Nonparametric Permutation Procedures for Evaluating Climate Model Accuracy

Joshua P. French[*], Piotr S. Kokoszka[+], Seth McGinnis

University of Colorado Denver, Colorado State University, National Center for Atmospheric Research

November 19, 2020

[*] Partially supported by NSF award 1915277. [+] Partially supported by NSF award 1914882.

Climate model evaluation

- Modern, large-scale climate models are popular for exploring the impact of humans on future climate (and the potential impact of future climate on humans).
- Climate models can be evaluated by comparing their faithfulness in mimicking the overall behavior of observed data.
- A model's deficiencies in describing observed climate will likely be amplified for future predictions and may weaken their usefulness in making decisions.
- We desire to evaluate the accuracy of the state-of-the-art NA-CORDEX climate models (Mearns et al., 2017) in representing the state-of-the-art ERA5 reanalysis data (Copernicus Climate Change Service (C3S), 2017).
- We propose two permutation-based approaches for comparing the models.

Reanalysis background

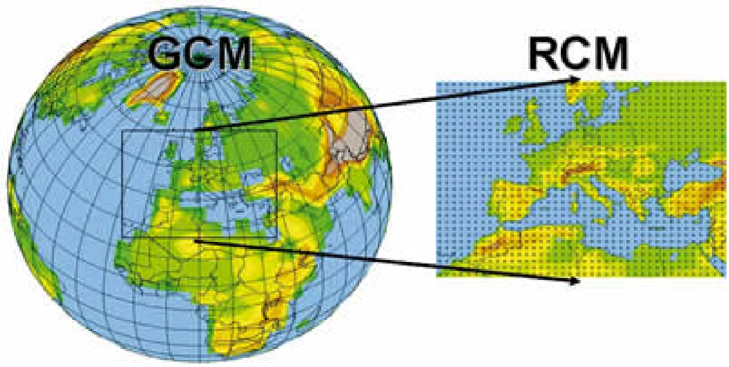
- A climate reanalysis feeds large amounts of observational data into data assimilation models to provide a summary of recent climate for many variables, e.g., surface air temperature, total precipitation, and wind speed.
- The response values are typically provided on a grid over substantial parts of the earth.
- A individual researcher is unlikely to be able to obtain the many data sources used for the product, nor will they be able to process the data using standard computing resources.

ERA5 information

- The ERA5 global reanalysis is the 5th generation of reanalysis produced by the European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020).
- The currently available data stretches from 1979 to approximately the present day.
- The data are available at several spatial and temporal resolutions.
- The data assimilate 74 data sources (European Centre for Medium-Range Weather Forecasts, 2020) using a 4D-Var ensemble data assimilation system.

GCM vs RCM resolution

- GCMs model environmental factors and climate dynamics on a coarse scale ($\approx$ 150-200 km spatial resolution).
- RCMs use information from the GCMs to make predictions on a much finer scale.

