The Stippled Gridpoints are Statistically Significant: (Mis)uses of False Discovery Rate Correction for Geospatial Data

Michael Konrad Schutte[1,2], Leonardo Olivetti[1,2], Flavio Maria Emanuele Pons[3,4], and Gabriele Messori[1,2,5]

[1]Department of Earth Sciences, Uppsala University, Uppsala, Sweden
[2]Swedish Centre for Impacts of Climate Extremes (climes), Uppsala University, Uppsala, Sweden
[3]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay, IPSL, 91191 Gif-sur-Yvette, France
[4]Science Partners, Paris, France
[5]Department of Meteorology, Stockholm University, Stockholm, Sweden

Correspondence: Michael Konrad Schutte (michael.schutte@geo.uu.se)

Abstract. Peer-reviewed articles in the geosciences routinely assess statistical significance in spatially distributed data. Statistical significance is often assessed independently at each grid point, while formal adjustment for multiple testing is applied less consistently. Although several approaches to account for multiple testing exist, their application to geosciences data is not always straightforward, as these data often exhibit spatially coherent signals.

In this work, we revisit multiple-testing correction in the context of spatially structured datasets. We first highlight how neglecting multiple testing correction can substantially inflate the number of false positives. We further show that the global false discovery rate (FDR) approach, proposed in literature for application in geosciences, can yield counterintuitive and potentially misleading results when applied to spatially coherent signals. To illustrate the latter point, we provide an example based on near-surface air temperature composites following sudden stratospheric warmings. We show that when anomalies are spatially coherent, restricting the spatial domain can increase the FDR-adjusted significance threshold. Consequently, the same underlying field can appear more statistically significant solely due to domain selection, despite unchanged data. We explain this behavior from the rank-based structure of the FDR procedure and discuss its implications for spatial inference and uncertainty quantification in the geosciences.

Building on these insights, we outline practical recommendations for transparent and robust significance assessment in geoscientific applications. These include clearly documenting multiple-testing corrections when adjusted pointwise significance is shown, cautious interpretation of adjusted thresholds, and considering spatially aware alternatives such as regional or cluster-based inference when appropriate. Overall, our results highlight both the need to account for multiple-testing and potential issues with a naïve application and interpretation of the FDR correction. We hope that our work may contribute to more robust statistical testing in the geosciences.

1 Introduction

Highlighting regions with statistically significant signals on maps is very common in the analysis of geospatial data, including model-evaluation, observational diagnostics and event composites (e.g., Deser et al., 2012; Eade et al., 2014; Butler et al., 2017). This approach supports hypothesis-testing and interpretation of results, and can also translate into real-world decision-making, e.g., in the case of forecast evaluation (e.g., Jeuring et al., 2024; Pappenberger et al., 2019; Taggart and Wilke, 2025). However, the common practice of applying a pointwise significance test at each data point (i.e. repeated local tests) introduces a fundamental statistical issue: when hypothesis tests are conducted multiple times, the probability of finding apparently significant results purely by chance increases sharply. For example, if 100 independent grid points are tested at the 5 % significance level, one would expect on average five points to appear "significant" even in the absence of any true signal (i.e. false rejections of the null hypothesis). However, for any single sample, this number may be higher. Moreover, the chance of at least one spurious rejection is extremely high: 99.41% when performing 100 local tests. This point was eloquently made by Wilks (2016). However, the geosciences community has been aware of this issue for well over a century (Walker, 1914) and the issue has since been revisited regularly (e.g., Livezey and Chen, 1983; Katz and Brown, 1991; Katz, 2002; Wilks, 2006, 2016).

To avoid inflating the number of false rejections, adjustment for multiple testing is required. A variety of approaches have been proposed to address this issue, including family-wise error rate control (e.g., Walker, 1914), field significance methods (e.g., Livezey and Chen, 1983), and, more recently, false discovery rate (FDR) control (Benjamini and Hochberg, 1995). FDR aims to limit the expected proportion of falsely rejected hypotheses among all tests, and is widely applied in many fields such as genomics or neuroimaging (e.g., Noble, 2009; Alberton et al., 2020). The original paper by Benjamini and Hochberg (1995) has accumulated more than 125,000 citations according to Google Scholar. Following the review and recommendations of Wilks (2016), FDR has also become the de facto golden standard for multiple-testing correction in geoscientific applications. Nonetheless, it is still implemented sporadically in the field. When randomly sampling fifty articles from meteorology and climate science cited by the IPCC Sixth Assessment Report, Chapter 11 (Seneviratne et al., 2021), discussing how anthropogenic climate change affects weather and climate extreme events, we found that 18 studies assessed statistical significance on a map, of which 17 without multiple testing correction. Only one article implemented a correction, using FDR. These numbers are consistent with previous empirical observations of the same issue (e.g., Wilks, 2016).

However, even the application and interpretation of FDR can be problematic. FDR is often misunderstood as adjusting the significance threshold for each individual grid point, making it "harder" for a point to be considered significant. The last author of this study has himself implicitly adopted this fallacious interpretation in some of his past work. While FDR does control the global error rate, it does not provide any guarantee for individual grid points. In some applications, the resulting FDR threshold may even be higher than the uncorrected threshold. That is, some gridpoints become significant only after the FDR correction. This effect is highly sensitive to the choice of domain. Interpreting FDR-corrected significance on maps as if each highlighted grid point were locally significant can therefore be misleading.

In this work, we aim to raise awareness of these issues and argue that: (i) Any depiction of pointwise statistical significance on maps must be accompanied by an explicit and appropriate multiple-testing correction; and (ii) the FDR procedure can yield

misleading thresholds if misunderstood or deliberately exploited. We illustrate these points with an example based on meteorological reanalysis data. Finally, we conclude with some practical recommendations for transparent and robust significance testing in the geosciences and its interpretation.

