Dealing with multiple testing problems in spatial trend analysis

José Cortés

Department of Geography, Friedrich Schiller University, Jena, Germany

International Summer School on Geospatial Data Science with R
25 August - 1 September 2019

Introduction

- We have a set of hypotheses that we want to test simultaneously.
- We think we can test each hypothesis separately, using some level of significance α.
- **But wait!** There is a chance that we might find at least one significant result just by chance *although our null hypothesis is true (i.e. false positive tests).*
- When performing a large number of statistical tests, some null hypotheses will be rejected by chance alone ($p - value < \alpha$) even if they are true.

Motivating example

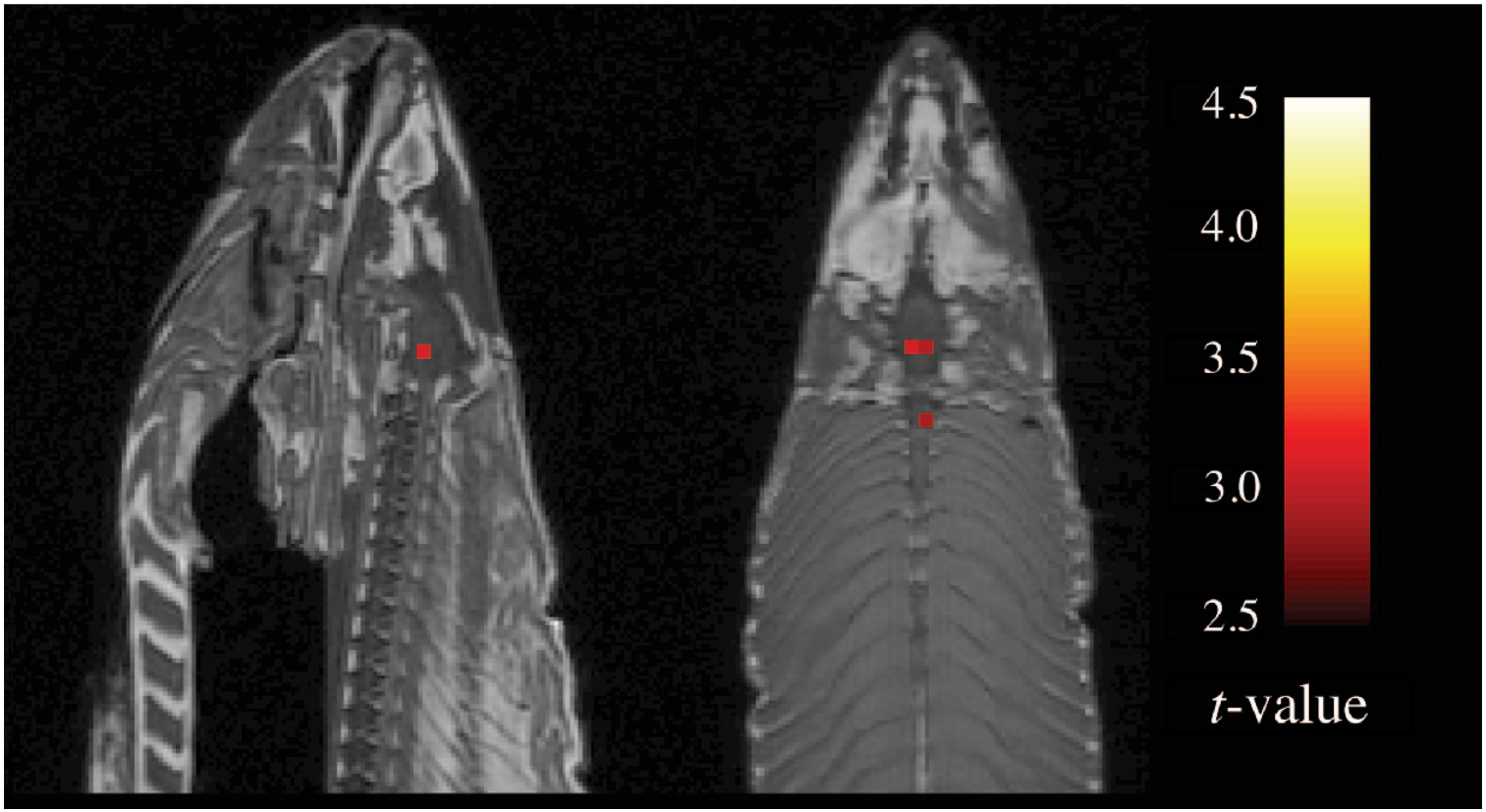


Figure: Sagittal and axial images of significant brain voxels in the task > rest contrast. The parameters for this comparison were t(131) > 3.15, p(uncorrected) < 0.001, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal column.
`https://www.semanticscholar.org/paper/Neural-correlates-of-interspecies-perspective-in-an-Bennett-Miller/3ace3864cc5ada47b31a74d8fea91edb48bc019d`

Introduction

- What's the probability of (erroneously) observing at least one significant result if all null hypotheses are true?

$$\begin{aligned} P(\text{at least one significant result}) &= 1\text{–}P(\text{no significant results}) \\ &= 1\text{–}(1-\alpha)^{\#\text{ of hypotheses to test}} \end{aligned}$$

- This probability is referred to as the Familywise Error Rate (FWER) of a family of hypothesis tests.

Introduction

Example

What is the probability of observing at least one significant result, if all null hypotheses are true, when testing 20 hypotheses with a significance level of 0.05?

$$\begin{aligned} P(\text{at least one significant result}) &= 1\text{–}(1\text{–}0.05)^{20} \\ &\sim 0.64 \end{aligned}$$

So, we have a 64% chance of observing at least one significant result, even if all of the null hypotheses are true. This is our familywise error rate.

Familywise Error Rate

What happens when we increase the number of hypothesis tests?

