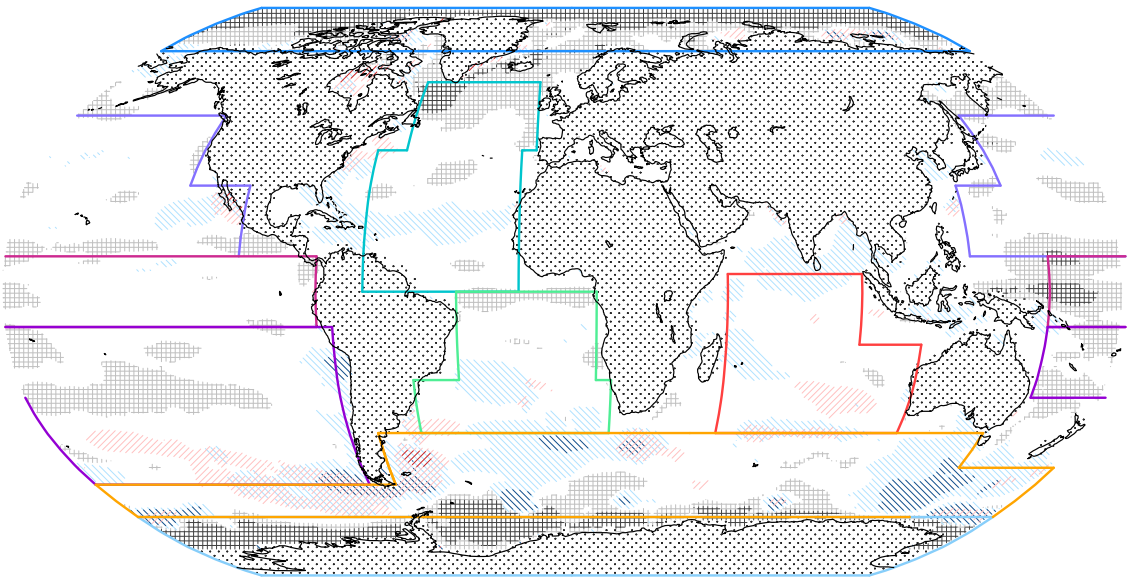
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

ACCESS vs ERSSTv5 White Area = 58.5 %



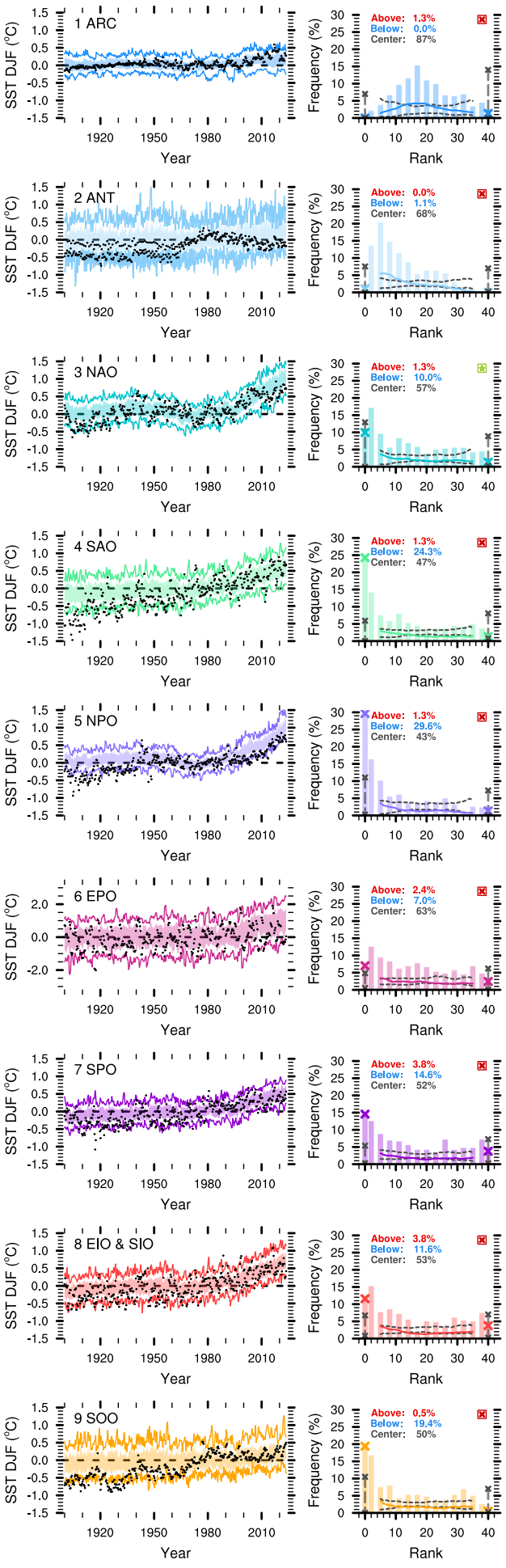
SST JJA Obs inside central 75th percentile (%)

80 95

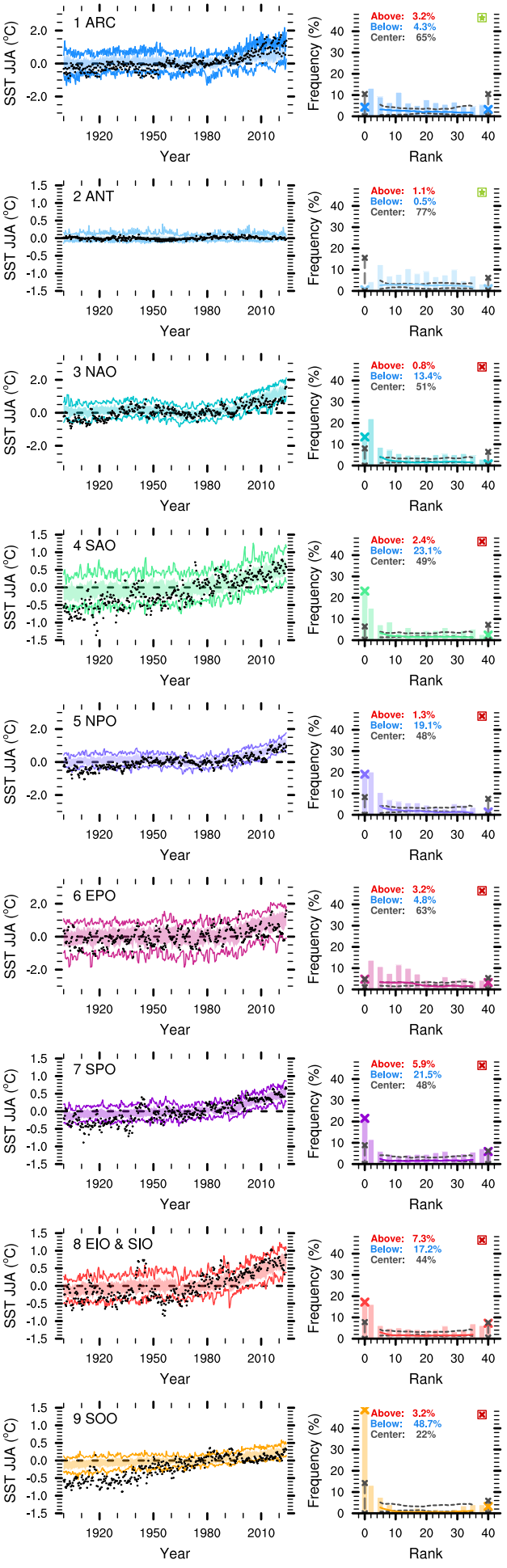
SST JJA Obs outside ensemble spread (%)

8 20

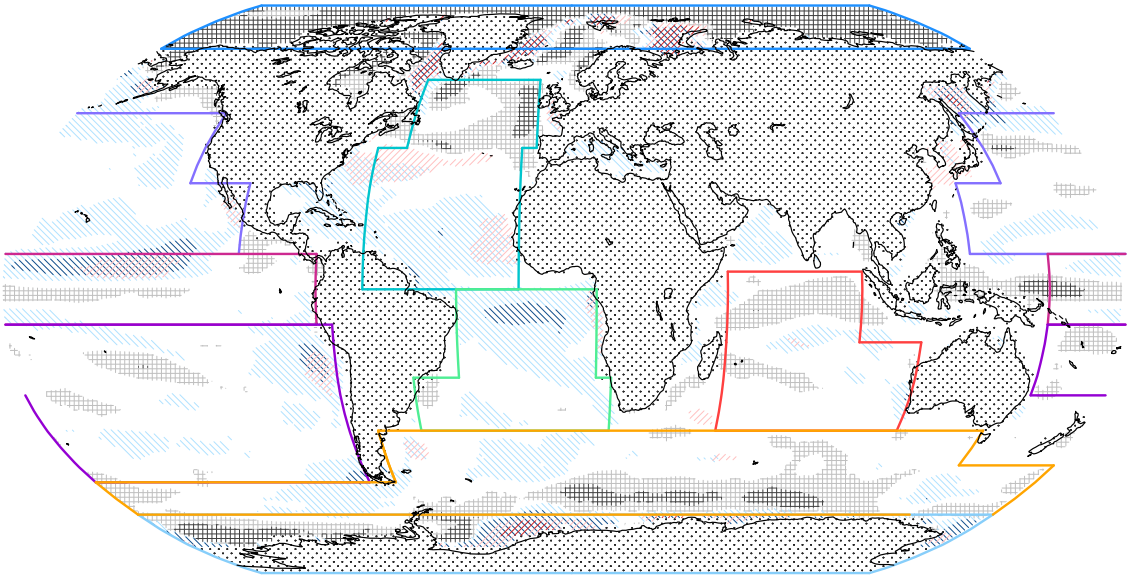
ACCESS vs ERSSTv5 SST DJF



ACCESS vs ERSSTv5 SST JJA



CanESM2 vs ERSSTv5 White Area = 51.8 %



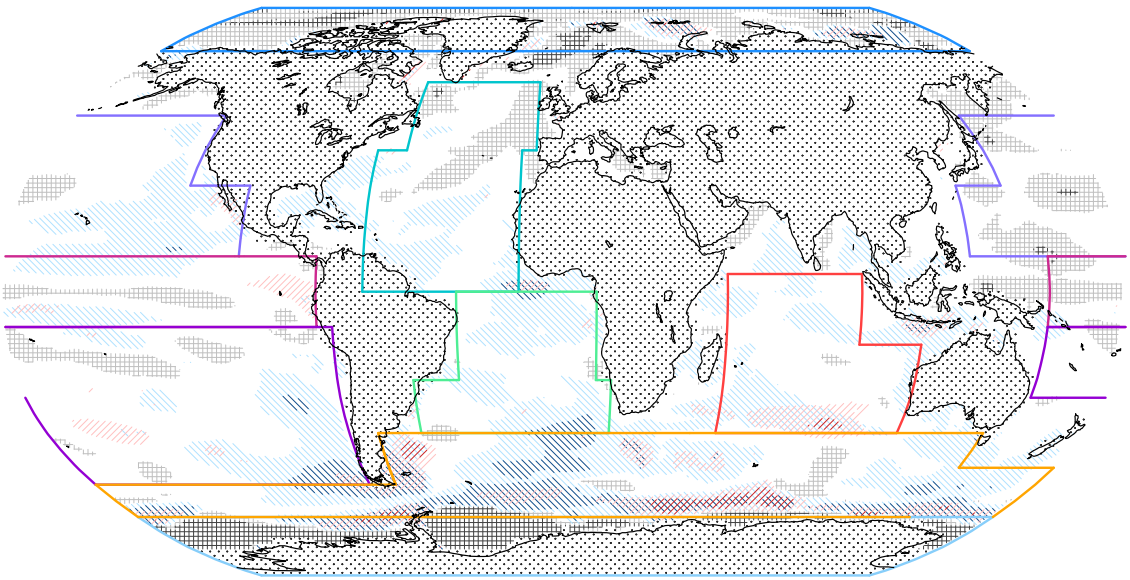
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

CanESM2 vs ERSSTv5 White Area = 48.5 %



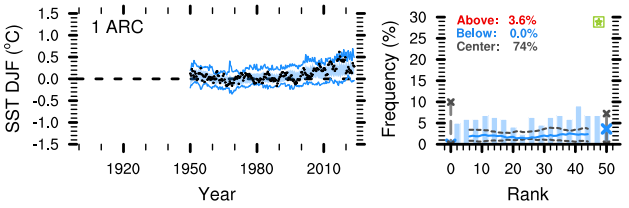
SST JJA Obs inside central 75th percentile (%)

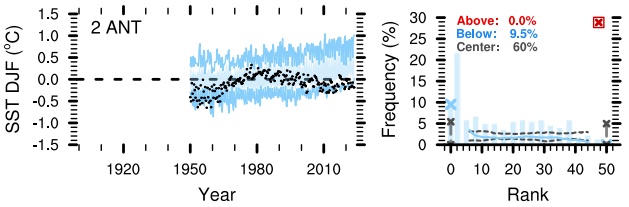
80 95

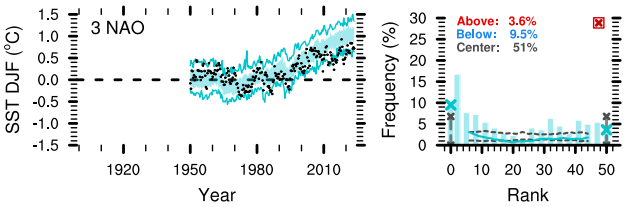
SST JJA Obs outside ensemble spread (%)