2 The false discovery rate

Under the assumption of independent tests, the number of falsely rejected null hypotheses is a random variable following a binomial distribution with probability α, the chosen significance level (Wilks, 2016). Consequently, the probability of obtaining at least one false positive approaches unity as the number of tests increases. For the commonly used 5% significance level, the probability of no false rejections is $(1-0.05)^n$, where n is the number of tests. In high-resolution gridded maps containing thousands of grid points, at least one false rejection is therefore almost certain. Moreover, the expected number of false rejections scales linearly with n (the expected value of the binomial distribution is $n \cdot \alpha$), implying that hundreds of spuriously significant grid points may appear in a single high-resolution map (e.g., on average about 500 for $n = 10{,}000$ grid-points).

A commonly-used approach in the geosciences to address this issue is the FDR correction (Benjamini and Hochberg, 1995). FDR is defined as the expected proportion of incorrectly rejected null hypotheses among all rejected hypotheses. For example, an FDR level of $\alpha_{\mathrm{FDR}} = 0.1$ implies that, on average, 10% of rejected null hypotheses are false positives. In practice, the p-values from the local tests are first sorted in ascending order and a FDR threshold is defined as

$$p^*_{\mathrm{FDR}} = \max_{i=1,\ldots,N_{\mathrm{tests}}} \left[p_i : p_i \leq \left(\frac{i}{N_{\mathrm{tests}}} \right) \alpha_{\mathrm{FDR}} \right], \tag{1}$$

so that all p-values smaller than or equal to the largest p-value satisfying this condition are considered statistically significant. If no local p-value satisfies this condition, the field as a whole does not reach statistical significance, implying that all apparent local discoveries may be spurious–an interpretation analogous to that of field significance tests.

The appeal of FDR for geoscientific applications lies primarily in its ease of implementation and apparent ease of interpretation, as well as its robustness to violations of test independence (Wilks, 2016). Indeed, numerical experiments by Wilks (2016) show that FDR performs well even in the presence of strong spatial autocorrelation – as is commonly the case in many geophysical fields – whereas earlier approaches such as field significance testing are highly sensitive to spatial dependence (Livezey and Chen, 1983).

However, while FDR provides a valid control of a global error rate, it does not necessarily render individual tests more conservative in a local sense. Based on simulations with spatially autocorrelated data, Wilks (2016) recommends selecting an α_{FDR} approximately twice as large as the target field significance level in order to avoid overly restrictive thresholds. This recommendation is well-motivated for large-scale maps in which statistically significant signals are sparse and spatially restricted, yet it exposes an important limitation of the FDR procedure. Because the FDR threshold p^*_{FDR} depends on the rank ordering of local p-values, it can become counterintuitively permissive when applied to a spatial subset exhibiting a prevalent signal. In practice, selecting α_{FDR} approximately twice as large as the target field significance level can lead to corrected thresholds that exceed the nominal local significance level. This may produce the counterintuitive outcome that more grid

points appear statistically significant following FDR correction than when using uncorrected local tests. This is particularly an issue if the region of interest is selected ad hoc following exploratory analysis.

FDR operates by controlling a set-level property of the rejected hypotheses, rather than adjusting the error probability at each grid point. When FDR results are visualized in the same manner as pointwise significance maps, this distinction is easily obscured, encouraging local interpretations that the method does not support. In contrast to pointwise testing, which bounds the false-positive probability at each location, FDR constrains only the expected fraction of false positives among all declared discoveries. As a result, FDR-based significance should not be interpreted as a local measure of statistical significance. This point is illustrated with a real-world case study in the next section.

3 Example of assessing statistical significance with meteorological data

We now illustrate how the FDR procedure by Benjamini and Hochberg (1995) can yield corrected thresholds that exceed the nominal local significance level and how markedly different apparent patterns of statistical significance can result from the same underlying data, solely as a consequence of changing the analysis domain. Employing an example from meteorology, we show composites of 2-m air temperature (t2m) anomalies, averaged over the 60 days following Sudden Stratospheric Warming events (SSWs; e.g., Baldwin et al., 2021; Lee et al., 2025). It is well-documented that SSWs often result in near-surface responses, for instance showing below-average temperatures over northern Eurasia (e.g., Kolstad et al., 2010; Butler et al., 2017). However, considerable uncertainty remains in the exact location and magnitude of these anomalies due to variability in stratosphere-troposphere coupling (e.g., Oehrlein et al., 2021; Kolstad et al., 2022). In the following, we illustrate how domain choice affects the visual interpretation of statistical significance in this example.

3.1 Temperature data and p-value computation

The temperature field is obtained from ERA5 reanalysis (Hersbach et al., 2020), a widely used data set in meteorology. We use t2m data over the Northern Hemisphere North of 20°N with a horizontal resolution of 0.5°. Temperature anomalies are computed by subtracting a smoothed seasonal cycle (15-day centered running mean). SSW dates are obtained from the Sudden Stratospheric Warming Compendium (Butler et al., 2017; Butler, 2025), based on the conventional definition of reversal of 10 hPa zonal winds at 60°N (Charlton and Polvani, 2007).

To test the statistical significance of the t2m anomalies, we assess at each grid point whether the 60-day mean t2m anomaly following the SSW dates is colder than climatology. The climatological distribution of t2m is estimated through a resampling approach with replacement: the mean of randomly selected 60-day periods from extended winter (November–March) is computed 10,000 times. Following this, at each grid point the p-value of the composite anomaly is computed using a one-sided Welch's t-test, assessing whether the t2m anomaly during SSWs is significantly colder than climatology.