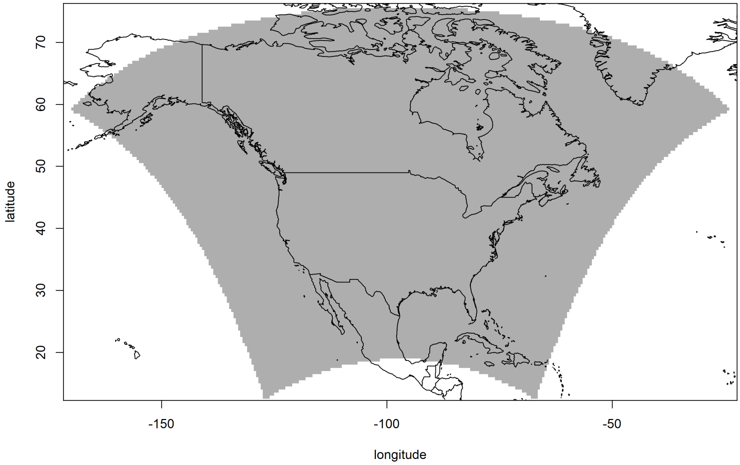


*From the World Meteorological Organization

NA-CORDEX information

- The North American Coordinated Regional Downscaling Experiment (NA-CORDEX) is focused on downscaling climate output in the North American domain using boundary conditions from the CMIP5 archive (Hurrell et al., 2011).
- Data are available from 1950-2100 at fine temporal and spatial resolutions.
  - The NARCCAP data have been bias-corrected using the Multivariate Bias Correction (Cannon, 2018)using the Daymet data product as a reference (Thornton et al., 2018).
- There are combinations of 6 different GCMs to provide the boundary conditions for 7 different RCMs under 2 sets of future conditions, though not all combinations are currently available.

NA-CORDEX domain



Comparing the ERA5 and NA-CORDEX data

- We restrict our use of the ERA5 data to the same subdomain as the NA-CORDEX data and land masses.
- Both data sets are available on the same 0.5° spatial grid.
- We utilize monthly average of maximum daily 2 meter temperature.
  - Other variables may not be reliable to compare.
- The historical period for the NA-CORDEX data runs from 1950-2005, while the ERA5 data we consider runs from 1979-present day.
  - We restrict our analysis to monthly temperature for the complete years 1979-2004 due to data problems in 2005.
- There is a single realization of the ERA5 data available and 15 realizations of NA-CORDEX data with these characteristics.

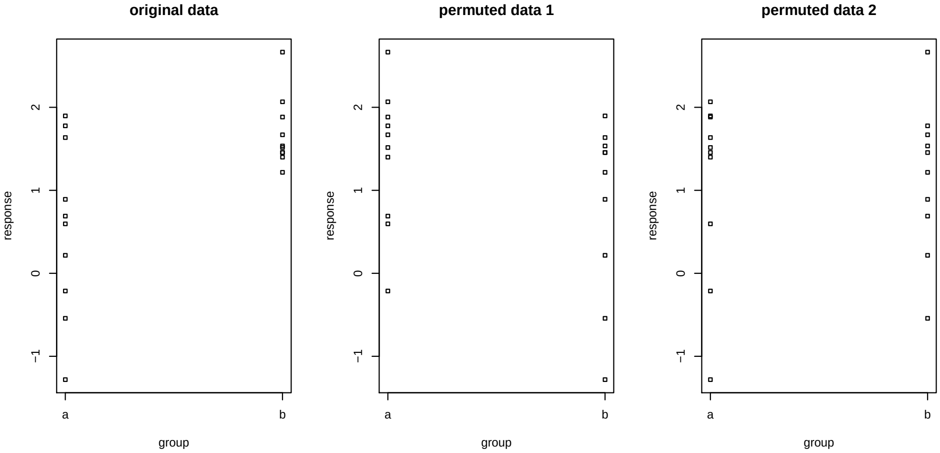
Tests we consider

- We consider a distributional equality test $H_0 : F^R(\mathbf{s}) = F^M(\mathbf{s})$ versus $H_a : F^R(\mathbf{s}) \neq F^M(\mathbf{s})$, where F denotes the distribution of the data at location $\mathbf{s}$.
  - R and M refer to the "Reanalysis data" and "Model" data, respectively.
  - The distributional test can't tell us HOW the distributions differ.
- Generically, let θ denote a well-defined characteristic of F.
  - $\theta(\mathbf{s})$ indicates that the characteristic is for location $\mathbf{s}$.
- To assess HOW the distributions differ, we test $H_0 : \theta^R(\mathbf{s}) = \theta^M(\mathbf{s})$ versus $H_a : \theta^R(\mathbf{s}) \neq \theta^M(\mathbf{s})$.

Standard permutation tests

- Permutation tests (Fisher, 1935) are a standard, nonparametric procedure for testing hypotheses while making minimal assumptions.
  - A test statistic is computed for the original data and for many permutations under the null hypothesis that the data are exchangeable across groups.
- In this context, a standard permutation test permutes the random fields across the reanalysis and climate model groups.
- A limitation in our context is that while there are 16! permutations of the data indices, there may only be 16 unique combinations of the data leading to different test statistics
  - For example, the sample mean of the model group will not change if the 15 models are permuted.
- Testing at a significance level of 0.10, the test statistic for the observed data will have to be more extreme than every test statistic resulting from a data permutation in order to conclude statistical significance.