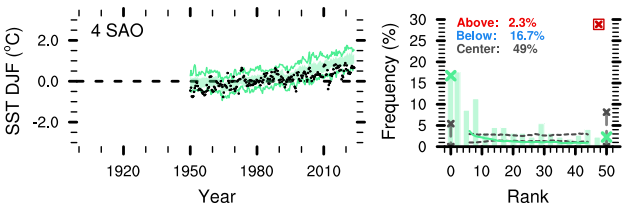
8 20

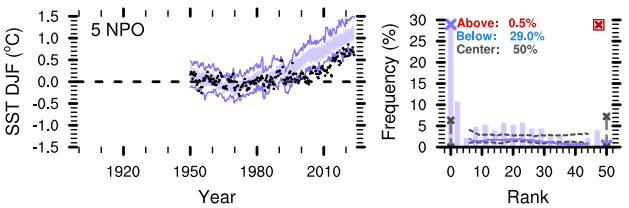
CanESM2 vs ERSSTv5 SST DJF

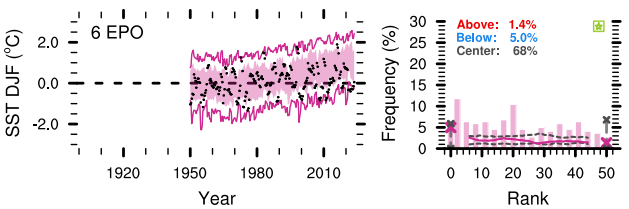


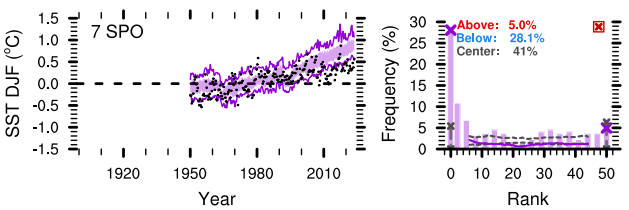


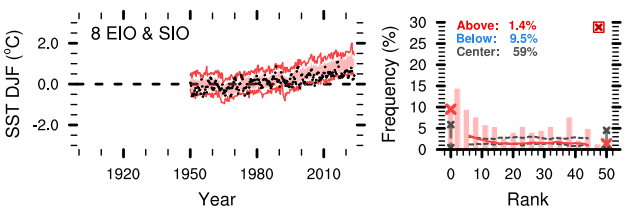


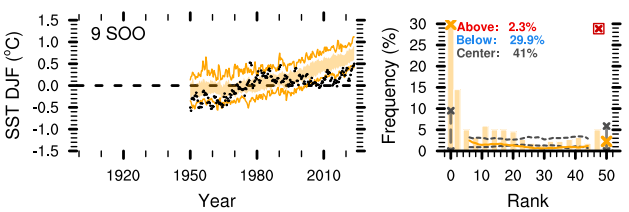




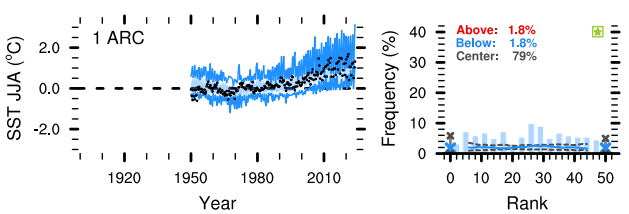


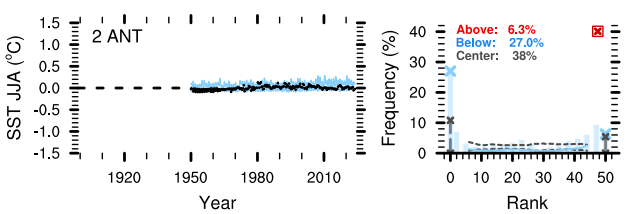


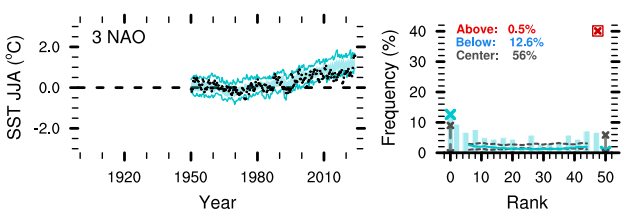


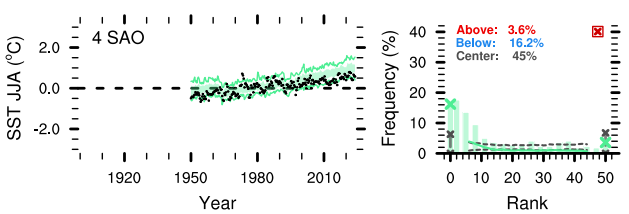


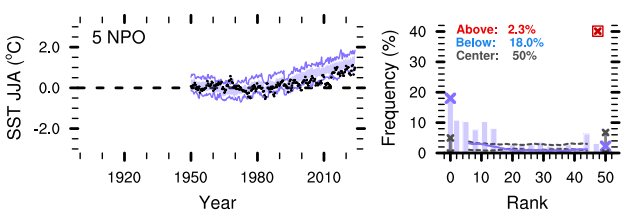
CanESM2 vs ERSSTv5 SST JJA

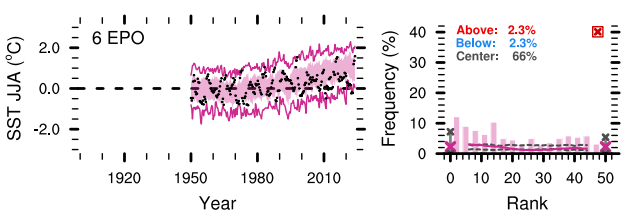


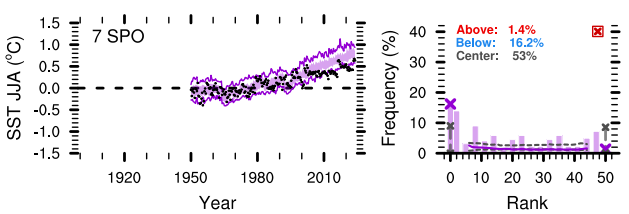


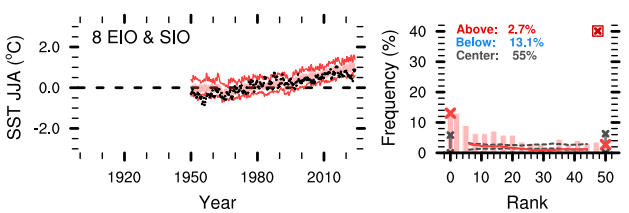


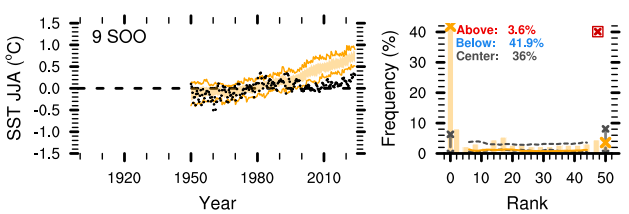




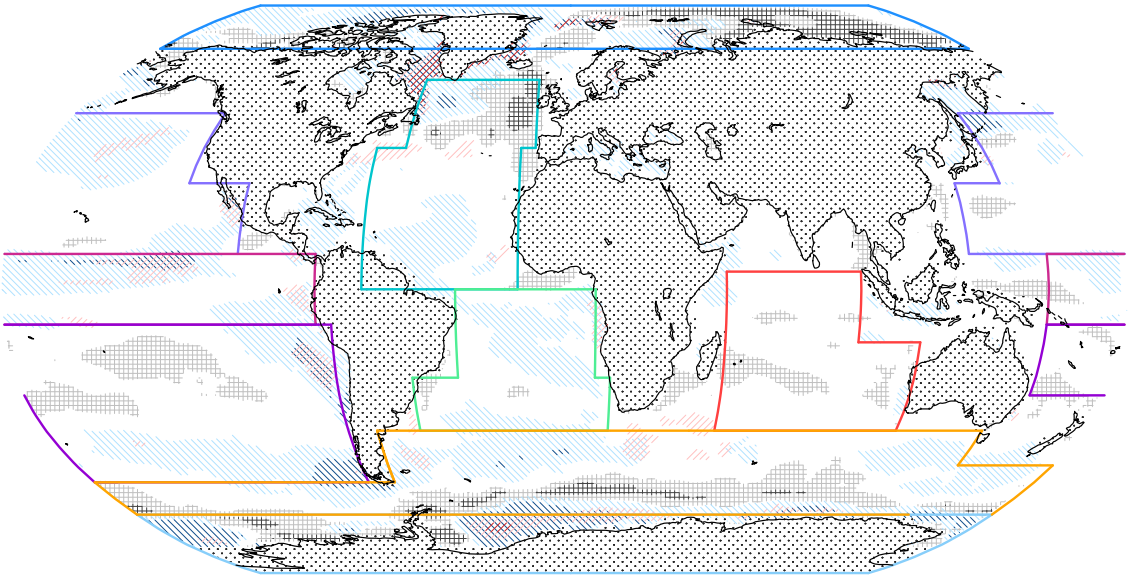








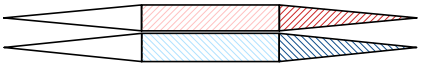
CanESM5 vs ERSSTv5 White Area = 55.0 %



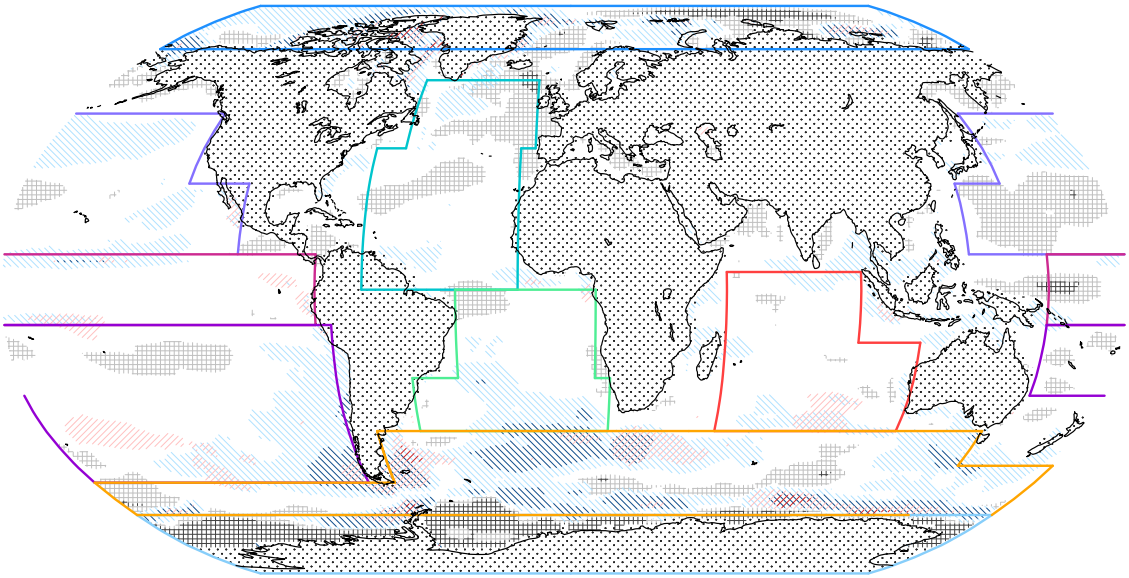
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



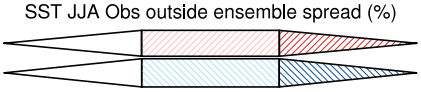
CanESM5 vs ERSSTv5 White Area = 55.9 %



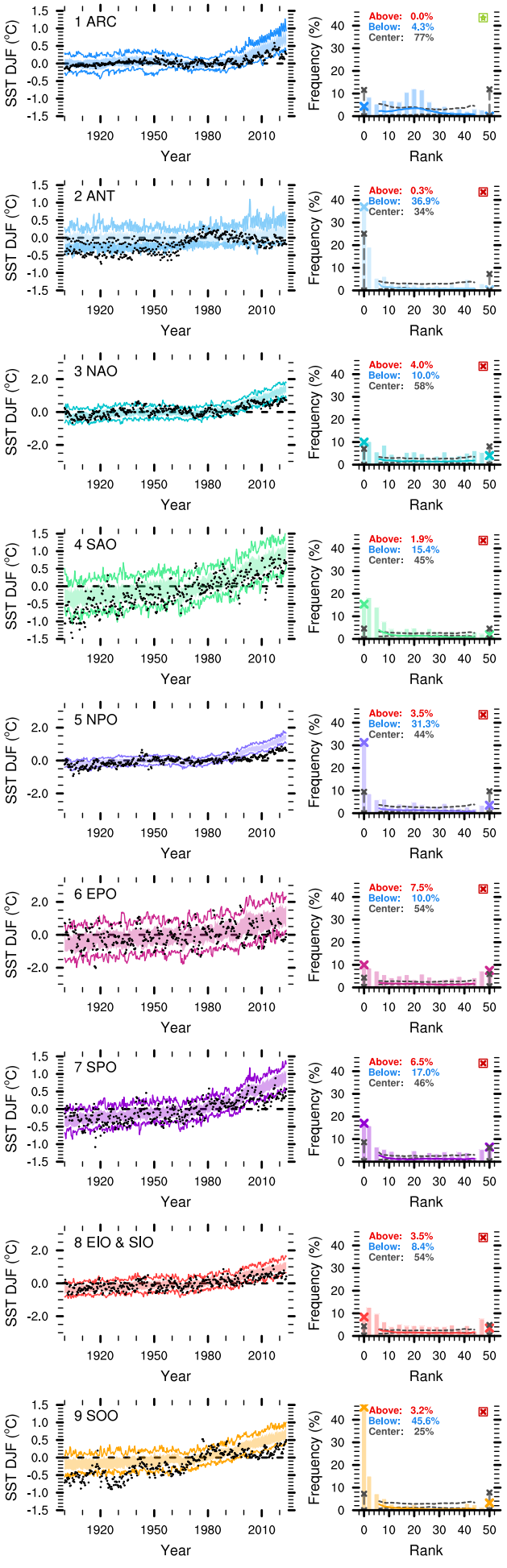
SST JJA Obs inside central 75th percentile (%)



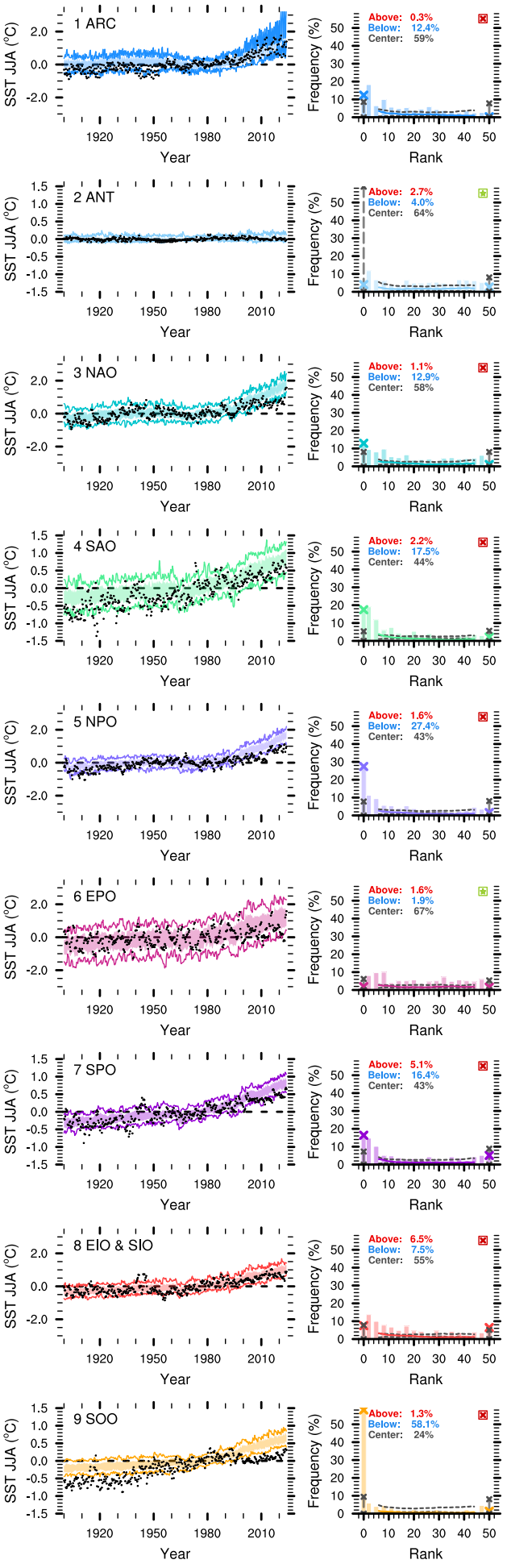
SST JJA Obs outside ensemble spread (%)