3.2 Statistical significance assessment

Assessing statistical significance with independent tests for a 5% threshold at each grid point highlights broad regions of significant cold temperature anomalies (Fig. 1a). Even relatively weak anomalies, for example over the North Atlantic or over North America, are marked as statistically significant. Regardless of the lower temperature variability over oceans than over land, one may suspect that the significance of some signals is spurious.

Applying the FDR correction as recommended by Wilks (2016) with $\alpha = 0.1$ to all grid points between 20°N and 90°N, reduces the number of points marked as statistically significant (Fig. 1b). In agreement with earlier studies, northern Eurasia stands out as the region with the clearest signal of significant below-average temperatures following SSWs (e.g., Kolstad et al., 2010; Butler et al., 2017). In this case, the FDR correction suppresses points with relatively weak signals, as these points could be linked to false rejections, while retaining signals with the highest amplitudes.

We now restrict our attention to t2m anomalies over Northern Europe (55°N - 70°N, 4°E - 32°E), which has been highlighted as a region experiencing colder than usual temperatures following SSWs in a number of studies (e.g., King et al., 2019; Kolstad et al., 2010; Monnin et al., 2022). If the FDR correction is applied only to grid points within northern Europe, the cold anomalies over nearly the entire region appears statistically significant (Fig. 1d). This is in sharp contrast to the case where FDR is applied on the full hemispheric analysis domain (cf. Fig. 1c to d). The number of stippled gridpoints is even larger than the uncorrected significance testing case (cf. Fig. 1d to Fig. A1). The discrepancy occurs even though the underlying temperature anomaly field is identical in all cases. It arises from the dependence of the FDR threshold p^*_{FDR} on the rank ordering of p-values. Restricting the analysis to a region dominated by low p-values increases the adjusted threshold, allowing a larger fraction of points to satisfy the corrected value.

To better understand this behavior, we examine the ranked p-values and corresponding thresholds of negative t2m anomalies (Fig. 2). When the correction is applied to all grid points between 20°N and 90°N, the intersection defining p^* (Eq. 1) is found at $p^* \approx 0.9\%$, well below 5% (Fig. 2a). In contrast, when the correction is applied only to Northern Europe, p^* is an order of magnitude higher (Fig. 2b). This difference arises because the p-values over Northern Europe are generally small, increasing the fraction of points below the $\frac{i}{N}\alpha$ line when fewer grid points are considered. While the absolute distribution of p-values over northern Europe remains unchanged, their relative ranks within the smaller test family differ, resulting in a substantially more permissive threshold.

4 The domain dependence of FDR correction

The example above illustrates that the false discovery rate (FDR) correction as commonly implemented in geoscience applications can yield less restrictive thresholds than the uncorrected statistical test when applied to spatial domains dominated by coherent signals. This behavior follows directly from the way the Benjamini–Hochberg procedure is defined (Benjamini and Hochberg, 1995), which identifies the critical p-value threshold p^* as the largest p-value satisfying Eq. 1.

When only a small fraction of p-values in the full set are low, the line $\frac{i}{N}\alpha$ intersects the sorted p-values at a low relative rank $\frac{i}{N}$, resulting in a relatively strict (low) threshold. Conversely, when the analysis is restricted to a domain dominated by low p-