Overview

- Assume H_0 is true, the probability that we reject H_0 erroneously is α. We refer to false rejections as false positive tests.
- When performing many tests, what is the probability of erroneously rejecting at least one null hypothesis when all null hypotheses are in fact true?
- We call this probability the Familywise Error Rate (FWER) of a family of hypothesis tests.
- (How) Can we guarantee that the probability of obtaining ≥ 1 false positive tests among a family of m tests is less than or equal to a desired level?
- **Alternative** Control the proportion of false positives $(\frac{fp}{fp+tp})$

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M
- Compare $p_{(M-1)}$ to u_{M-1} and so on

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M
- Compare $p_{(M-1)}$ to u_{M-1} and so on
- The first $p_{(i)} \leq u_i$ means we declare $p_{(1)}, \ldots, p_{(i)}$ significant

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M
- Compare $p_{(M-1)}$ to u_{M-1} and so on
- The first $p_{(i)} \leq u_i$ means we declare $p_{(1)}, \ldots, p_{(i)}$ significant

Step-down methods

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M
- Compare $p_{(M-1)}$ to u_{M-1} and so on
- The first $p_{(i)} \leq u_i$ means we declare $p_{(1)}, \ldots, p_{(i)}$ significant

Step-down methods

- Start with $p_{(1)}$ and compare it to u_1

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M
- Compare $p_{(M-1)}$ to u_{M-1} and so on
- The first $p_{(i)} \leq u_i$ means we declare $p_{(1)}, \ldots, p_{(i)}$ significant

Step-down methods

- Start with $p_{(1)}$ and compare it to u_1
- Compare $p_{(2)}$ to u_2 and so on

Bonferroni and related methods

Compare the (ordered) p-values, $p_{(i)}$, (sequentially) to $p_{(i)} < u_i$

Step-up methods

- Start with $p_{(M)}$ and compare it to u_M
- Compare $p_{(M-1)}$ to u_{M-1} and so on
- The first $p_{(i)} \leq u_i$ means we declare $p_{(1)}, \ldots, p_{(i)}$ significant

Step-down methods

- Start with $p_{(1)}$ and compare it to u_1
- Compare $p_{(2)}$ to u_2 and so on
- The first $p_{(i)} > u_i$ means we declare $p_{(1)}, \ldots, p_{(i-1)}$ significant