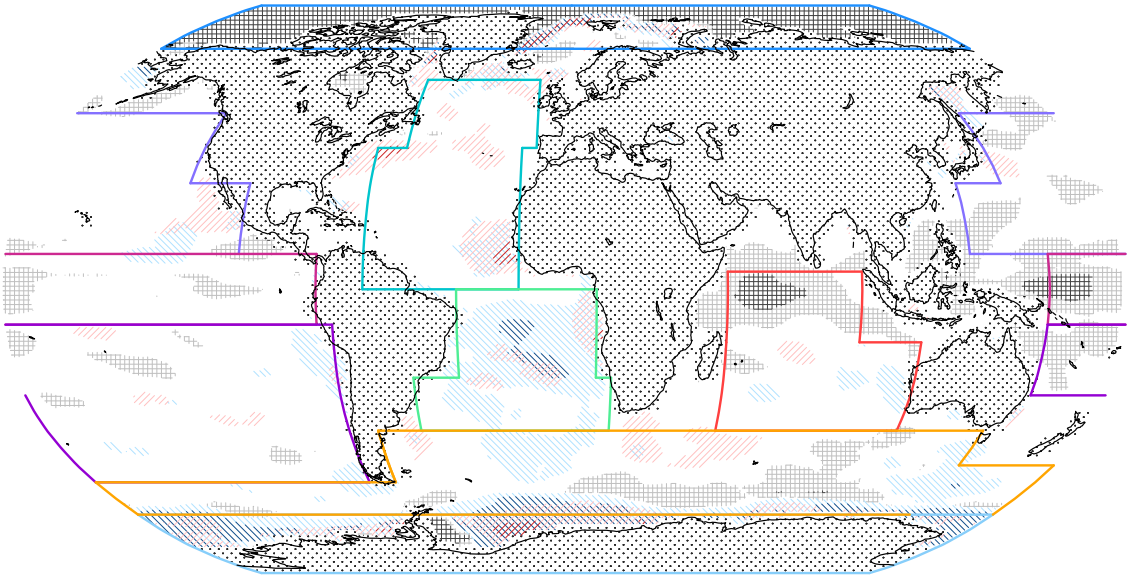
CanESM5 vs ERSSTv5 SST DJF



CanESM5 vs ERSSTv5 SST JJA



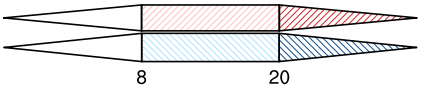
CESM-LE vs ERSSTv5 White Area = 61.8 %



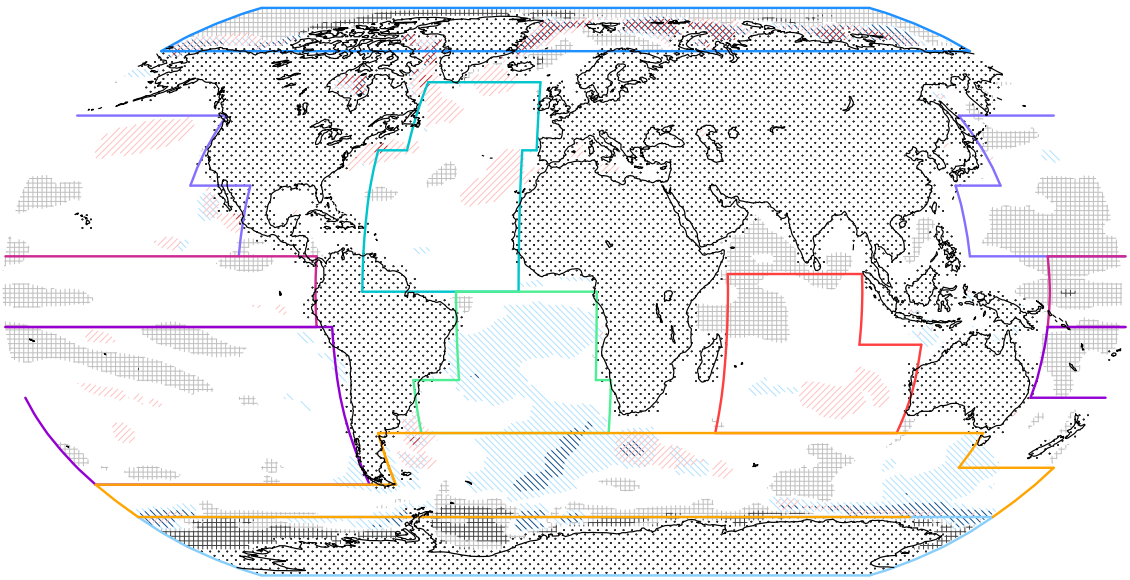
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



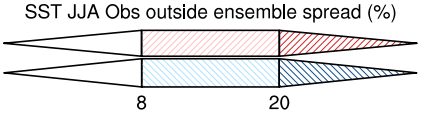
CESM-LE vs ERSSTv5 White Area = 64.7 %



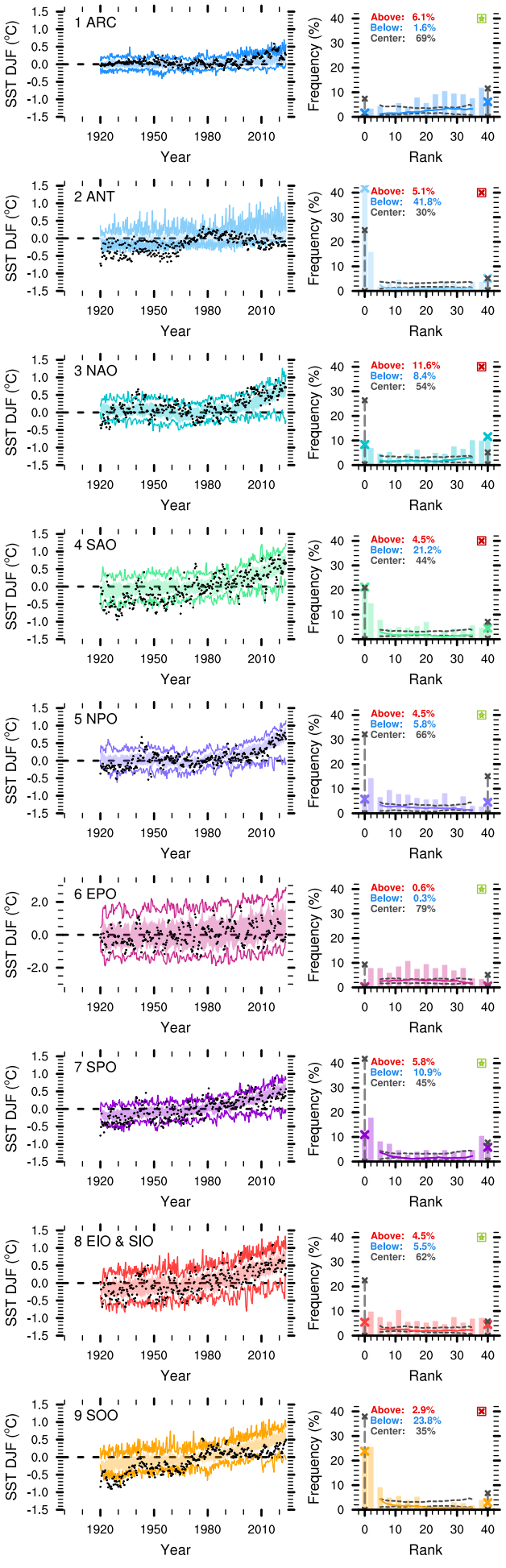
SST JJA Obs inside central 75th percentile (%)



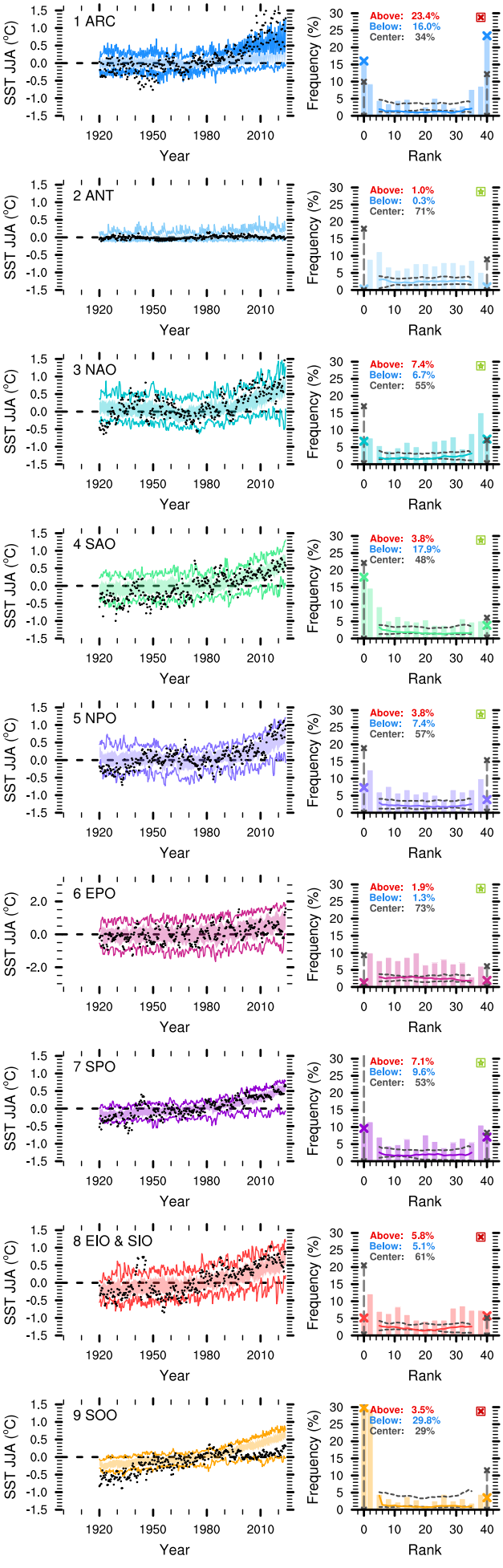
SST JJA Obs outside ensemble spread (%)



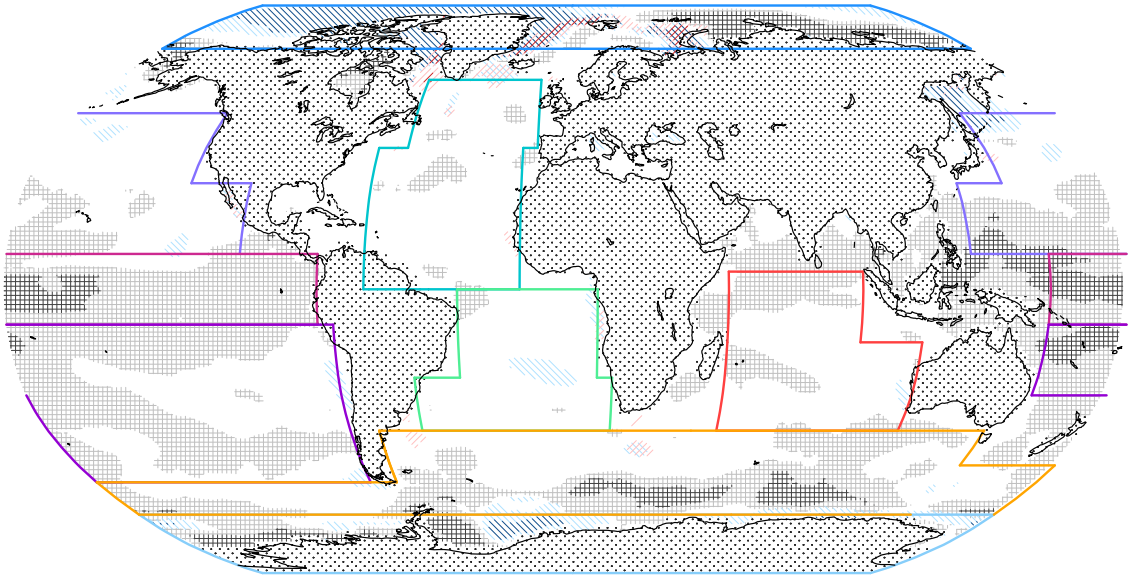
CESM-LE vs ERSSTv5 SST DJF



CESM-LE vs ERSSTv5 SST JJA



CESM2-LE vs ERSSTv5 White Area = 52.8 %



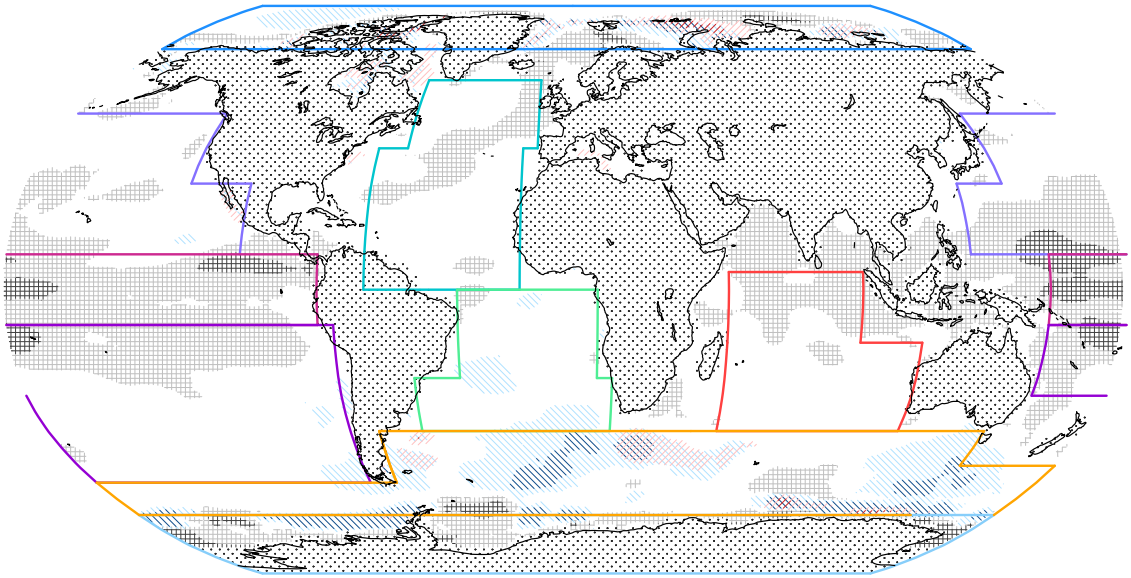
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