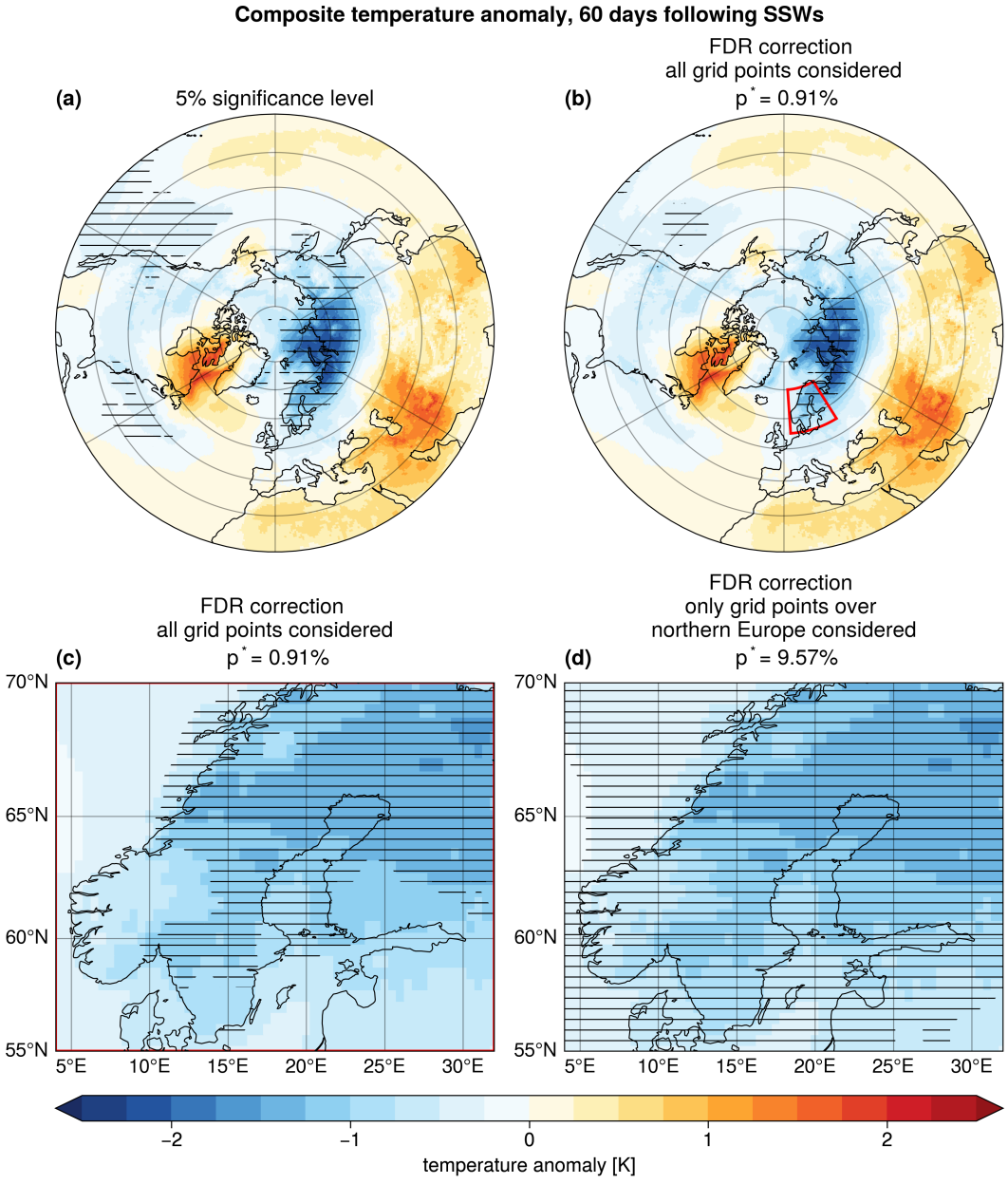

Figure 1. Composites of 2-m air temperature anomalies averaged over 60 days following SSWs. Statistical significance of cold anomalies is indicated by horizontal hatching for all grid points north of 20°N with respect to: (a) A local 5% significance level using a one-sided Welch's t-test and (b) adjusted with the FDR correction. Panel (c) shows the same hatching as in the red box in panel (b), while panel (d) shows horizontal hatching for the FDR corrected significance with respect to only northern European grid points. The subtitles in panels (b) to (d) further indicate the FDR adjusted p-value as p*. A zoomed-in map of northern Europe for the local 5% significance level is found in Fig. A1.

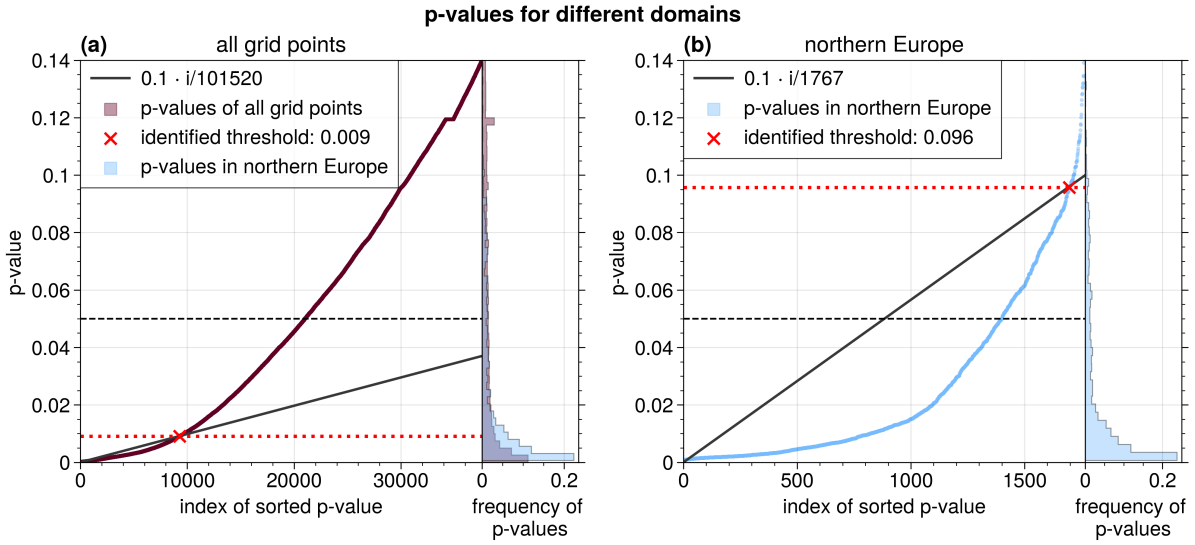

Figure 2. Ranked p-values (colored line) and the FDR threshold (black line) for: (a) All grid points north of 20°N (Fig. 1 b) and (b) Northern Europe (Fig. 1 d). In both panels, the red dotted line indicates the significance threshold obtained from the respective FDR correction, the black dashed line indicates the 5% significance level. Histograms to the right of each panel show the distribution of p-values. Note that both panels only show p-values below 0.14 for improved readability. Fig. A2 shows all p-values, including those above 0.14.

values, the ratio $\frac{i}{N}$ changes such that the intersection occurs at a higher p-value, effectively relaxing the threshold for declaring significance. In other words, the same underlying field can yield an apparently higher fraction of "significant" gridpoints if reducing the spatial domain than both the FDR correction applied to a larger domain and the uncorrected case.

Crucially, this behavior reflects a fundamental tension between a set-level error criterion and the local interpretation encouraged by significance maps. Coherent anomalies, such as large-scale temperature responses, generate clusters of similar p-values that alter the rank structure on which the Benjamini–Hochberg procedure operates. While FDR remains valid under many forms of spatial dependence, its sensitivity to rank ordering implies that spatial context and domain choice can strongly influence the resulting significance threshold, and hence the apparent robustness and spatial extent of detected signals. Yet, this behavior only applies to domains dominated by points with relatively large amplitudes, i.e., the FDR correction reduces the significance threshold p^* as expected for domains containing only few grid points with strong anomalies.

5 Some practical recommendations

The previous sections demonstrate that statistical significance assessment with geospatial data can be misleading, and controlling for the global error rate comes with its own complexities and pitfalls. In particular, FDR correction can invite inappropriate local interpretations when results are visualized on maps, leading to strongly domain-dependent conclusions. Table 1 summa-

rizes a set of practical recommendations aimed at improving robustness, interpretability, and transparency in geoscientific significance assessment. We briefly elaborate on these points below.