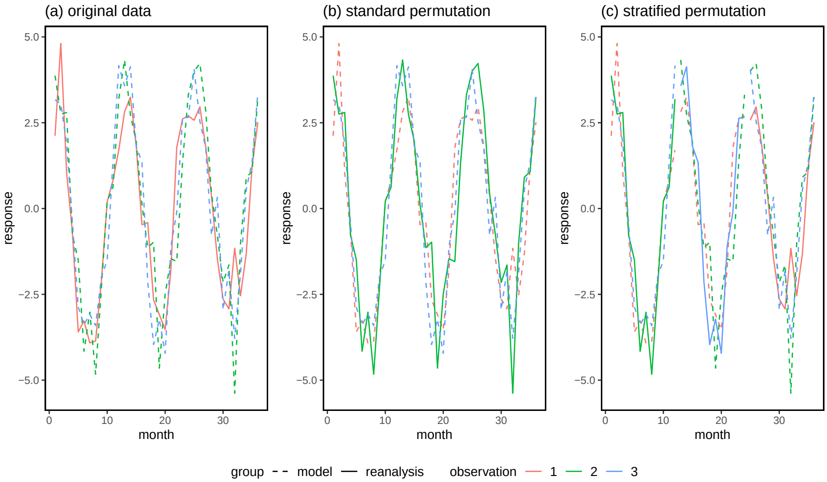
Permutation example



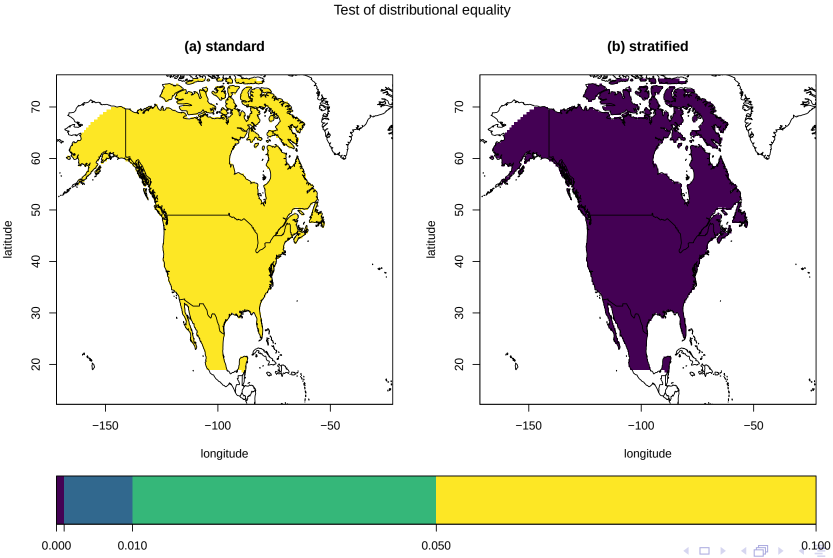
Stratified permutation tests

- Matchett et al. (2015) introduced a general stratified permutation test to test whether rare stressors had impact on certain animal species after controlling for certain covariates.
  - After stratifying their data, responses within each strata were exchangeable under the null hypothesis.
- We permute whole spatio-temporal random fields within the same year across climate models but permute each year independently.
  - Spatio-temporal dependence is preserved within each year.
  - Potential non-stationarity between years is respected.
  - Wilks (1997): "Simultaneous application of the same resampling patterns to all dimensions of the data vectors will yield resampled statistics reflecting the cross correlations in the underlying data, without the necessity of explicitly modeling those cross correlations."
- We have $16^{26} > 2 \times 2^{31}$ effective permutations, substantially increasing power.