CESM2-LE vs ERSSTv5 White Area = 56.3 %



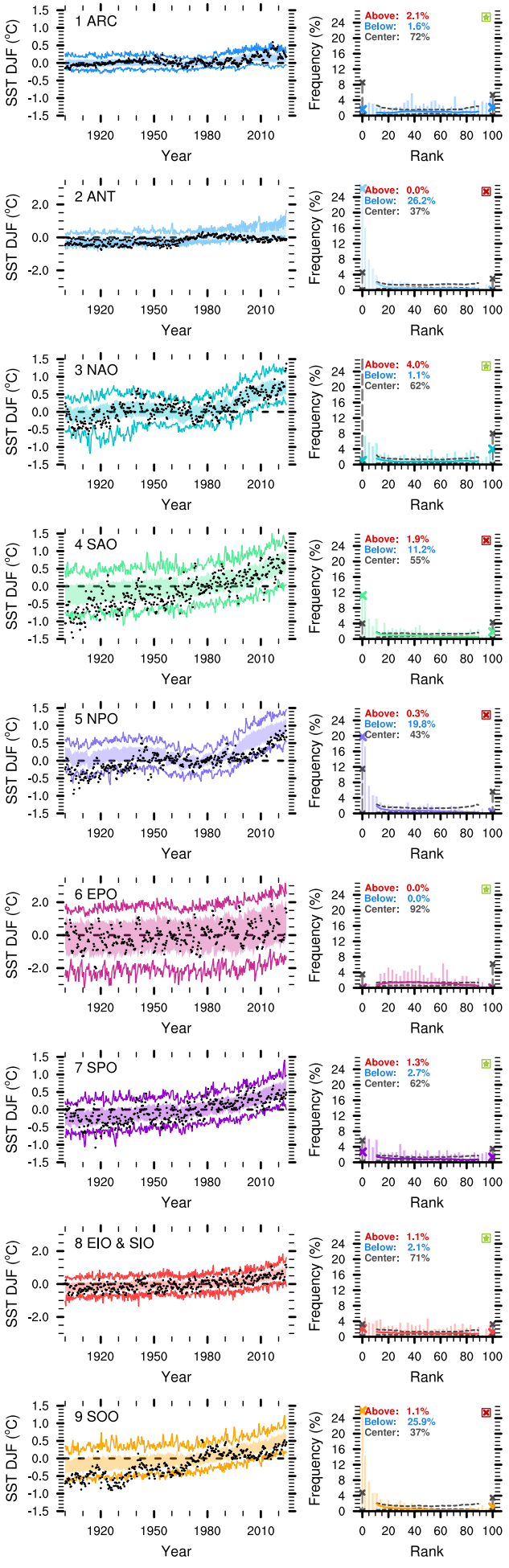
SST JJA Obs inside central 75th percentile (%)

80 95

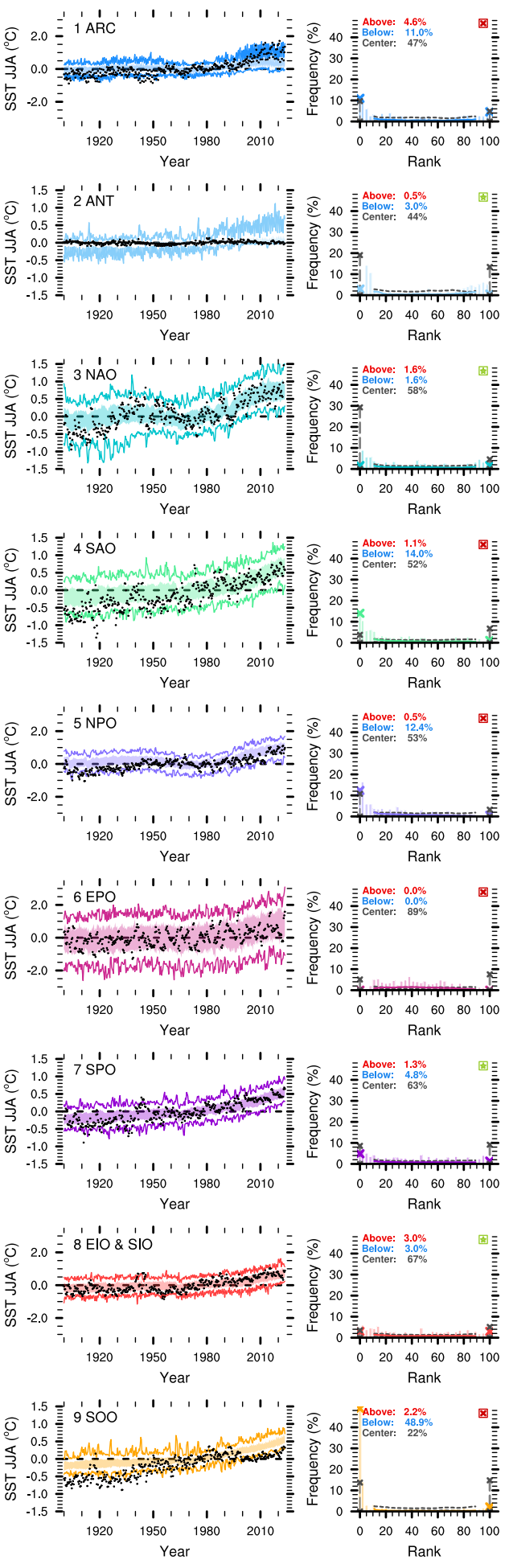
SST JJA Obs outside ensemble spread (%)

8 20

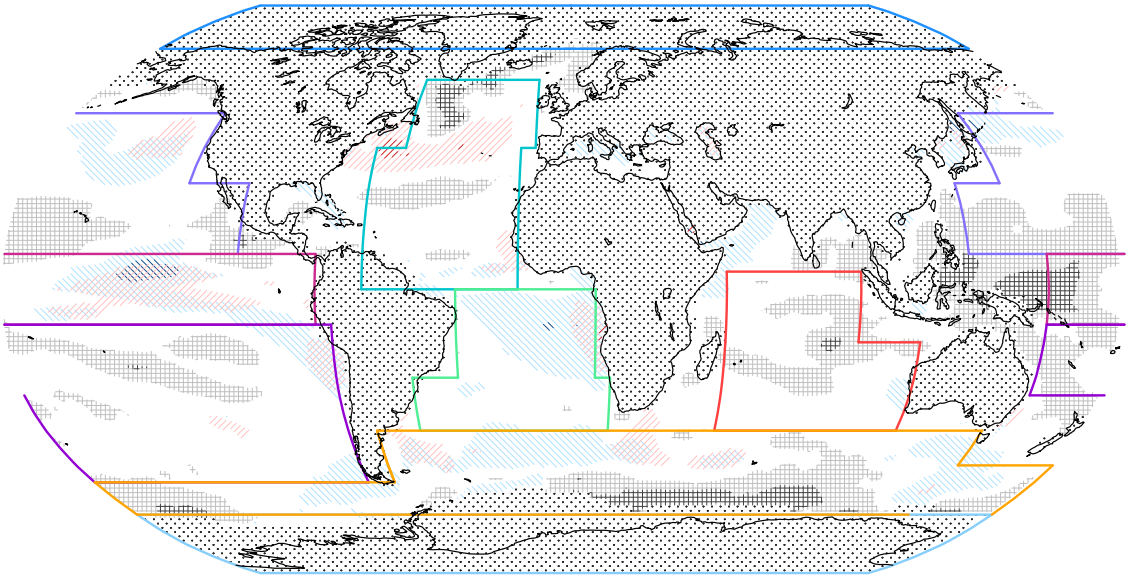
CESM2-LE vs ERSSTv5 SST DJF



CESM2-LE vs ERSSTv5 SST JJA



CSIRO-Mk360 vs ERSSTv5 White Area = 54.9 %



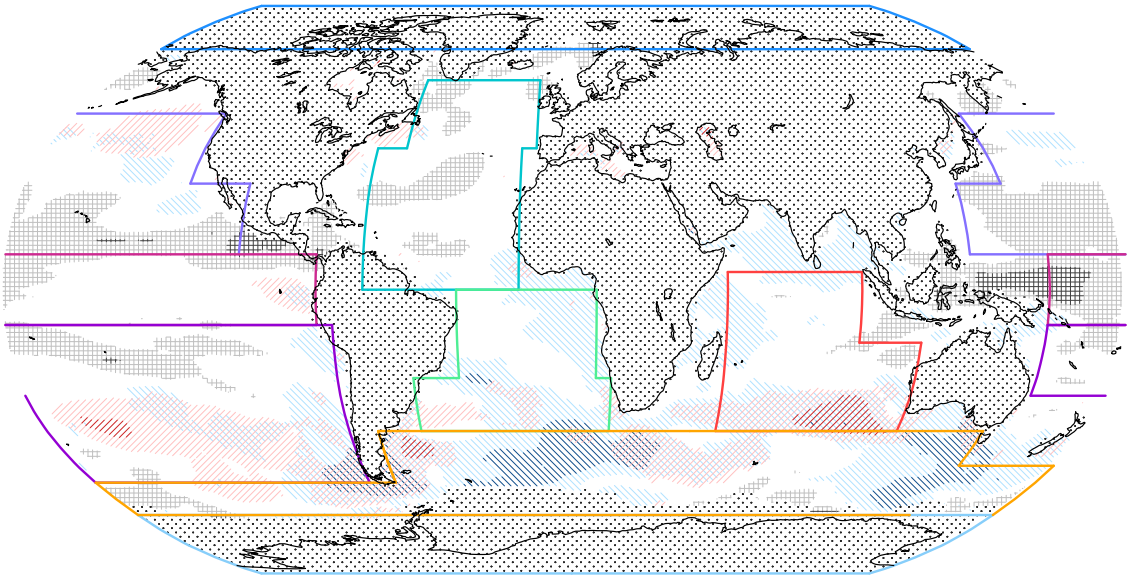
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

CSIRO-Mk360 vs ERSSTv5 White Area = 47.3 %



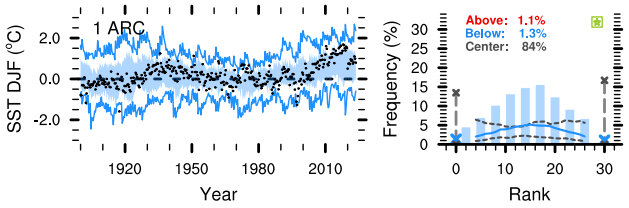
SST JJA Obs inside central 75th percentile (%)

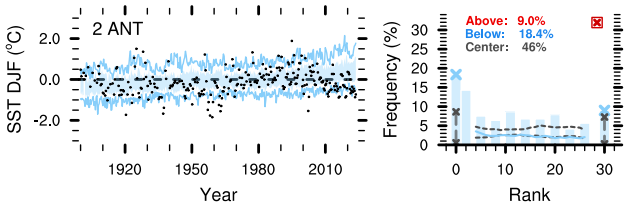
80 95

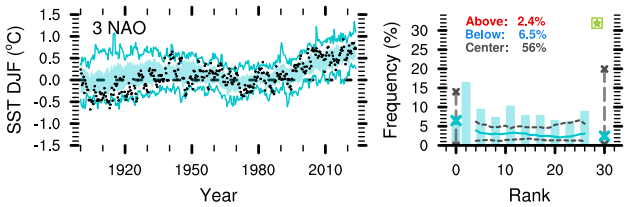
SST JJA Obs outside ensemble spread (%)