A first requirement is transparent reporting of all statistical assumptions. Studies should clearly specify the null hypothesis, test sidedness, test statistic, and significance level. Furthermore, if point-wise testing is appropriate, the point-wise significance level, and the number of tests performed should be reported. Likewise, one should explicitly state whether multiple-testing correction has been applied and, if so, the resulting adjusted threshold. Visualizing the behavior of the FDR procedure, for example through ranked p-value plots as shown in Fig. 2, can substantially improve interpretability and facilitate open discussion of statistical decisions.

Equally important is defining the testing domain a priori. Selecting or modifying spatial domains after inspecting results changes the family of hypotheses being tested and can substantially alter the adjusted significance threshold. As demonstrated in Sect. 3.2, such post-hoc choices may result in misleading interpretations. When widespread anomalies produce unusually permissive FDR thresholds, authors may consider capping the adjusted threshold at a predefined level (e.g. 5%), provided this choice is stated explicitly.

An additional question is whether pointwise testing is appropriate at all. Geophysical signals may exhibit strong spatial coherence, while sample sizes remain limited. Sometimes, the sample size may even be too low to perform statistical testing with sufficient statistical power, which should be stated clearly in that case. Furthermore, grid-point testing may overemphasize spatial detail relative to statistical certainty. In many applications the scientific question concerns the existence of a regional-scale anomaly rather than the significance of individual grid points. When spatial coherence is expected, spatially aware approaches can provide a more consistent alternative to pointwise hypothesis testing. Instead of evaluating each grid point independently, the analysis can be performed on spatially aggregated quantities that better reflect the underlying physical structure of the system. Examples include regional averages, spatial clusters, or object-based approaches, in which coherent features such as anomalies, circulation patterns, or contiguous regions exceeding a predefined threshold are identified and treated as the statistical units of analysis. By reducing the dimensionality of the problem, these approaches decrease the number of independent tests and align statistical inference more closely with physically meaningful structures.

The regional mean temperature analysis presented above provides a simple illustration. Considering only land points, the regional mean 60-day t2m anomaly over northern Europe lies well below its resampled climatological distribution and can be assessed using a single hypothesis test ($p = 3.98 \cdot 10^{-9}$ for a mean anomaly of $-0.94°$C). In this case, inference focuses directly on the question whether a coherent regional response exists, rather than on hundreds of individual grid-point tests whose outcomes are strongly correlated.

Another practical aspect concerns the choice of the FDR level itself. In most applications, the FDR level α is selected a priori (e.g. $\alpha = 0.1$ following Wilks (2016)) and treated as independent of the data structure. However, our example suggests that the effective behavior of the FDR procedure depends not only on α, but also on the proportion of nominally significant tests (e.g., at the 5% level) and the spatial coherence of the field. Thus, one may consider adapting the FDR level based on properties of the tested field, such as the fraction of significant nominal tests or the ratio between domain size and spatial autocorrelation length scale. For example, one could reduce an initial $\alpha_{in} = 0.1$ to $\alpha_{eff} = \alpha_{in} \cdot exp(-N_{sig,0.05}/N)$ which accounts for the fraction

Table 1. Practical recommendations for assessing statistical significance in geospatial data.

Recommendation	Explanation	Examples
Specify statistical assumptions transparently	State null hypothesis, test sidedness, test statistic, and significance level. If pointwise significance testing is appropriate, report the number of tests performed. If multiple test correction is applied, also report all relevant information for correct interpretation and reproducibility (methodology, adjusted threshold p^* and significance level).	(Scientific best practice)
Visualize FDR behavior	If FDR correction is applied, show sorted p-values together with the $\alpha \frac{i}{N}$ decision line to illustrate how the rejection threshold is obtained.	Wilks (2016); Fig. 2
Avoid post-hoc domain changes	Define spatial domain and masks before significance testing; changing domains alters the hypothesis family, inferred threshold and inflates the de facto number of tests being performed (Gelman and Loken, 2019)	Sec. 3.2.
Interpret large adjusted thresholds cautiously	When anomalies with relatively high amplitudes cover large fractions of the domain, FDR-adjusted threshold can become unrealistically permissive; optionally cap thresholds (e.g. at 5%) and report this explicitly.	Schutte et al. (2025)
Check applicability of point-wise tests	Avoid point-wise testing when the sample size is limited or spatial coherence dominates the signal.	Cortés et al. (2020); Kolstad et al. (2022)
Use spatially aware methods when appropriate	When spatial coherence is expected, consider regional averages, cluster-based tests, or object-based approaches reducing either the FDR level α or N.	Lembo et al. (2026); Sun et al. (2015); Faranda and Alberti (2026); Wilks (2006)

of significant grid points at the point-wise 5% level $N_{sig,0.05}$ and all grid points of a given domain N (Fig. A3). We note that the function to adjust α is just an example and may reduce the statistical interpretability of the FDR correction method.

Motivated by the role of spatial dependence highlighted above, one may also consider intermediate approaches that retain grid-point information while accounting for spatial dependence more explicitly. One possible extension is a modified field-significance criterion based on the Walker approach (Wilks, 2006), in which the total number of tests N is replaced by an effective number of independent tests:

$$p^* = 1 - (1 - \alpha_0)^{1/N_{eff}} \tag{2}$$

Here, $N_{\text{eff}} < N$ represents the reduced, effective number of degrees of freedom implied by spatial dependence among grid points. In practice, N_{eff} may be estimated from the spatial autocorrelation structure, for example using an integrated correlation

area ($N_{\text{eff}} = N/N_{ICA}$; Faranda and Alberti 2026), or from other definitions of spatial objects or clusters. Framed in this way, Eq. (2) can be interpreted as a pragmatic compromise between fully pointwise inference and purely regional testing.