Bonferroni and related methods

Method	u_i	Control of FWER
Bonferroni	α/n	Strong
Walker	$1-(1-\alpha)^{(1/n)}$	Strong
Holm (step-up)	$\alpha(1/(n-i+1))$	Strong
Hochberg (step-down)	$\alpha(1/(n-i+1))$	Strong
BH (step-up)	$(i/n)\alpha$	Weak
BY (step-up)	$i\alpha/n\sum_{j=1}^{n}(1/j)$	Weak

Bonferroni and related methods

Example with five p-values:

i	*p-value*	**Bonferroni**	**Benjamini-Hochberg**
1	.002	$u_i = 0.05/5 = 0.01$ **sig**	$u_i = 1 * 0.05/5 = 0.01$ **sig**
2	.015	$u_i = 0.05/5 = 0.01$ **nonsig**	$u_i = 2 * 0.05/5 = 0.02$ **sig**
3	.022	$u_i = 0.05/5 = 0.01$ **nonsig**	$u_i = 3 * 0.05/5 = 0.03$ **sig**
4	.045	$u_i = 0.05/5 = 0.01$ **nonsig**	$u_i = 4 * 0.05/5 = 0.04$ **nonsig**
5	.1	$u_i = 0.05/5 = 0.01$ **nonsig**	$u_i = 5 * 0.05/5 = 0.05$ **nonsig**

The Maximum Statistic

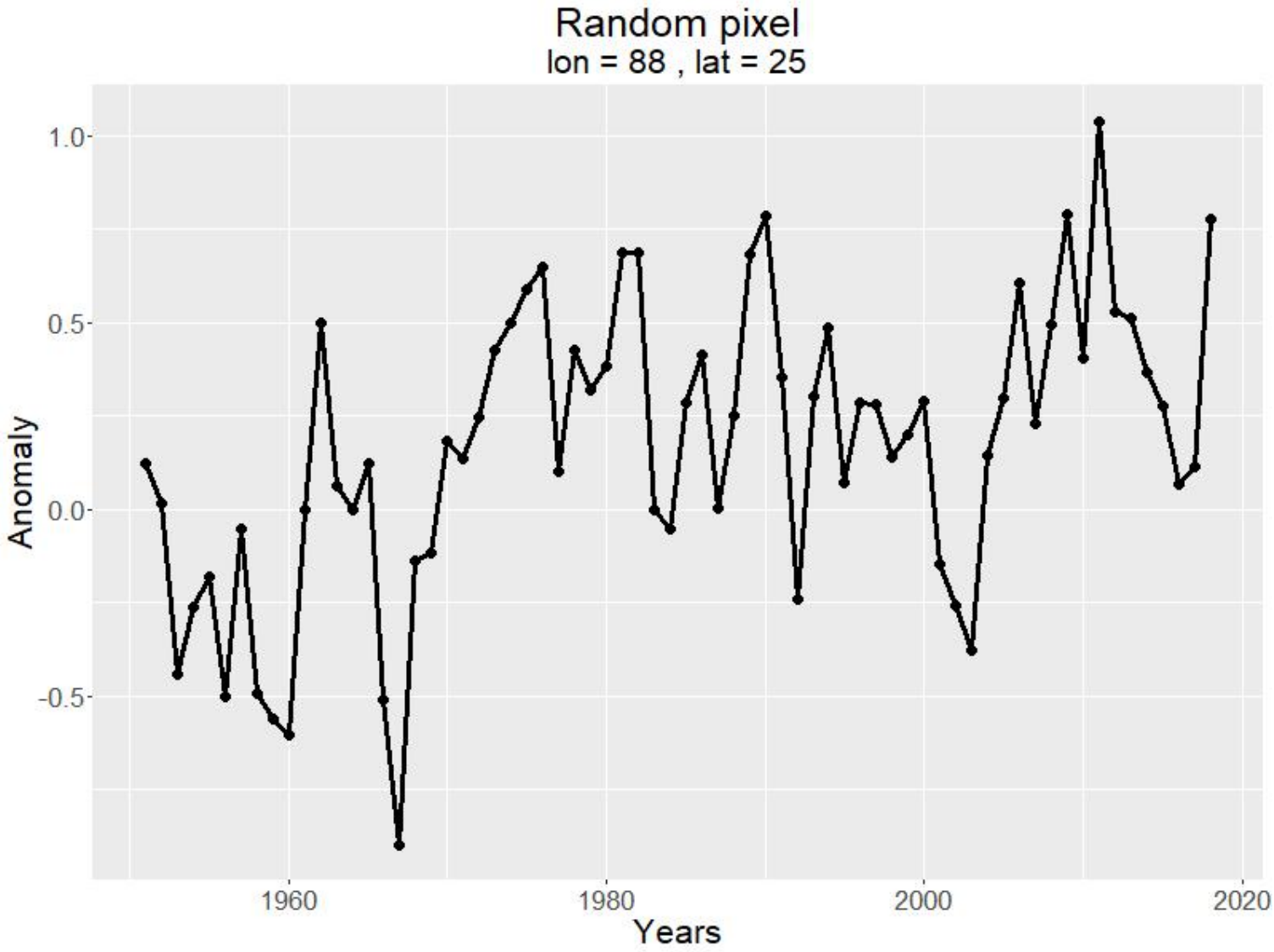
- If the maximum statistic exceeds the threshold then we know that at least one grid cell is significant.
- **Why?** P(at least one significant result) = P(maximum statistic is significant)
- We can derive the distribution of the maximum statistic with permutation methods!
- General idea - permute data N times and recalculate the test statistic for each permutation. These values form the distribution of our test statistic.
- We calculate two maximum statistics