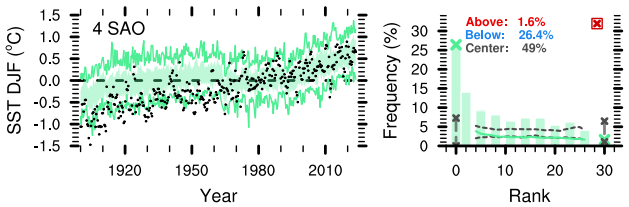
8 20

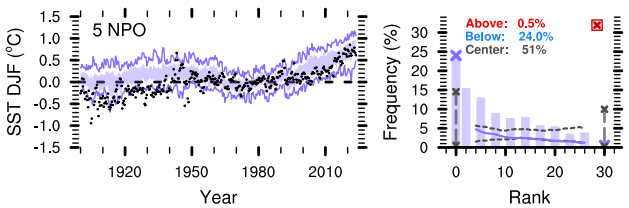
CSIRO-Mk360 vs ERSSTv5 SST DJF

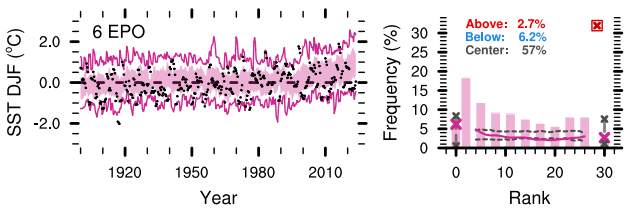


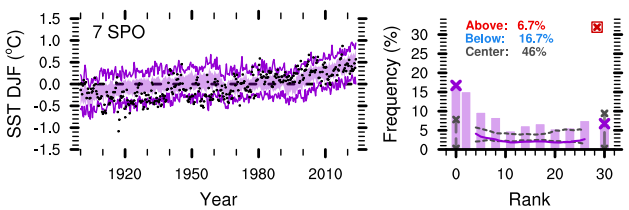


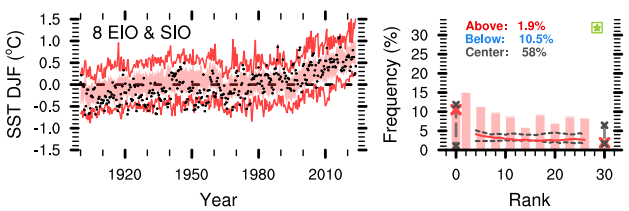


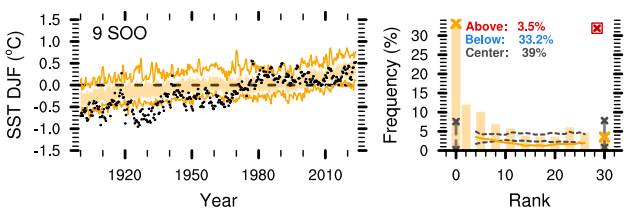




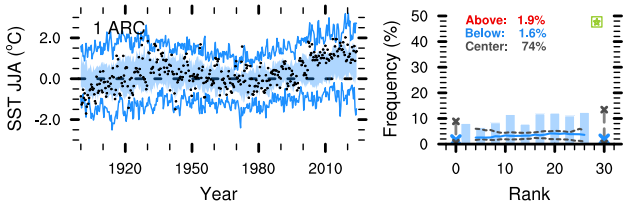


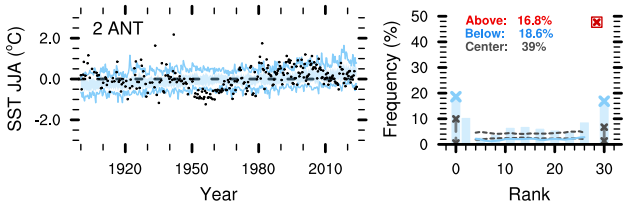


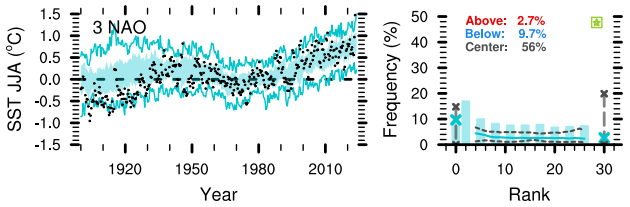


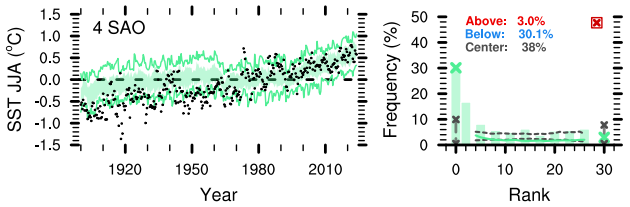


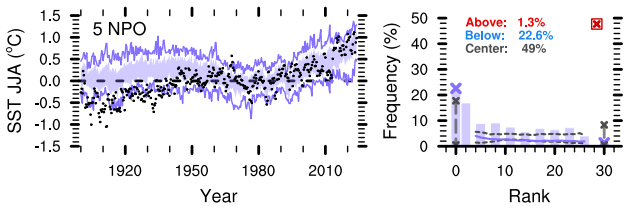
CSIRO-Mk360 vs ERSSTv5 SST JJA

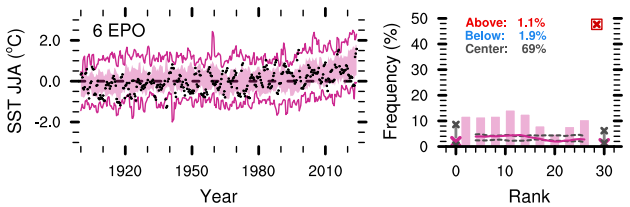


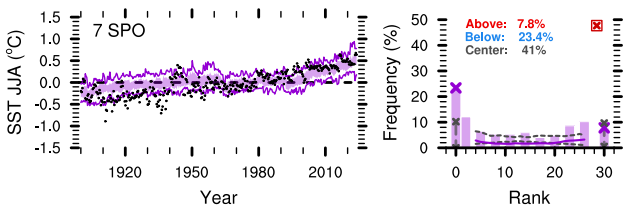


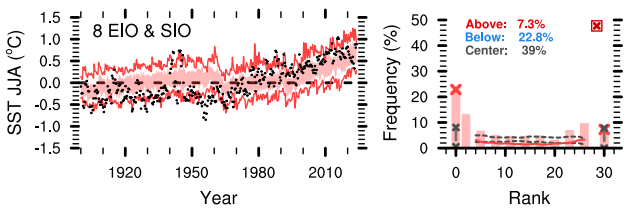


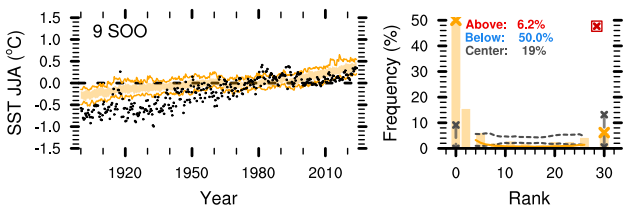




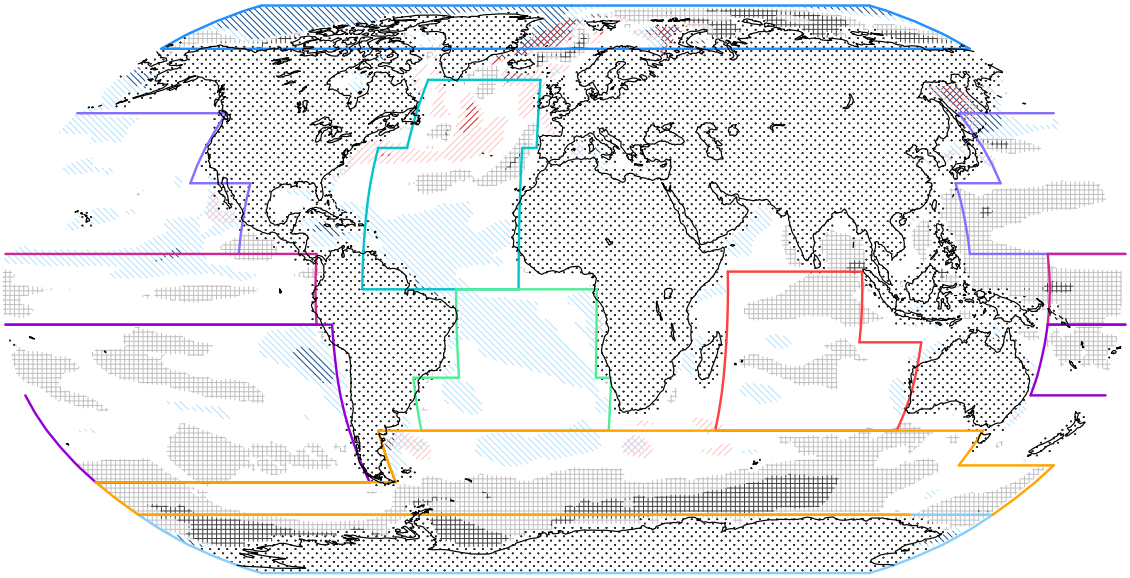








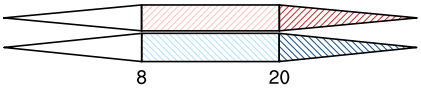
GFDL-ESM2M vs ERSSTv5 White Area = 53.6 %



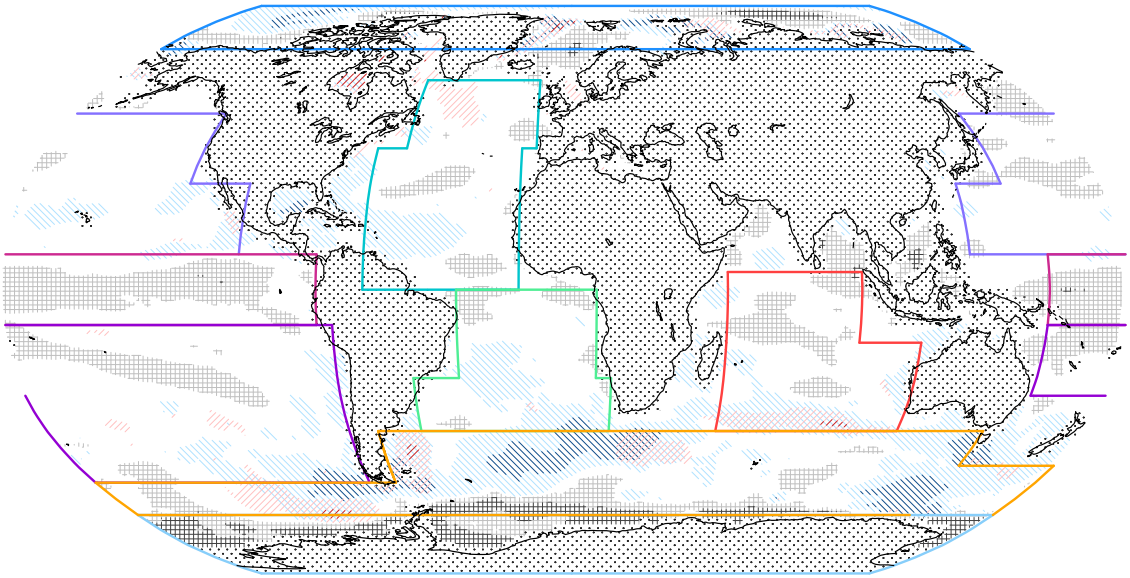
SST DJF Obs inside central 75th percentile (%)



SST DJF Obs outside ensemble spread (%)



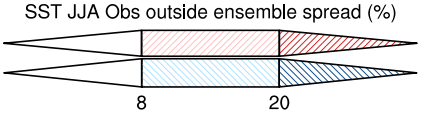
GFDL-ESM2M vs ERSSTv5 White Area = 54.0 %



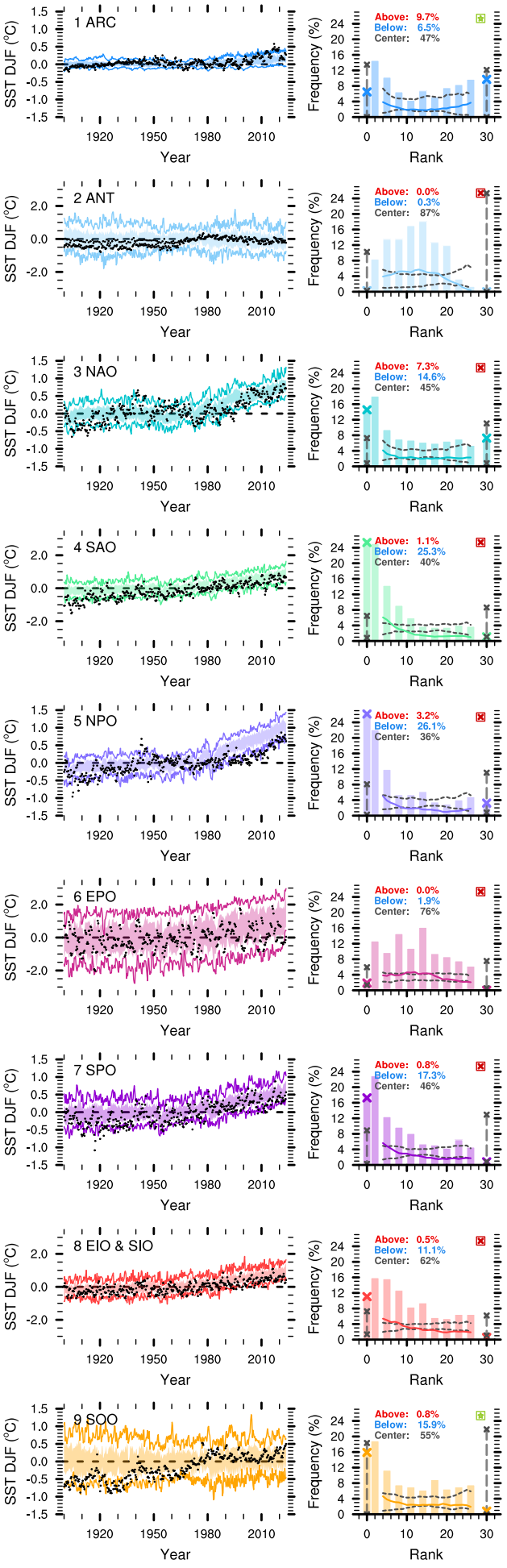
SST JJA Obs inside central 75th percentile (%)



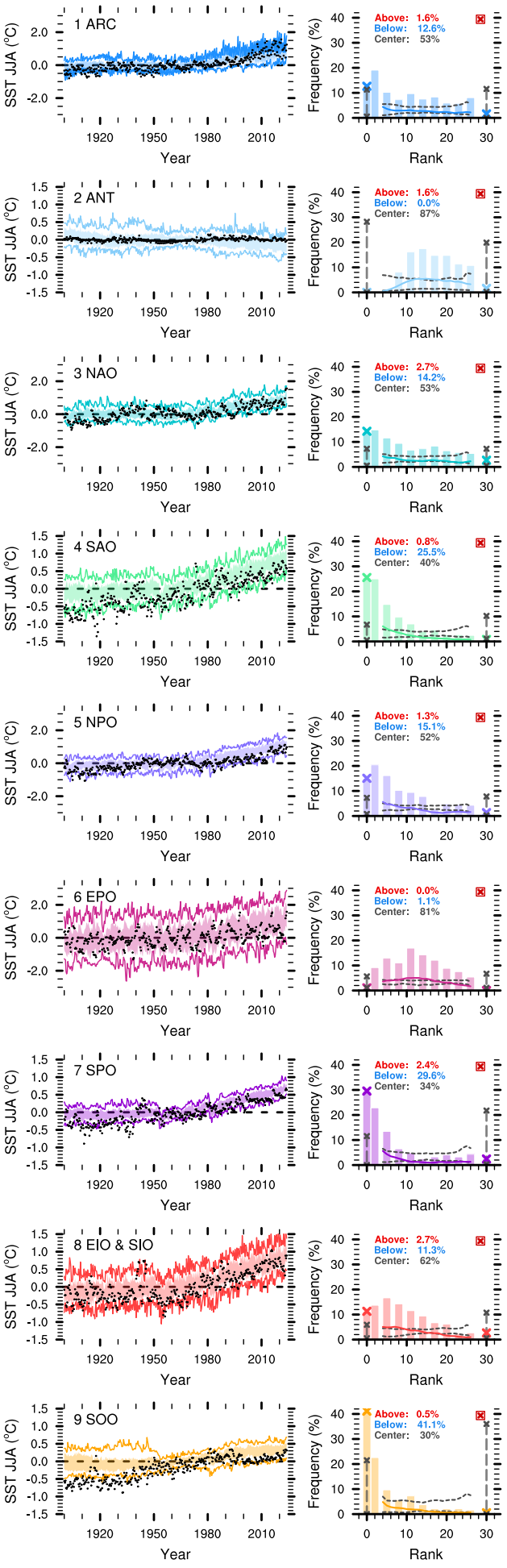
SST JJA Obs outside ensemble spread (%)



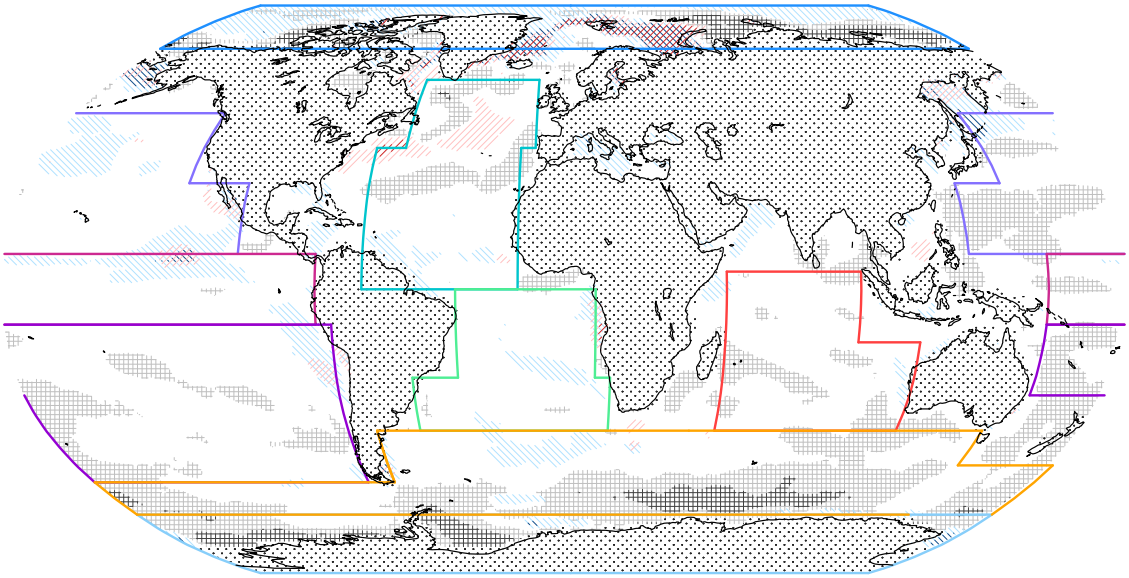
GFDL-ESM2M vs ERSSTv5 SST DJF



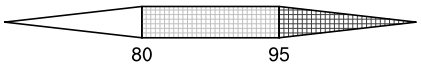
GFDL-ESM2M vs ERSSTv5 SST JJA



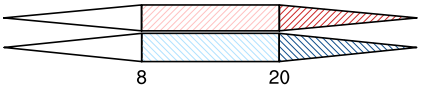
GFDL-SPEAR vs ERSSTv5 White Area = 62.7 %