However, estimating N_{eff} reliably remains challenging. Spatial correlations are typically non-stationary and anisotropic, and their structure varies across variables, regions, and temporal scales. As noted by Wilks (2016), these difficulties limit the robustness of methods that rely on a single global estimate of independence. While such approaches could be useful if N_{eff} can be estimated reliably, they should not currently be viewed as universal replacements for established multiple-testing procedures. In this sense, both adapting the FDR level and estimating N_{eff} can be viewed as attempts to reconcile multiple-testing control with the effective dimensionality imposed by spatial dependence.

6 Conclusions

The peer-reviewed geosciences literature routinely presents statistical significance on spatial maps. Multiple testing correction is essential for any inference drawn from such analyses. Yet, such correction should be applied with care. In this work, we have highlighted a specific caveat of the widely adopted false discovery rate (FDR) correction proposed by Wilks (2016): when a large fraction of grid points exhibit strong anomalies, the resulting FDR threshold can become relatively permissive, potentially leading to overconfident conclusions if significance is interpreted grid point-wise.

This behavior does not imply that FDR procedures are inherently flawed. Rather, it reflects a mismatch between the aim of the FDR correction method suggested by Wilks (2016) and the interpretation that is typically given of statistical significance in a spatial map. Because the FDR threshold depends on the global rank distribution of p-values, it is sensitive to spatial coherence, domain definition, and the proportion of affected grid points. As a result, statistical conclusions derived from maps can change in ways that appear counterintuitive when the tested spatial domain is modified, even if the underlying physical signal remains unchanged.

We have proposed a set of practical recommendations to improve the robustness and transparency of statistical assessments in geospatial analyses. These include clearly reporting the applied correction method along with all relevant information, cautious interpretation of adjusted thresholds, and considering spatially aware alternatives when appropriate. Such considerations are crucial to ensure reproducibility, comparability and interpretability across studies.

Ultimately, significance assessment in geospatial data involves a trade-off between interpretability and statistical rigor. Pointwise inference provides intuitive spatial detail but risks overstating confidence, whereas spatially aggregated approaches offer stronger error control at the expense of spatial specificity. Developing methods that balance these competing goals while remaining computationally efficient and interpretable remains an important direction for future research.

Code and data availability. ERA5 data is freely available from https://doi.org/10.24381/cds.bd0915c6 (Hersbach et al., 2025). The code to create the figures is available from the authors upon request.

Appendix A: Supplementary figures

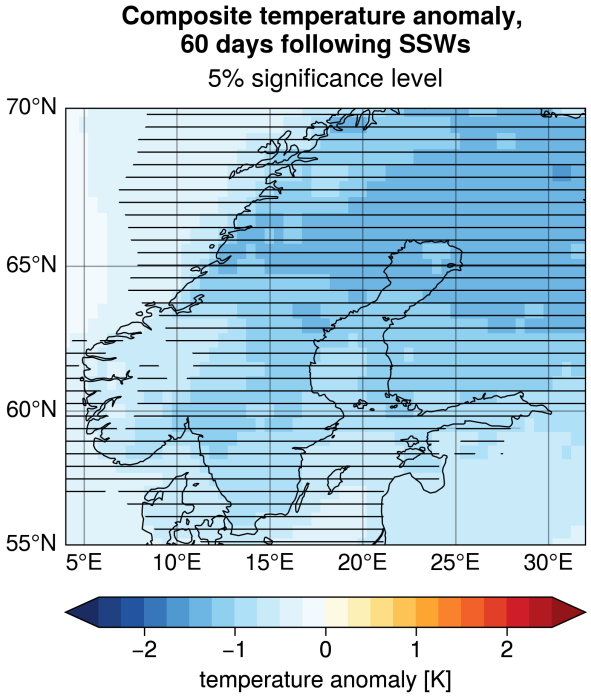

Figure A1. Composites of 2-m air temperature anomalies averaged over 60 days following SSWs. Horizontal hatching indicates statistical significance of cold anomalies with respect to a local 5% significance level.

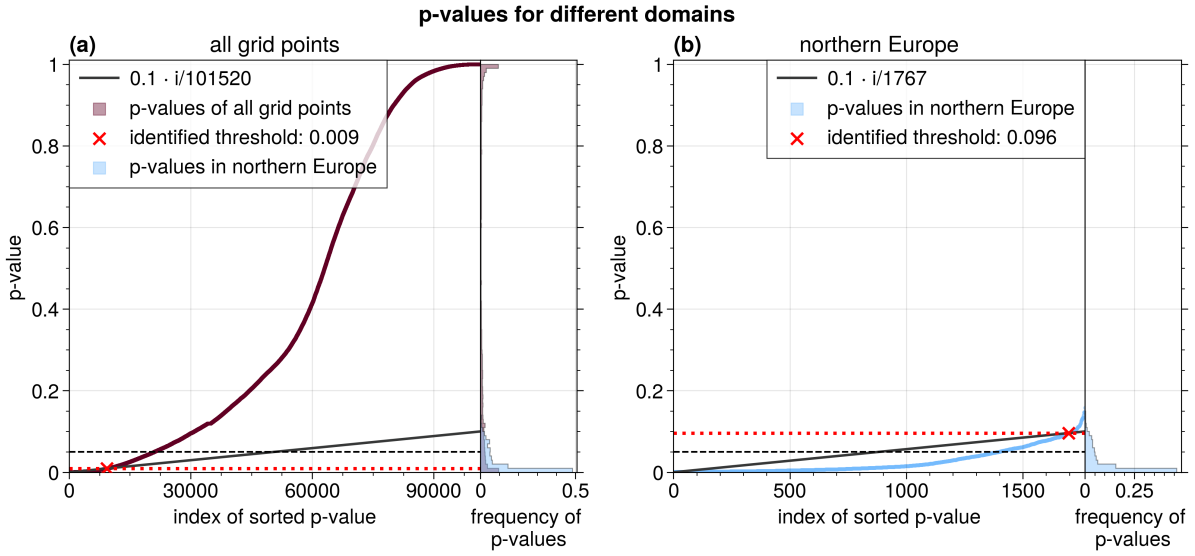

Figure A2. As Fig. 2, but showing all p-values, including those above 0.14.