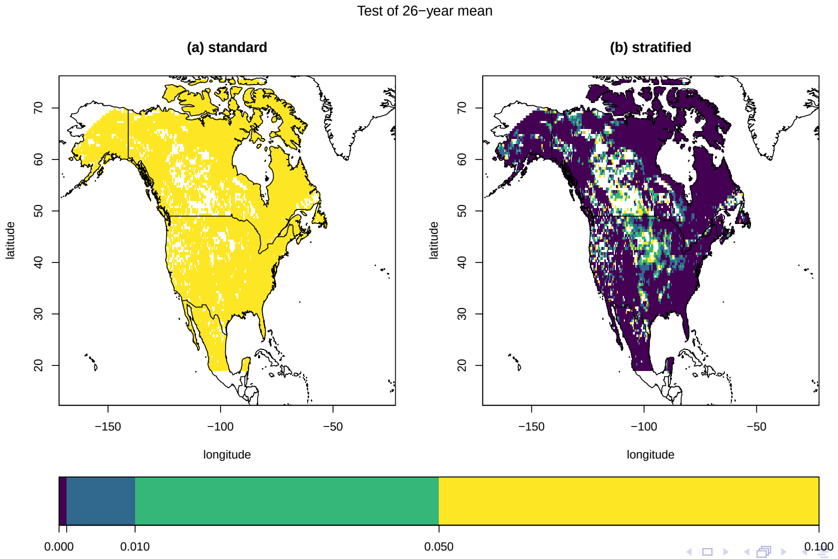
Permutation examples



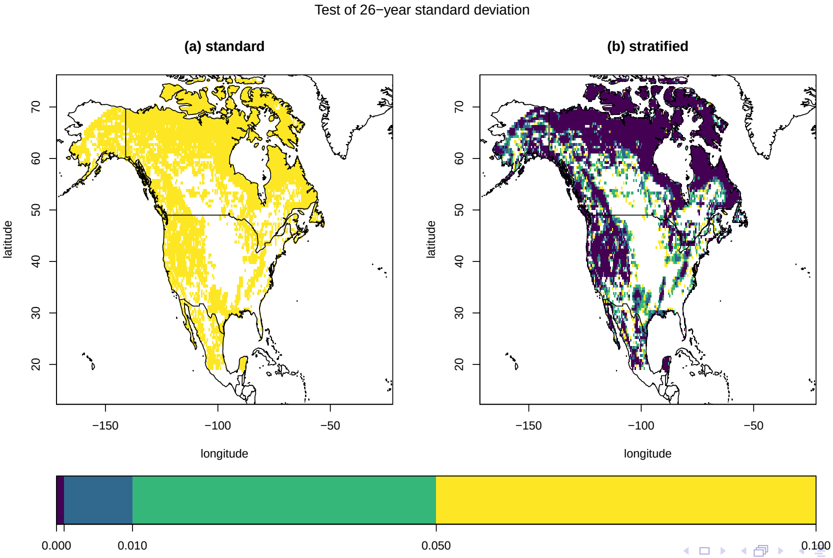
Test of distributional equality



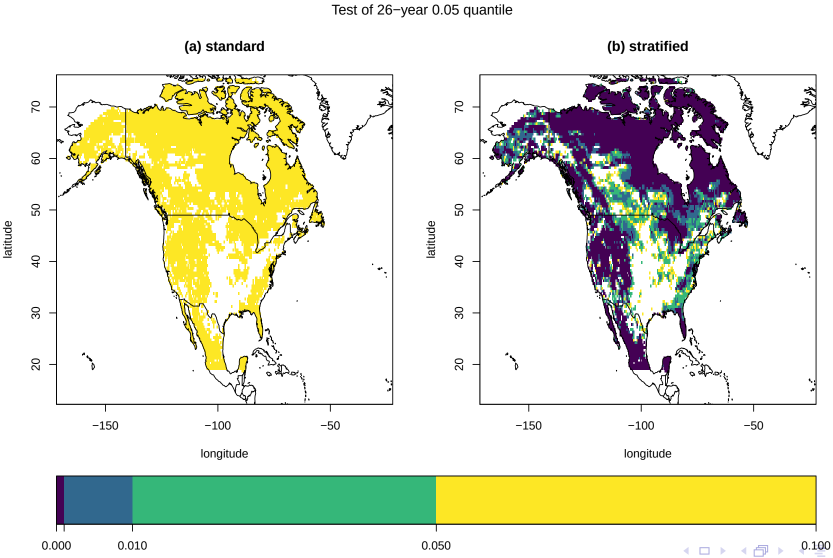
Tests of 26-year mean



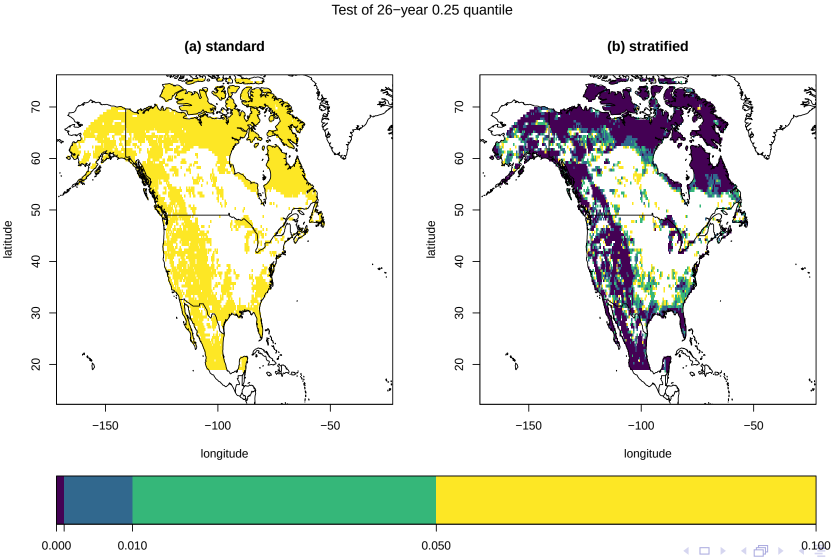
Tests of 26-year standard deviation



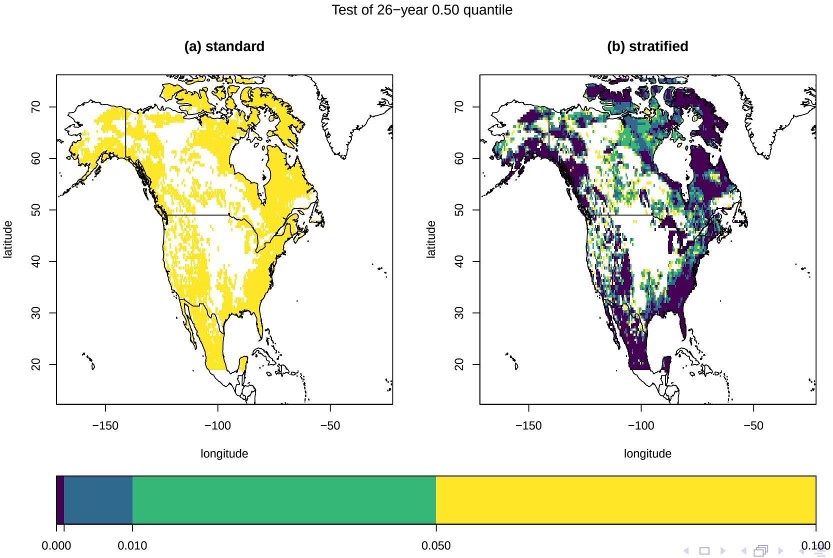
Tests of 26-year 0.05 quantile



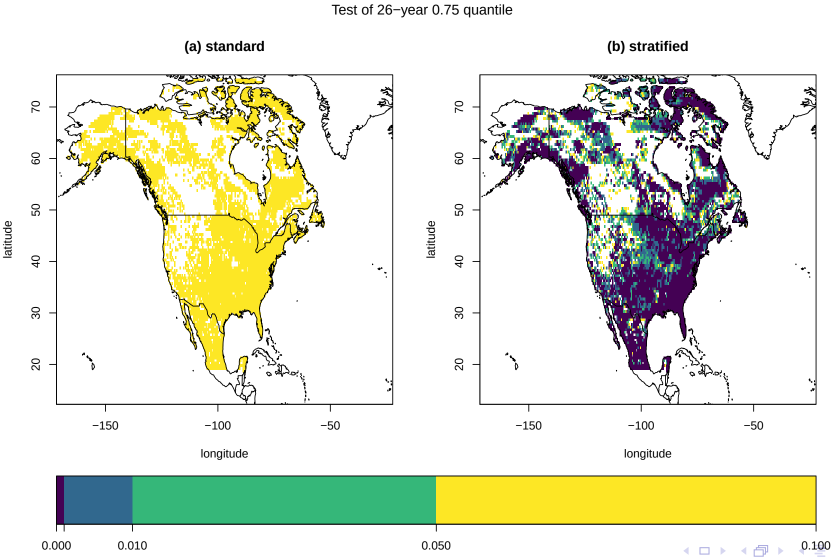
Tests of 26-year 0.25 quantile



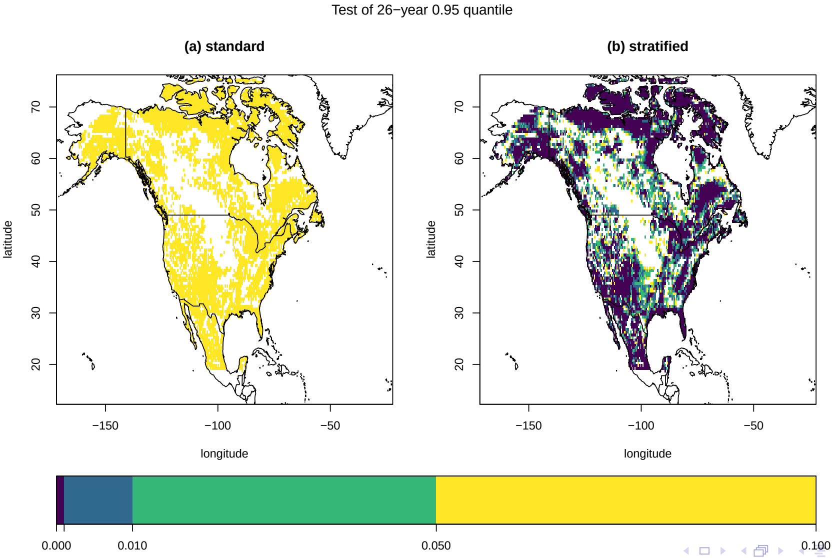
Tests of 26-year 0.50 quantile



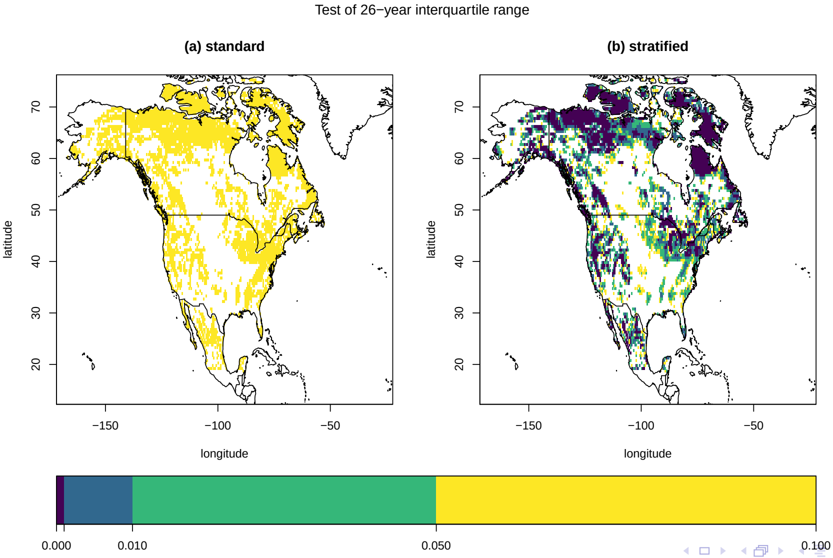
Tests of 26-year 0.75 quantile



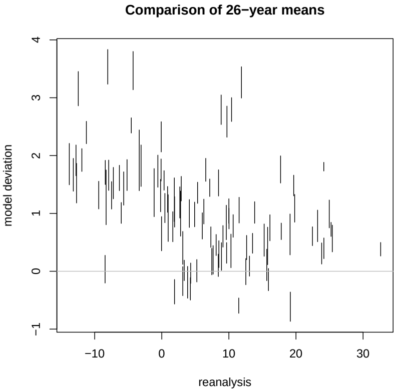
Tests of 26-year 0.95 quantile

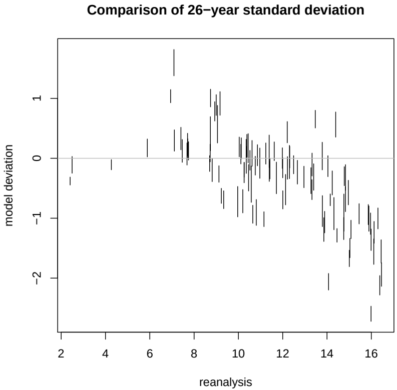


Tests of 26-year interquartile range

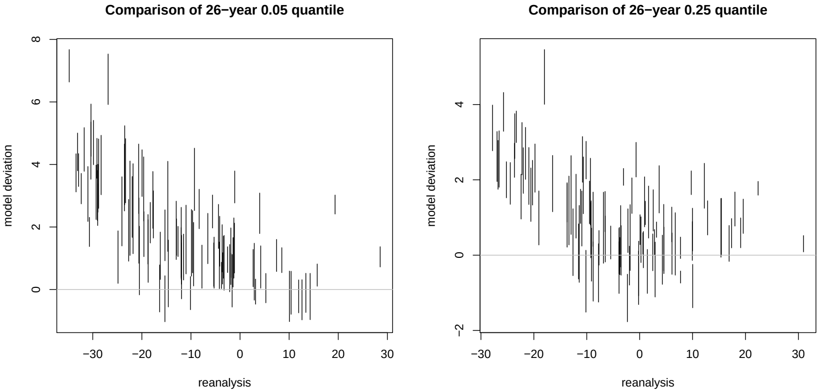


Deviations of 26-year means and standard deviations

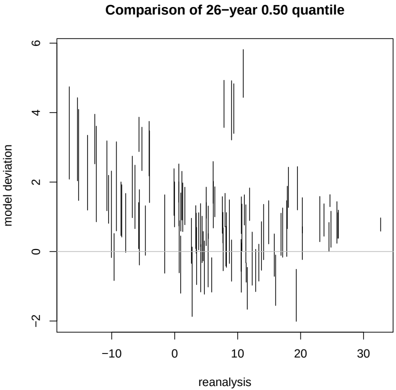

Comparison of 26−year means