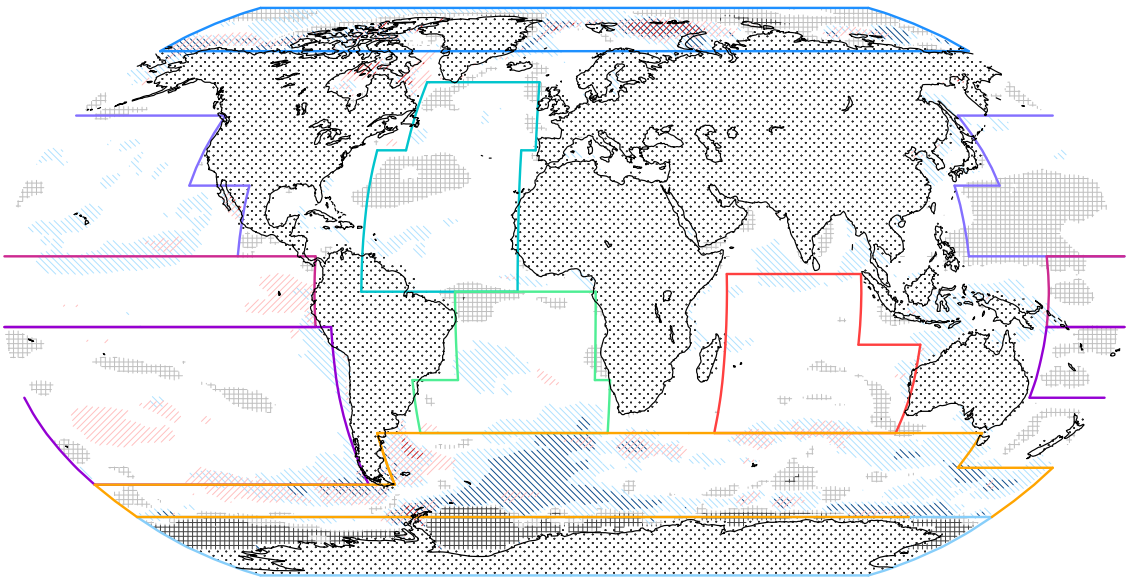
SST DJF Obs inside central 75th percentile (%)



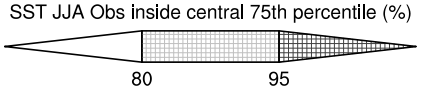
SST DJF Obs outside ensemble spread (%)



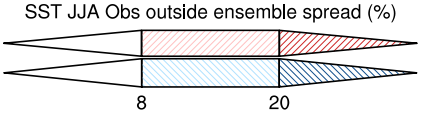
GFDL-SPEAR vs ERSSTv5 White Area = 62.5 %



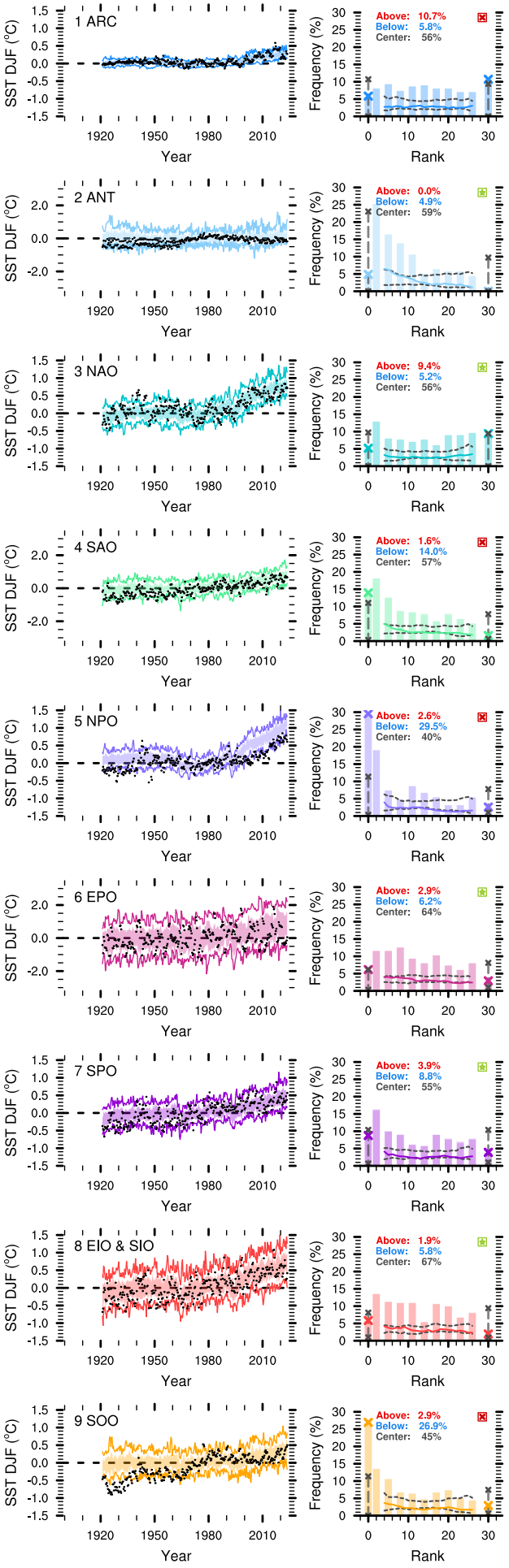
SST JJA Obs inside central 75th percentile (%)



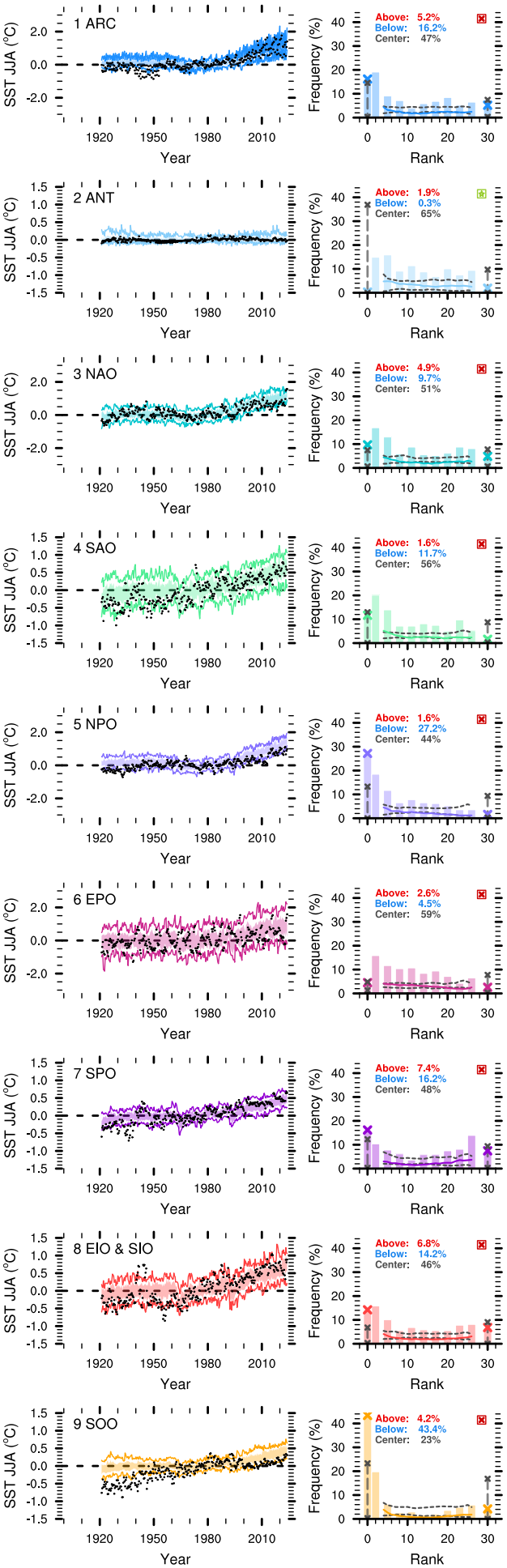
SST JJA Obs outside ensemble spread (%)



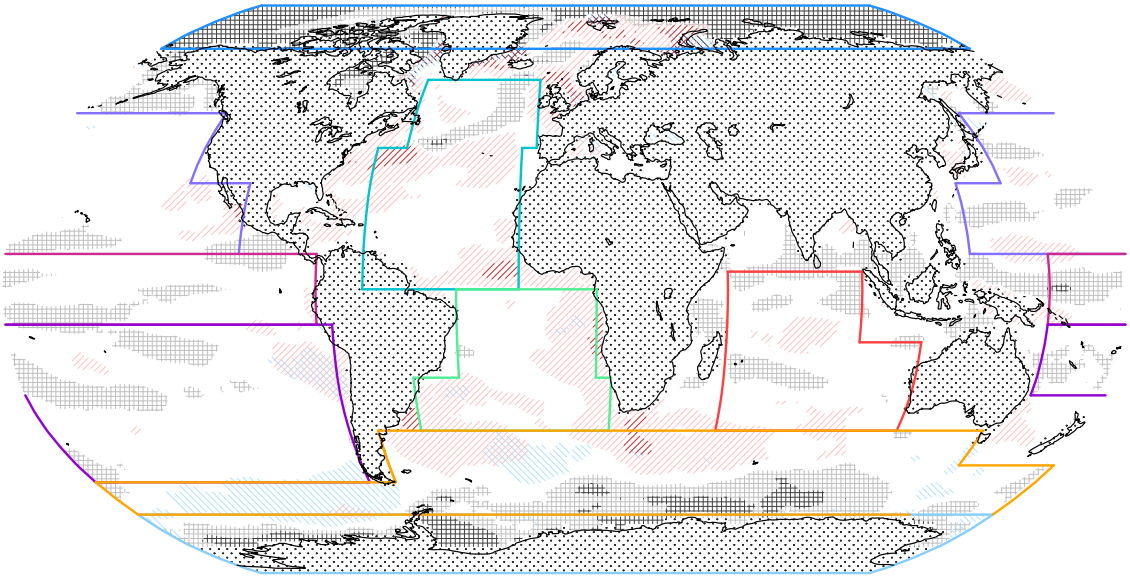
GFDL-SPEAR vs ERSSTv5 SST DJF



GFDL-SPEAR vs ERSSTv5 SST JJA



MIROC6 vs ERSSTv5 White Area = 57.4 %



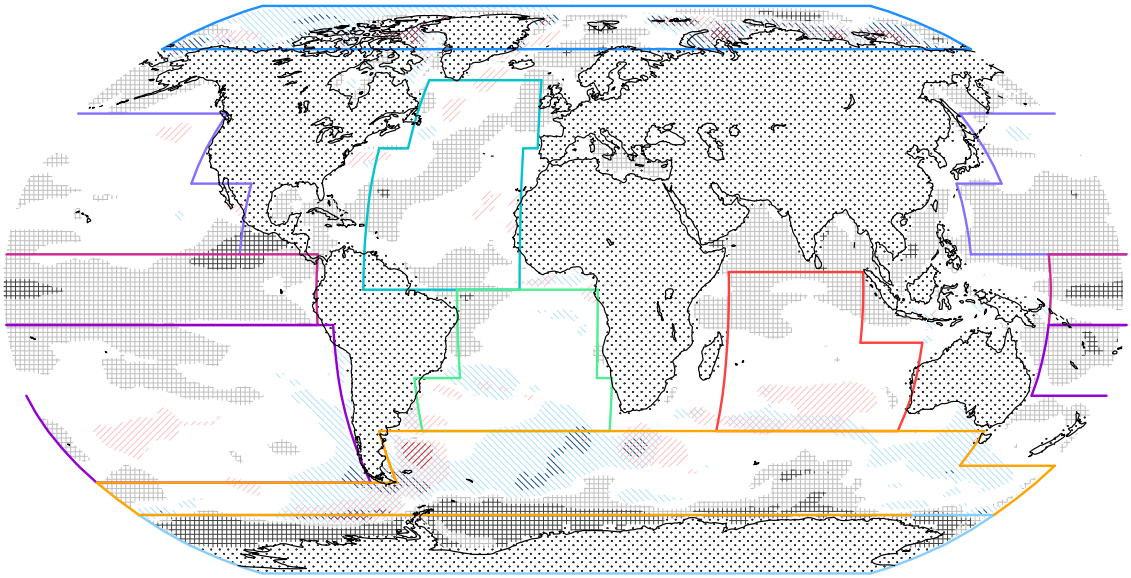
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

MIROC6 vs ERSSTv5 White Area = 48.6 %



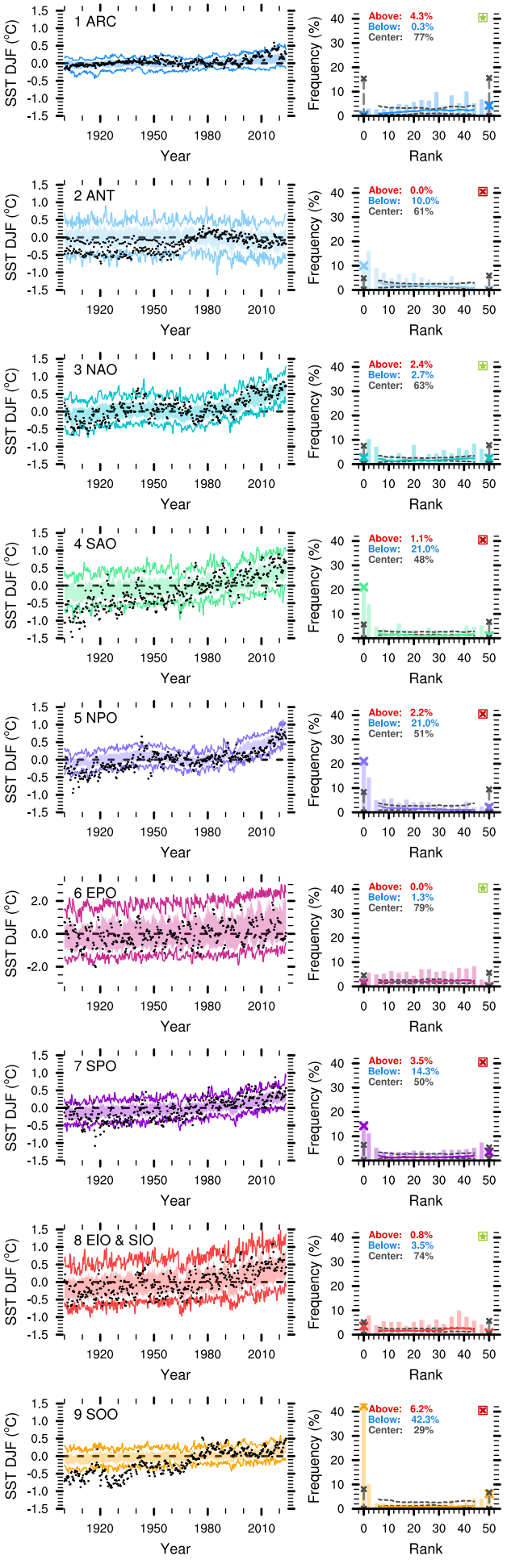
SST JJA Obs inside central 75th percentile (%)