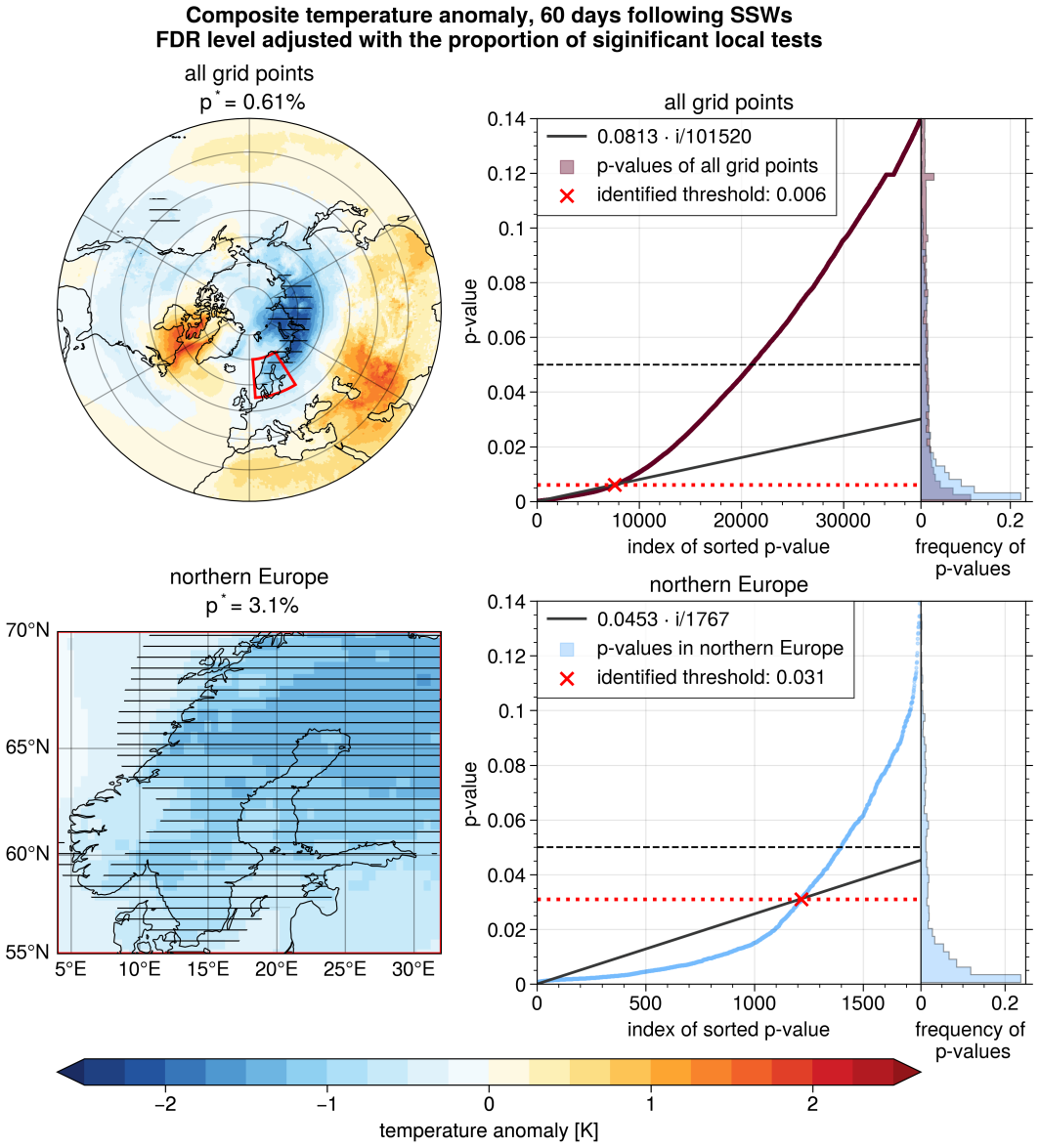

Figure A3. Example of using FDR correction with the FDR level α being adjusted with respect to the fraction of grid points exceeding the local 5% significance level to all grid points as described in Sec. 5. The first row shows results for all grid points north of 20°N, the second row shows results for northern Europe, as defined in Sec. 3.2 and indicated by the red box in panel (a). The left column shows composites of 2-m air temperature anomalies averaged over 60 days following SSWs. Statistical significance of cold anomalies is indicated by horizontal hatching. The right column shows the ranked p-values (colored line) and the FDR threshold (black line). In both panels, the red dotted line indicates the significance threshold obtained from the respective FDR correction, the black dashed line indicates the 5% significance level. Histograms to the right of each panel show the distribution of p-values. Note that both panels only show p-values below 0.14 for improved readability.

Author contributions. **Schutte, M. K.:** Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Olivetti, L.:** Conceptualization, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Pons, F. M. E.:** Conceptualization, Writing - Review & Editing, Visualization. **Messori, G.:** Conceptualization, Methodology, Writing - Review & Editing, Visualization, Supervision, Funding acquisition.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We would like to thank Daniel Stephen Wilks and Davide Faranda for their helpful feedback on the manuscript. MS, LO, and GM acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 948309, CENÆ project). During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. The data handling was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council Vetenskapsrådet through grant agreement no. 2022-06725.