  max T — The maximum test statistic among all test statistics

  max STCS — The maximum supra-threshold cluster size (STCS) is the largest number of adjacent significant pixels.
- Our significance threshold is the $(1^{-}\alpha)\%$ percentile of these distributions.

Case study - temperature trends

Research question Are temperatures increasing or decreasing?



Case study - temperature trends

Test statistic Mann Kendall's S (trend test)

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(x_j - x_i)$$

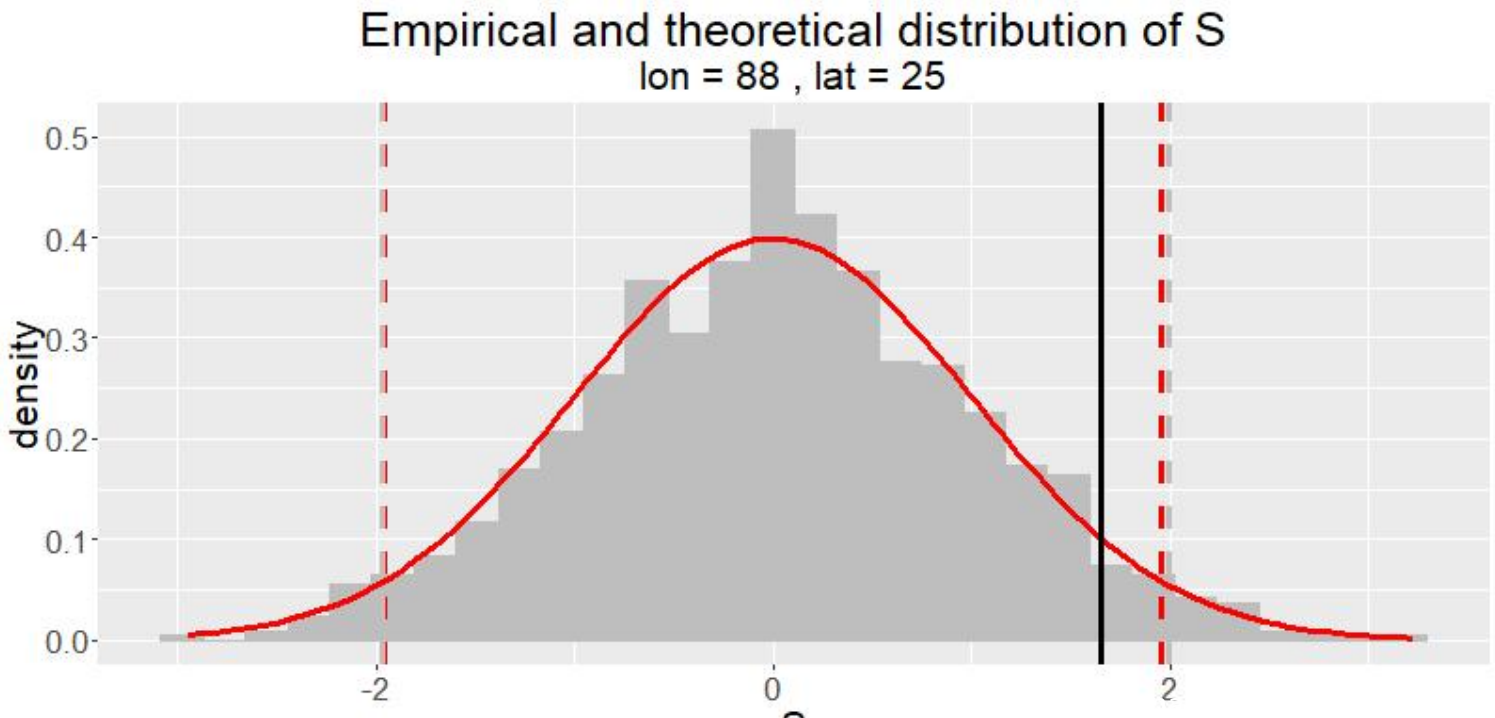
$$var(S) = \frac{n(n-1)(2n+5)}{18}$$

The S statistic can be converted to a standard normal distribution by

$$Z = \begin{cases} \frac{S-1}{\sqrt{var(S)}} & \text{for } S > 0 \\ 0 & \text{for } S = 0 \\ \frac{S+1}{\sqrt{var(S)}} & \text{for } S < 0 \end{cases}$$

Case study - temperature anomalies

- We define two hypothesis
  - $H_0 : Z = 0$ The null hypothesis - there is no change in temperature
  - $H_1 : Z \neq 0$ The alternative hypothesis - there is a change in temperature (either increasing or decreasing)
- Based on our observed Z, we accept or reject H_0
- The decision threshold is set by the distribution of the test statistic under the null hypothesis



Case study - temperature anomalies

temperature movie

Case study - temperature anomalies

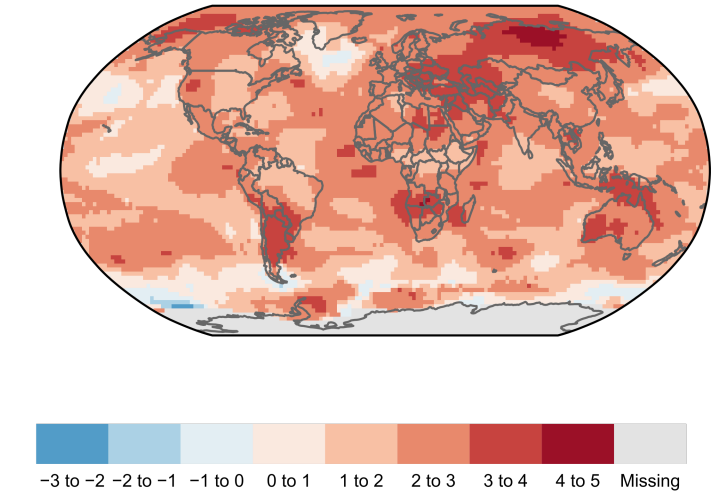
- Analysis of gridded spatio-temporal data $\rightarrow$ Fit a statistical model at each grid cell.
- Multiple hypothesis testing $\rightarrow$ Uncontrolled amount of false positives.
- Spatial correlation $\rightarrow$ Grouped false positives.
- Control of false positives $\rightarrow$ Familywise Error Rate (FWER).