80 95

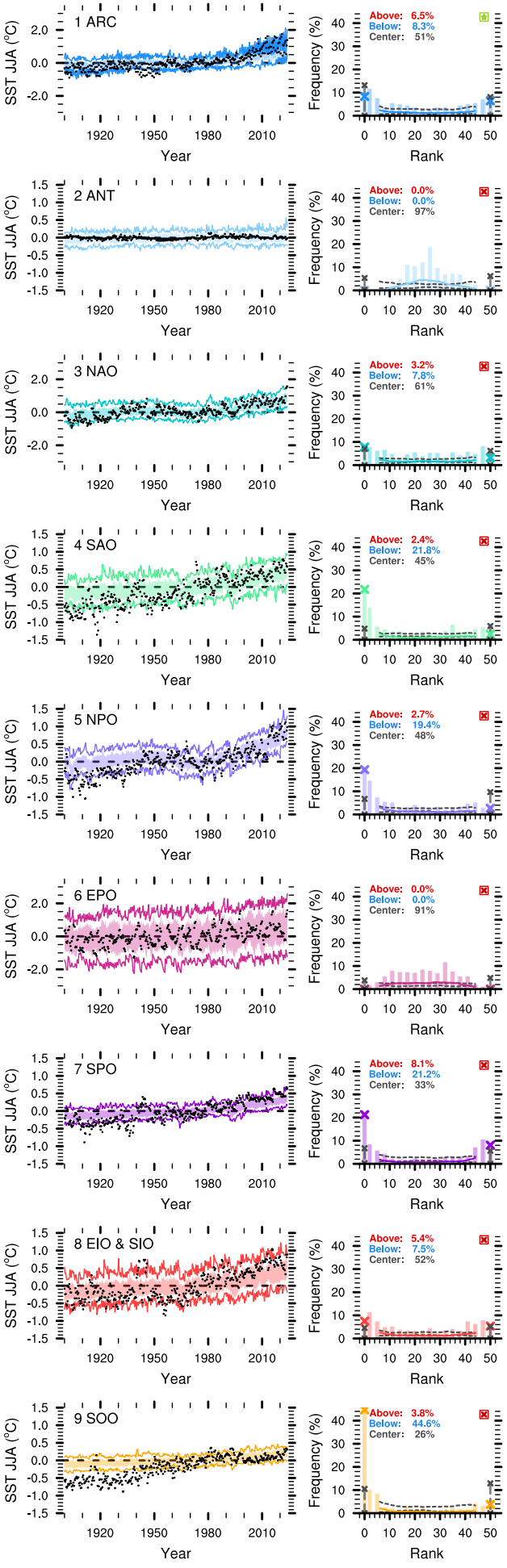
SST JJA Obs outside ensemble spread (%)

8 20

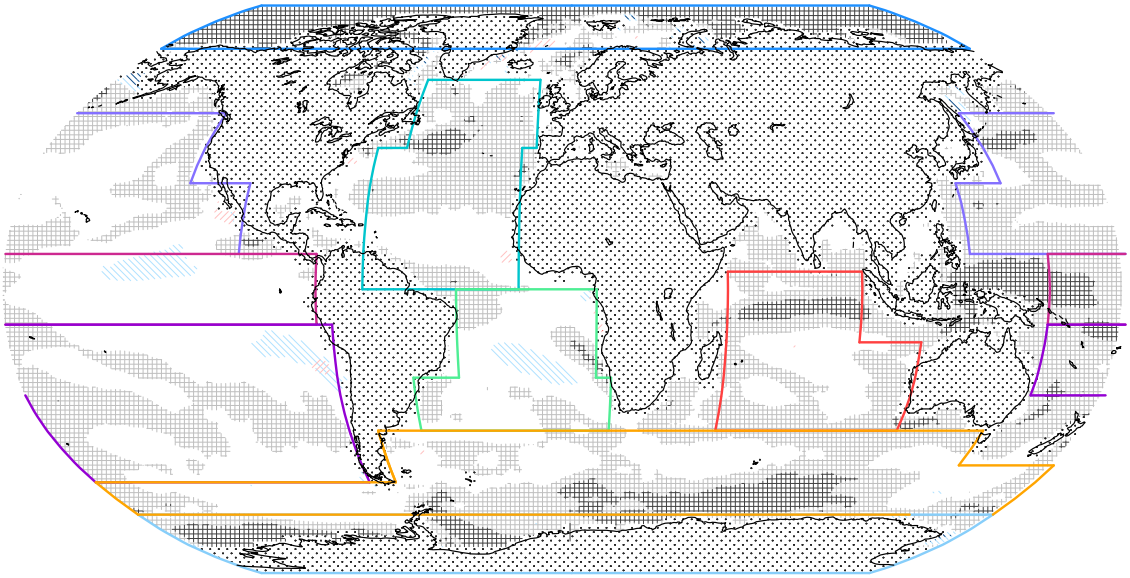
MIROC6 vs ERSSTv5 SST DJF



MIROC6 vs ERSSTv5 SST JJA



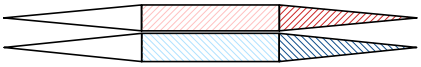
MPI-GE5 vs ERSSTv5 White Area = 43.9 %



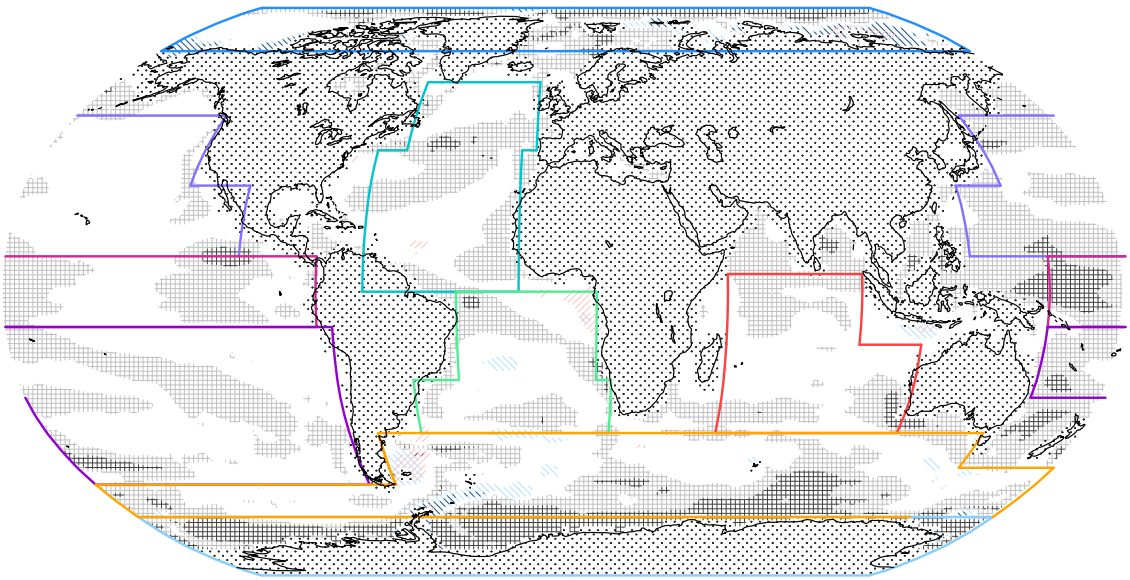
SST DJF Obs inside central 75th percentile (%)



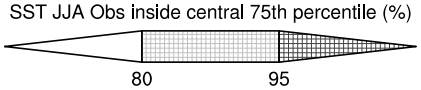
SST DJF Obs outside ensemble spread (%)



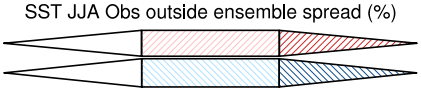
MPI-GE5 vs ERSSTv5 White Area = 51.0 %



SST JJA Obs inside central 75th percentile (%)

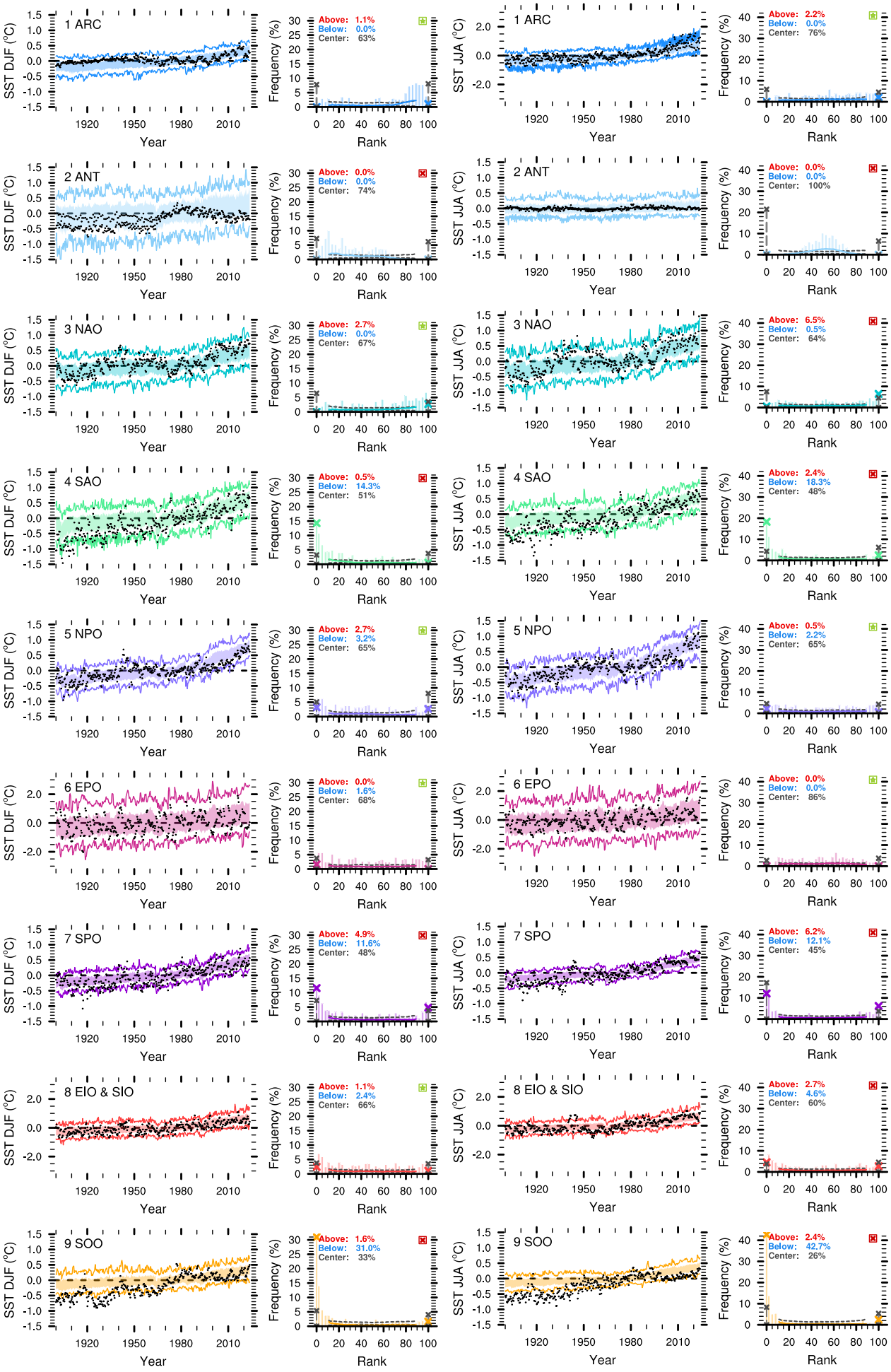


SST JJA Obs outside ensemble spread (%)

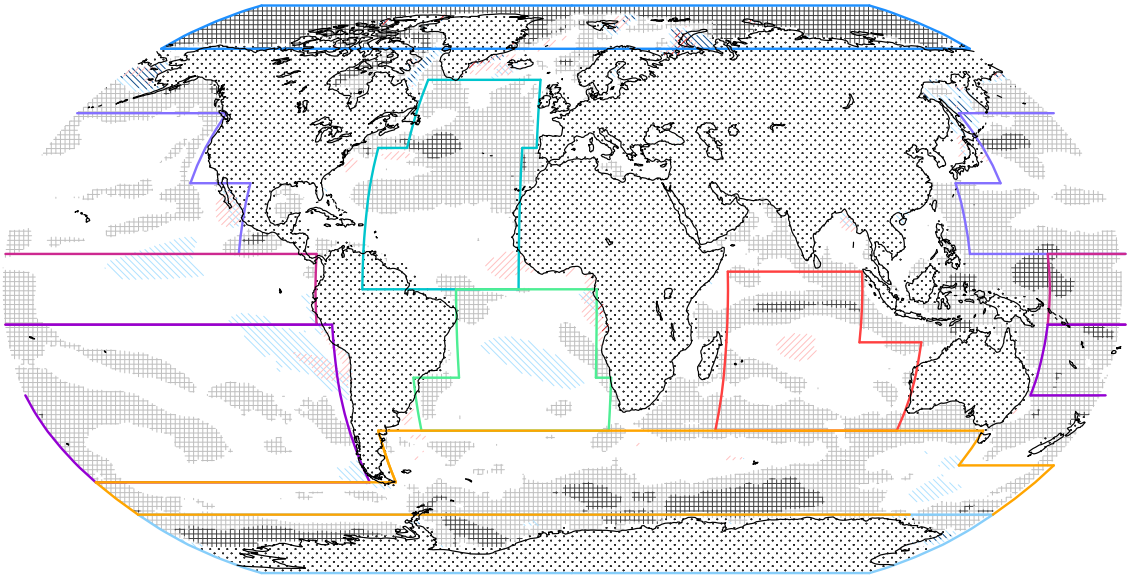


MPI-GE5 vs ERSSTv5 SST DJF

MPI-GE5 vs ERSSTv5 SST JJA



MPI-GE6 vs ERSSTv5 White Area = 49.3 %



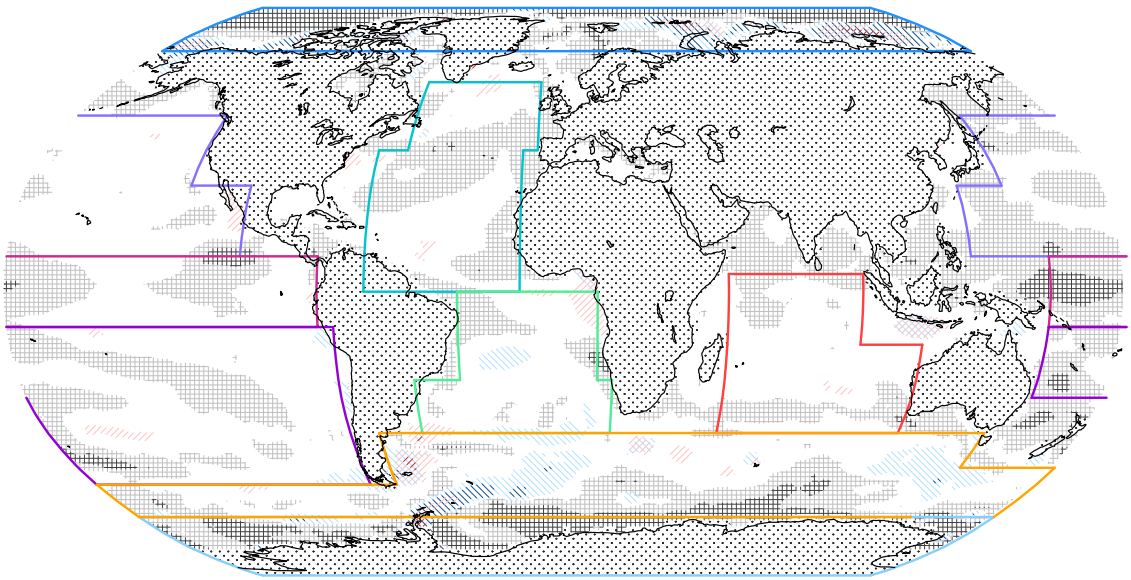
SST DJF Obs inside central 75th percentile (%)

80 95

SST DJF Obs outside ensemble spread (%)