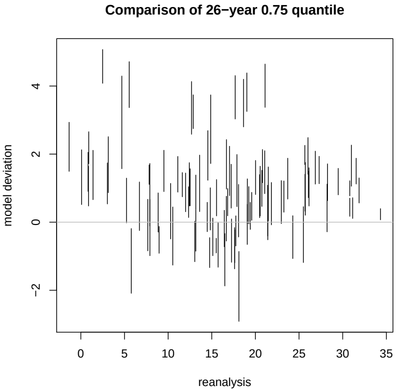

Comparison of 26−year standard deviation

Deviations of 26-year 0.05 and 0.25 quantiles

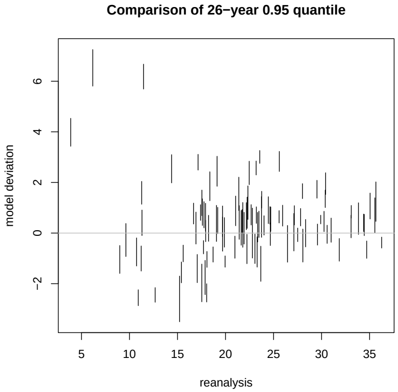


Deviations of 26-year 0.50 and 0.75 quantiles

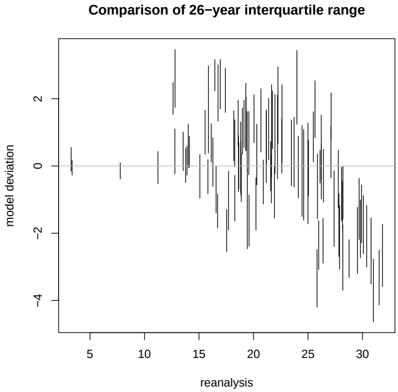




Deviations of 26-year 0.95 quantiles and IQRs



Comparison of 26−year 0.95 quantile



Comparison of 26−year interquartile range

Discussion

- The distribution and various characteristics of the reanalysis and climate model distributions are statistically different throughout much of the spatial domain.
- No spatial tests adjusted for multiple comparisons.
  - Off-the-shelf multiple comparisons adjustments are difficult because the precision of the p-value is directly related to the number of simulations and adding a decimal place of precision (even without accounting for Monte Carlo uncertainty) requires 10 times the computational cost.