Case study - temperature trends

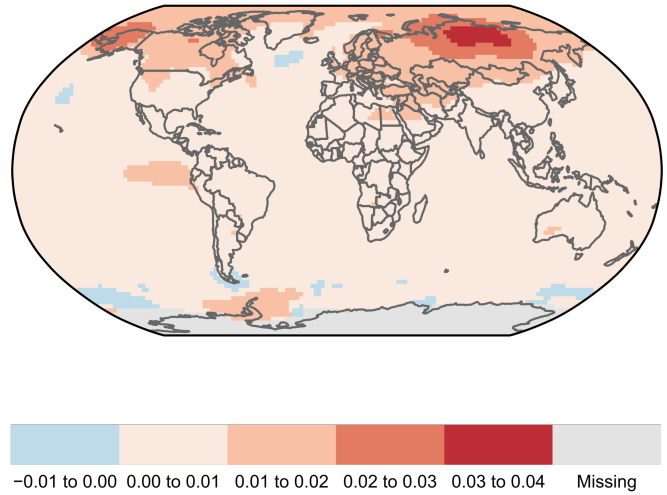
- NASA GISS Surface Temperature Analysis (GISTEMP) version 5.
- Aggregated yearly from 1951 to 2018 (68 years).
- Spatial resolution of $2°x2°$ (14,295 grid cells).
- Mann-Kendall trend test at $\alpha = 0.05$
- von Storch's correction for temporal autocorrelation.