8 20

MPI-GE6 vs ERSSTv5 White Area = 53.0 %



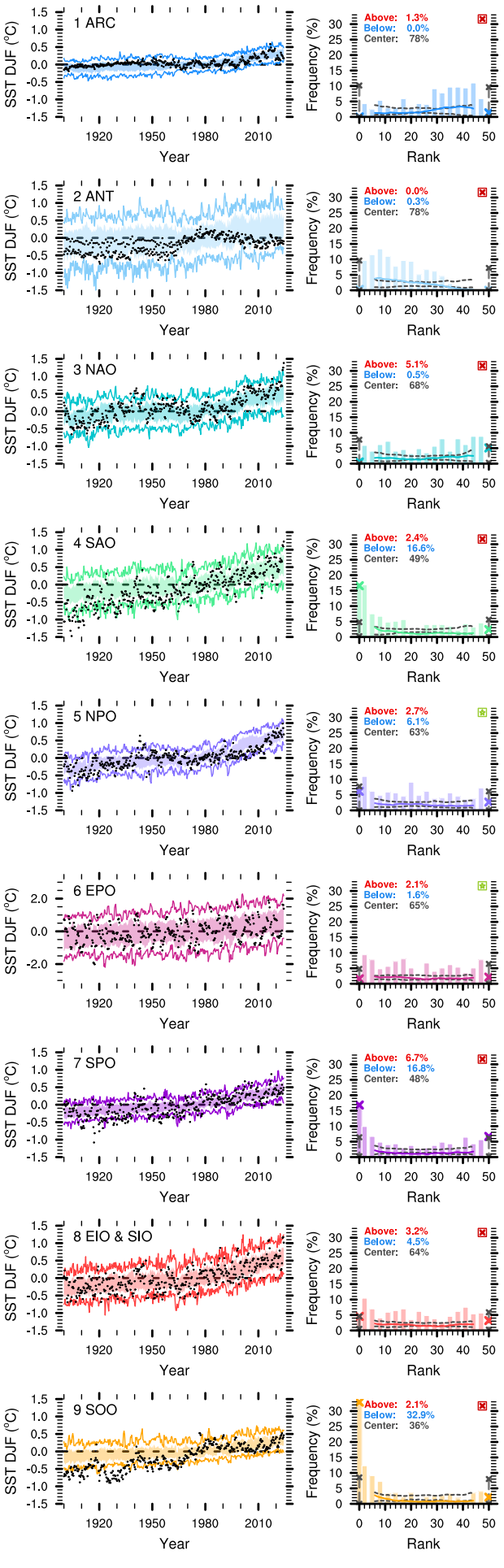
SST JJA Obs inside central 75th percentile (%)

80 95

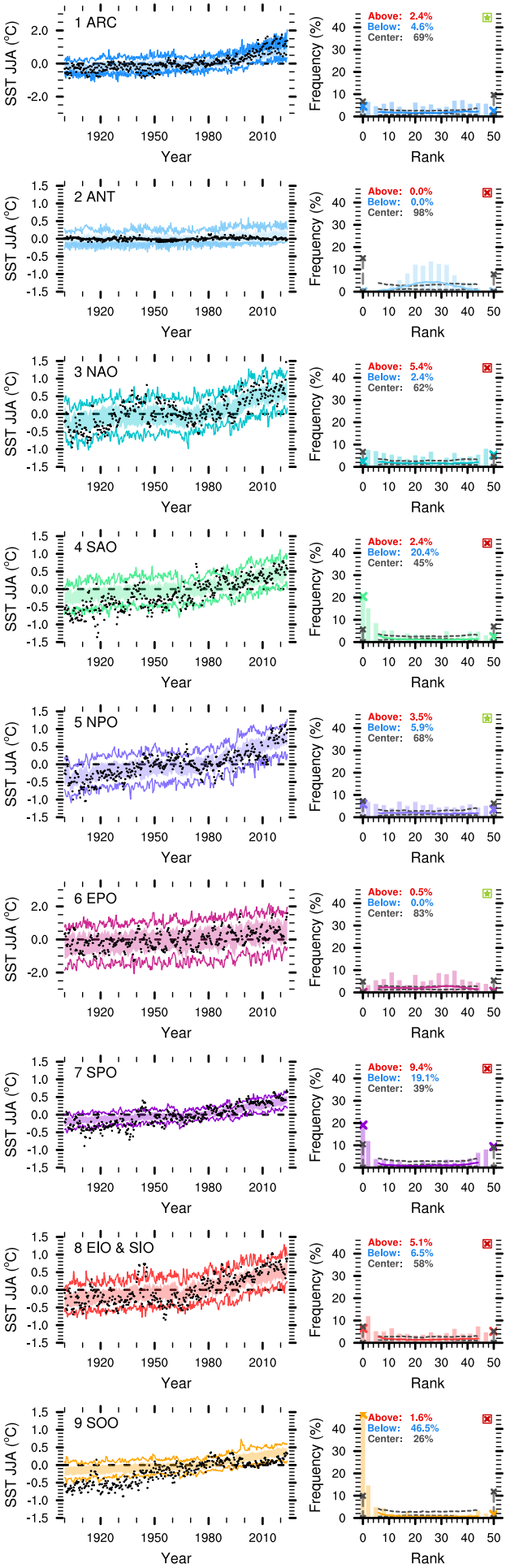
SST JJA Obs outside ensemble spread (%)

8 20

MPI-GE6 vs ERSSTv5 SST DJF



MPI-GE6 vs ERSSTv5 SST JJA



2. Global Warming Levels

This section encloses the information of the specific years in which each SMILE reaches the five global mean warming levels used in this study (Table 1). Warming levels are defined relative to a baseline period of 1850-1899 in each SMILE. Each warming level is computed as the year when the global mean surface temperature (annual) from the SMILE ensemble mean crosses an individual warming level.

SMILE	1°C	1.5°C	2°C	3°C	4°C
ACCESS	2017	2028	2041	2061	2080
CanESM2	2002	2015	2028	2049	2067
CanESM5	1999	2010	2020	2039	2054
CESM-LE	2017	2031	2042	2060	2079
CESM2-LE	2011	2030	2044	2068	2088
CSIROMK3.6	2018	2033	2045	2066	2084
GFDL-ESM2M	2008	2035	2051	2081	n/a
GFDL-SPEAR-MED	2013	2029	2041	2061	2081
MIROC6	2022	2038	2052	2075	2095
MPI-GE5	2008	2029	2046	2071	2095
MPI-GE6	2010	2034	2048	2071	2091

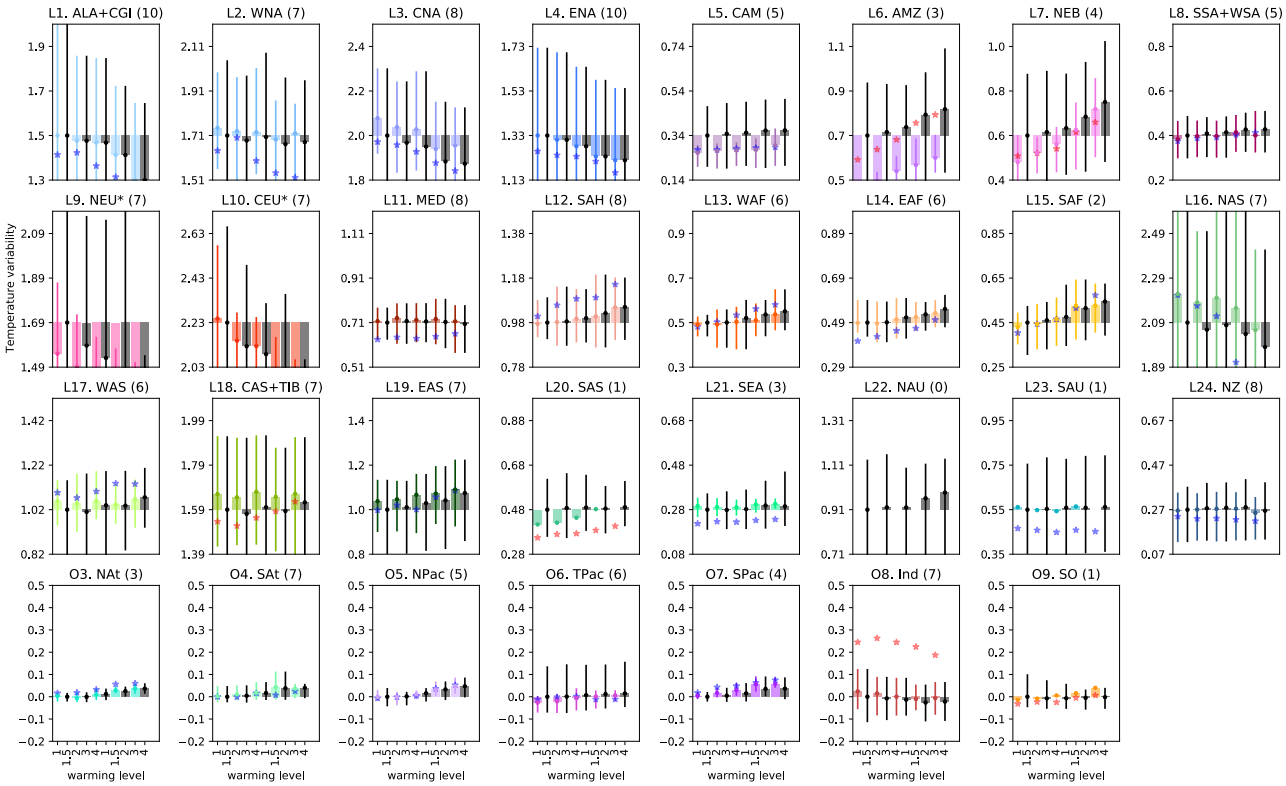
Supplementary Table 1: Warming levels for each single model initial-condition large ensembles (SMILE)

3. Performance Assessment of Constrained and Unconstrained Projections

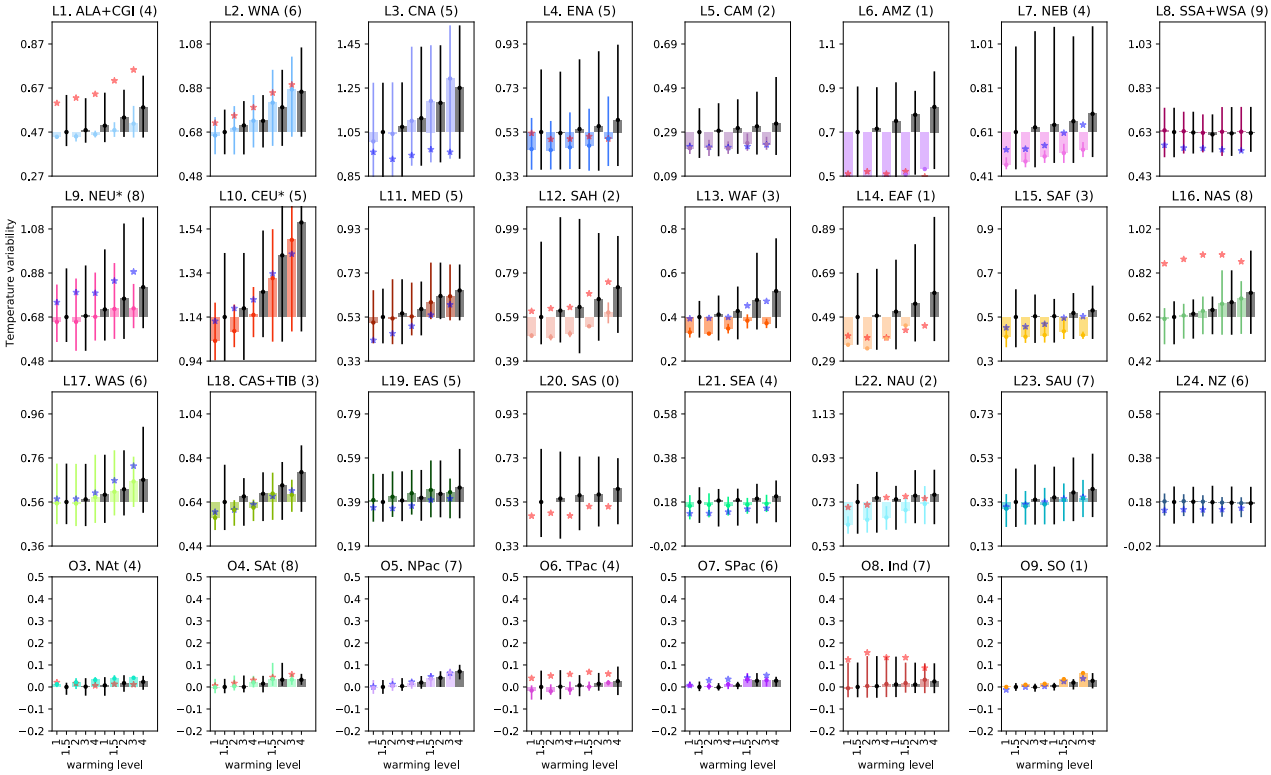
To assess whether the constrained and unconstrained ensemble projections based on historical performance are more adequate to capture future changes is challenging, as we do not have observations of the future. To partially overcome this, we can compare the projections from the most realistic model over key poorly represented areas in the historical period, CESM2-LE, against projections from the full and constrained ensembles, excluding CESM2-LE (e.g., Main Fig. 8 and Supplementary Figs. 3.1 and 3.2)

CESM2-LE is one of the models closest to historical observations, particularly over the most poorly represented regions, including Australia and South-East Asia (L21-23), South America (L6-8), and Africa (L12-16; the other being CESM-LE), with 15 "good performances" out of 22 across both regions and seasons. Removing CESM2-LE, we recalculate the full and constrained ensemble for future projections and assess how closely it resembles the CESM2-LE future projections, distinguishing between the cases where CESM2-LE adequately captures observations from those where it does not. We explore this comparison of the full vs. constrained projections against the projections of the 'good performing model' CESM2 (Supplementary Figs. 3.1 and 3.2, which replicate Main Figs. 3 and 4, respectively, but now excluding CESM2.

These results highlight how the new constrained (excluding CESM2) and the full ensemble (also excluding CESM2) compare with CESM2. In some cases, the constrained ensemble of adequately performing models is closer to CESM2 when CESM2 is a good performer (e.g., L5 or O7 in Fig. 3.1). However, in other cases it is not (e.g., L3 or L13 in Fig. 3.2). This is likely caused by future variability being more closely linked to past variability performance in some cases. In contrast, in other cases climatic and environmental changes intrinsic to each model and region may affect the changes in the temperature variability distribution differently across different models, even for those that show a common and adequate historical representation. Disentangling the drivers of these diverging behaviours for future variability change across historically adequate performing models for key remaining uncertainty hotspots (e.g., the Amazon basin or the polar Oceans) remains a critical knowledge gap.



Supplementary Figure 3.1: Comparison of 'new' constrained ensemble without CESM2 against CESM2 projections for December, January, Feberuary (DJF). Blue stars mark regions for which CESM2 offers an adequate representation of historical variability, red stars represent where it does not, as shown in Main Fig. 2a. Multi-ensemble mean DJF temperature variability defined as the standard deviation over each region at each warming level for both the full ensemble (black) and the constrained ensemble (colour). Results are shown relative to the variability at 1 degree of warming in the full ensemble. Errorbars show the full model spread (i.e. the end members of each ensemble), with the fatter errorbars highlighting the 25th and 75th percentiles. The number of models that accurately represent observed variability are shown in brackets. We choose to exclude the polar oceans from these plots due to issues around the ice edges.



Supplementary Figure 3.2: Comparison of 'new' constrained ensemble without CESM2 against CESM2 projections for June, July, August (JJA). Same as Supplementary Fig. 3.1, but for JJA. Blue stars mark regions for which CESM2 offers an adequate representation of historical variability, red stars represent where it does not, as shown in Main Fig. 2a.