- We might want to compare the distributional equality test that we proposed with the the T-metric proposed by Tian et al. (2017), which was custom-made to compare climate model simulations to each other.
- Angélil et al. (2016) recommend using multiple reanalyses data sets when performing climate model evaluation, so perhaps we should augment our current analysis by including reanalysis data from NASA's MERRA2 program and possibly the NCEP Climate Forecast System Reanalysis and the Japanese 55-year Reanalysis.

Discussion

- RCM outputs forced by the same GCM boundary conditions are NOT independent.
  - Rerun using only one RCM per GCM.
  - Lose some power.
- Better way to quantify the functional discrepancy between the datum?

Bibliography I

Angélil, O., Perkins-Kirkpatrick, S., Alexander, L. V., Stone, D., Donat, M. G., Wehner, M., Shiogama, H., Ciavarella, A., and Christidis, N. (2016). Comparing regional precipitation and temperature extremes in climate model and reanalysis products. *Weather and Climate Extremes*, 13:35 – 43.

Cannon, A. (2018). Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables. *Climate Dynamics*, 50:31–49.

Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS). Accessed March 10, 2020.

European Centre for Medium-Range Weather Forecasts (2020).

Fisher, R. A. (1935). *Design of Experiments*. Oliver and Boyd, Edinburgh.

Bibliography II

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*.

Hurrell, J., Visbeck, M., and Pirani, P. (2011). WCRP coupled model intercomparison project - Phase 5. *CLIVAR Exchanges Newsletter*, 15(56).

Matchett, J. R., Stark, P. B., Ostoja, S. M., Knapp, R. A., McKenny, H. C., Brooks, M. L., Langford, W. T., Joppa, L. N., and Berlow, E. L. (2015). Detecting the influence of rare stressors on rare species in Yosemite National Park using a novel stratified permutation test. *Scientific Reports*, 5(1):1–12.

Mearns, L. O. et al. (2017). The NA-CORDEX dataset, version 1.0. NCAR Climate Data Gateway, Boulder CO. Accessed January 27, 2020.

Bibliography III

Thornton, M., Thornton, P., Wei, Y., Mayer, B., Cook, R., , and Vose, R. (2018). Daymet: Monthly Climate Summaries on a 1-km Grid for North America, Version 3. ORNL DAAC, Oak Ridge, Tennessee, USA.

Tian, B., Lee, H., Waliser, D. E., Ferraro, R., Kim, J., Case, J., Iguchi, T., Kemp, E., Wu, D., Putman, W., et al. (2017). Development of a model performance metric and its application to assess summer precipitation over the us great plains in downscaled climate simulations. *Journal of Hydrometeorology*, 18(10):2781–2799.

Wilks, D. S. (1997). Resampling hypothesis tests for autocorrelated fields. *Journal of Climate*, 10(1):65–82.