Exploratory maps

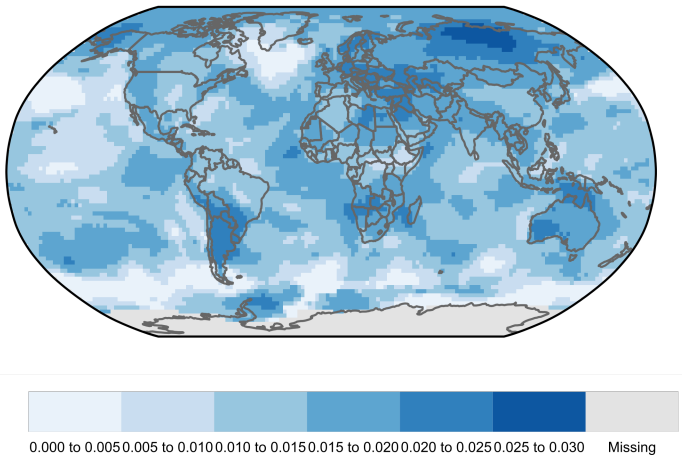
MK Z stat



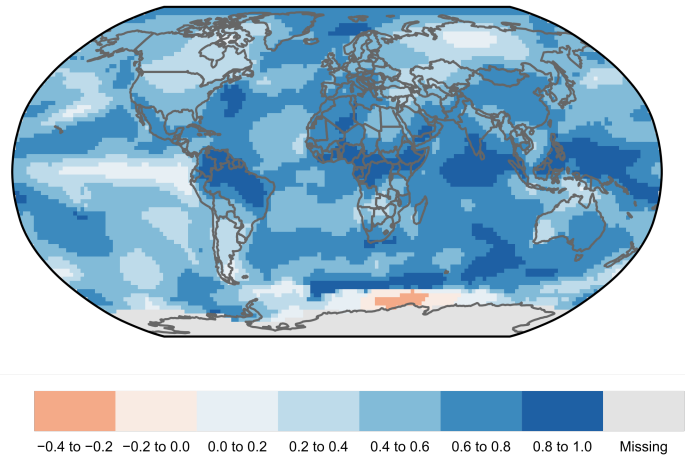
Sen slope



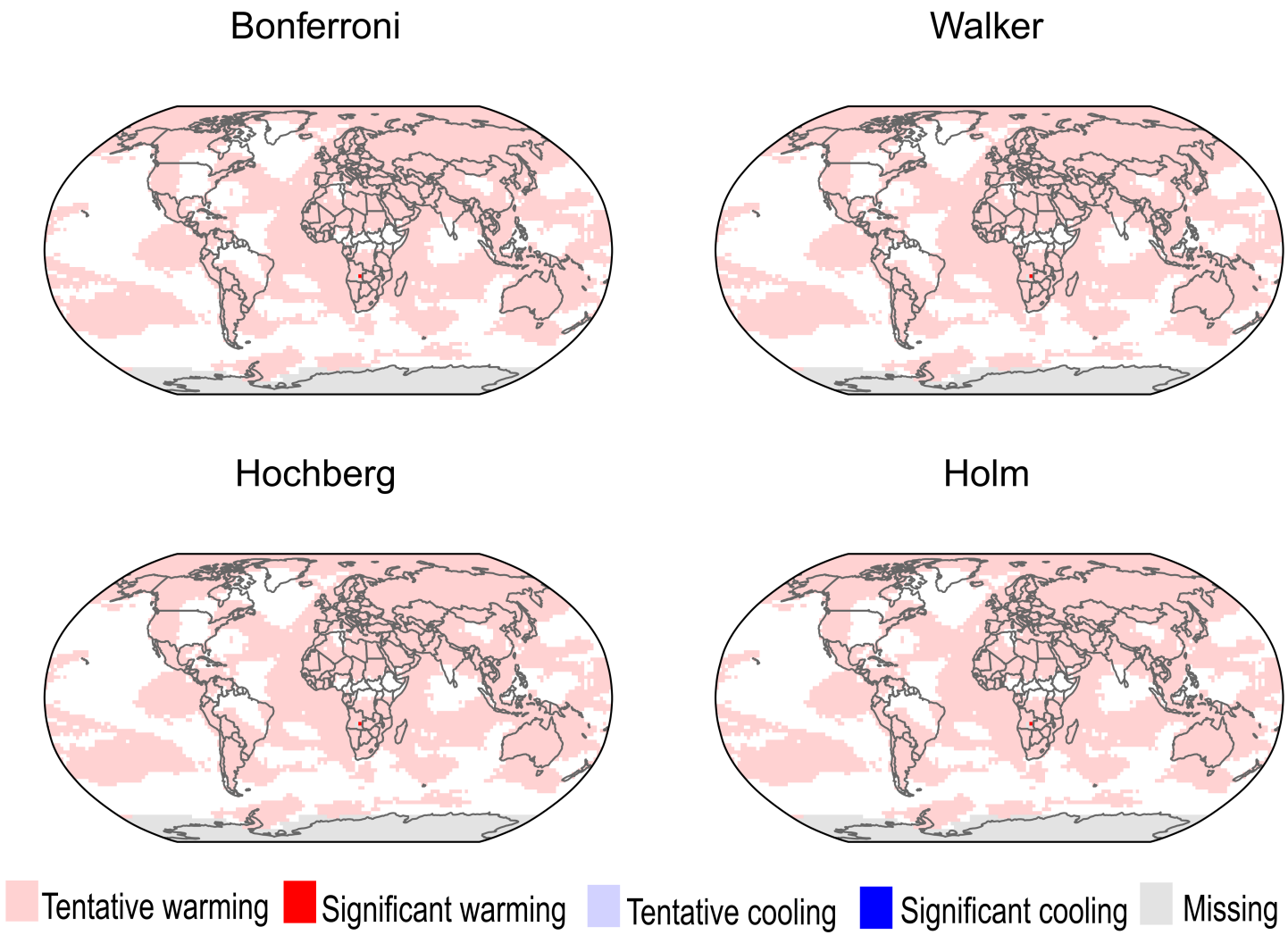
SNR



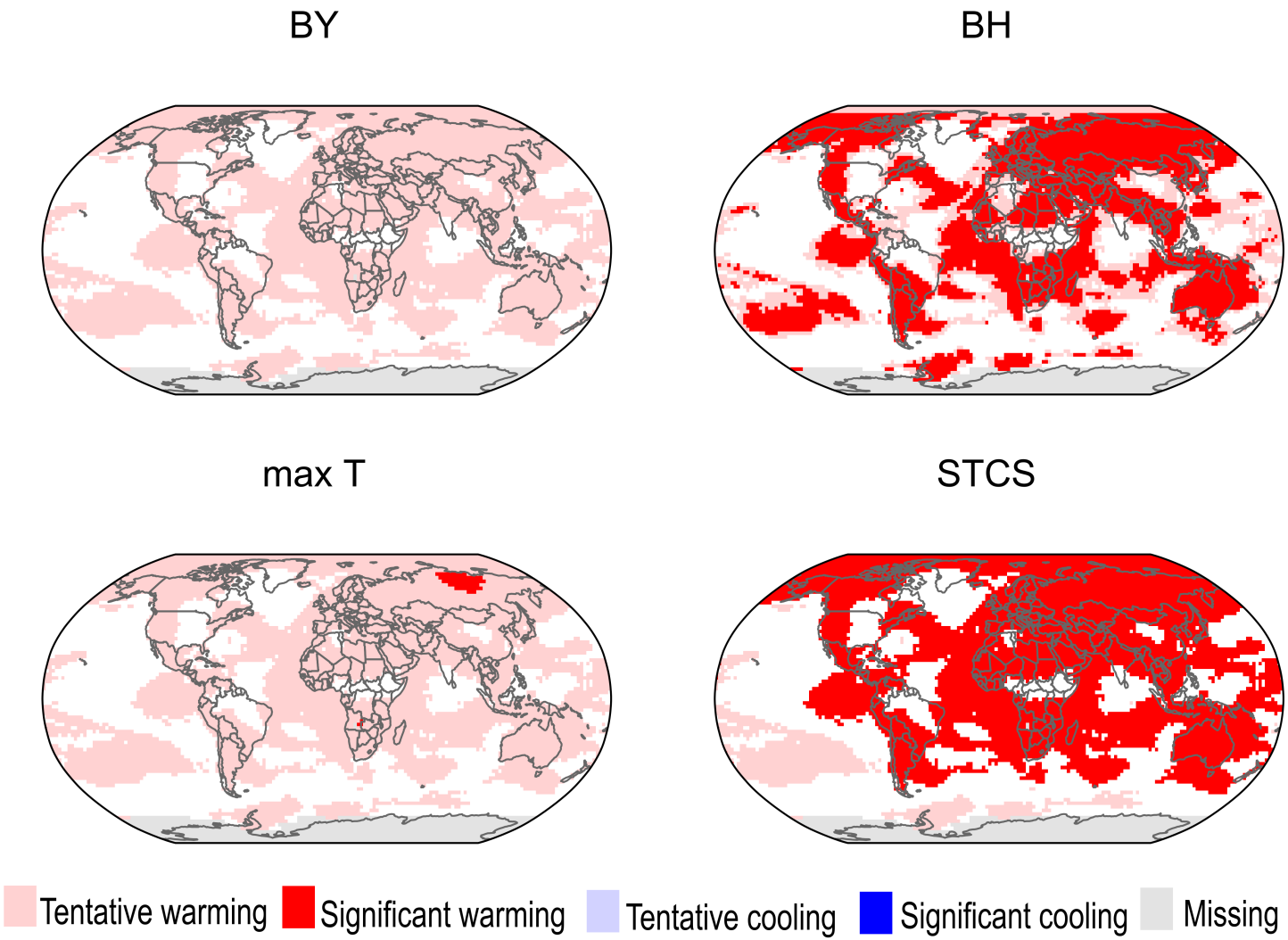
Rho



Results


Bonferroni
Walker
Hochberg
Holm
Tentative warming
Significant warming
Tentative cooling
Significant cooling
Missing

Results


BY
BH
max T
STCS
Tentative warming
Significant warming
Tentative cooling
Significant cooling
Missing

Conclusion

- Multiple testing corrections ensure the analysis is done in a statistically rigorous way $\rightarrow$ Enhanced reliability of results.
- Can help better identify statistically significant spatial patterns.
- Post-hoc methods $\rightarrow$ can be applied to existing results.
- Permutation methods - especially the supra-threshold cluster size - offer increased statistical power compared to Bonferroni related methods.

Conclusion

- Multiple testing corrections ensure the analysis is done in a statistically rigorous way $\rightarrow$ Enhanced reliability of results.
- Can help better identify statistically significant spatial patterns.
- Post-hoc methods $\rightarrow$ can be applied to existing results.
- Permutation methods - especially the supra-threshold cluster size - offer increased statistical power compared to Bonferroni related methods.

Thank you!

contact: `jose.cortes@uni-jena.de`