References

Alberton, B. A., Nichols, T. E., Gamba, H. R., and Winkler, A. M.: Multiple testing correction over contrasts for brain imaging, NeuroImage, 216, 116 760, https://doi.org/10.1016/j.neuroimage.2020.116760, 2020.

Baldwin, M. P., Ayarzagüena, B., Birner, T., Butchart, N., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I. V., Garfinkel, C. I., Garny, H., Gerber, E. P., Hegglin, M. I., Langematz, U., and Pedatella, N. M.: Sudden Stratospheric Warmings, Rev. Geophys., 59, e2020RG000 708, https://doi.org/10.1029/2020RG000708, 2021.

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B, 57, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x, 1995.

Butler, A.: Table of major mid-winter SSWs in reanalyses products, https://csl.noaa.gov/groups/csl8/sswcompendium/majorevents.html, 2025.

Butler, A. H., Sjoberg, J. P., Seidel, D. J., and Rosenlof, K. H.: A sudden stratospheric warming compendium, Earth Syst. Sci. Data, 9, 63–76, https://doi.org/10.5194/essd-9-63-2017, 2017.

Charlton, A. J. and Polvani, L. M.: A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks, J. Climate, 20, 449–469, https://doi.org/10.1175/JCLI3996.1, 2007.

Cortés, J., Mahecha, M., Reichstein, M., and Brenning, A.: Accounting for multiple testing in the analysis of spatio-temporal environmental data, Environ. Ecol. Statistics, 27, 293–318, https://doi.org/10.1007/s10651-020-00446-4, 2020.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Clim. Dynam., 38, 527–546, https://doi.org/10.1007/s00382-010-0977-x, 2012.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., and Robinson, N.: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?, Geophys. Res. Lett., 41, 5620–5628, https://doi.org/10.1002/2014GL061146, 2014.

Faranda, D. and Alberti, T.: Investigating the Role of Climate Change in the 3 May 2025 Western Europe Hailstorm Using Atmospheric Analogs, Atmos. Sci. Lett., 27, e70 016, https://doi.org/doi.org/10.1002/asl2.70016, 2026.

Gelman, A. and Loken, E.: The garden of forking paths : Why multiple comparisons can be a problem , even when there is no " fishing expedition " or " p-hacking " and the research hypothesis was posited ahead of time $_*$, https://api.semanticscholar.org/CorpusID:198164638, 2019.

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.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on pressure levels from 1940 to present, https://doi.org/10.24381/cds.bd0915c6, 2025.

Jeuring, J., Samuelsen, E. M., Lamers, M., Müller, M., Åge Hjøllo, B., Bertino, L., and Hagen, B.: Map-Based Ensemble Forecasts for Maritime Operations: An Interactive Usability Assessment with Decision Scenarios, Weather Clim. Soc., 16, 235 – 256, https://doi.org/10.1175/WCAS-D-23-0076.1, 2024.

Katz, R. W.: Sir Gilbert Walker and a Connection between El Niño and Statistics, Statistical Science, 17, https://doi.org/10.1214/ss/1023799000, 2002.

Katz, R. W. and Brown, B. G.: The problem of multiplicity in research on teleconnections, International Journal of Climatology, 11, 505–513, https://doi.org/10.1002/joc.3370110504, _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3370110504, 1991.

King, A. D., Butler, A. H., Jucker, M., Earl, N. O., and Rudeva, I.: Observed relationships between sudden stratospheric warmings and European climate extremes, J. Geophys. Res. Atmospheres, 124, 13 943–13 961, https://doi.org/10.1029/2019JD030480, 2019.

Kolstad, E. W., Breiteig, T., and Scaife, A. A.: The association between stratospheric weak polar vortex events and cold air outbreaks in the Northern Hemisphere, Q. J. Roy. Meteorol. Soc., 136, 886–893, https://doi.org/https://doi.org/10.1002/qj.620, 2010.

Kolstad, E. W., Lee, S. H., Butler, A. H., Domeisen, D. I., and Wulff, C. O.: Diverse surface signatures of stratospheric polar vortex anomalies, J. Geophys. Res.-Atmos., 127, e2022JD037 422, https://doi.org/10.1029/2022JD037422, 2022.

Lee, S. H., Butler, A. H., and Manney, G. L.: Two major sudden stratospheric warmings during winter 2023/2024, Weather, 80, 45–53, https://doi.org/10.1002/wea.7656, 2025.

Lembo, V., Messori, G., Faranda, D., Galfi, V. M., Graversen, R. G., and Pons, F. E.: Concurrent heat waves and their linkage to large-scale meridional heat transports through planetary-scale waves, Weather Clim. Dynam., 7, 453–473, https://doi.org/10.5194/wcd-7-453-2026, 2026.

Livezey, R. E. and Chen, W. Y.: Statistical Field Significance and its Determination by Monte Carlo Techniques, Monthly Weather Review, 111, 46–59, https://doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2, publisher: American Meteorological Society Section: Monthly Weather Review, 1983.

Monnin, E., Kretschmer, M., and Polichtchouk, I.: The role of the timing of sudden stratospheric warmings for precipitation and temperature anomalies in Europe, Int. J. Climatol., 42, 3448–3462, https://doi.org/10.1002/joc.7426, 2022.

Noble, W. S.: How does multiple testing correction work?, Nat. Biotechnol., 27, 1135–1137, https://doi.org/10.1038/nbt1209-1135, 2009.

Oehrlein, J., Polvani, L. M., Sun, L., and Deser, C.: How well do we know the surface impact of sudden stratospheric warmings?, Geophys. Res. Lett., 48, e2021GL095 493, https://doi.org/10.1029/2021GL095493, 2021.

Pappenberger, F., Cloke, H. L., and Baugh, C. A.: Cartograms for use in forecasting weather-driven natural hazards, The Cartographic Journal, 56, 134–145, https://doi.org/10.1080/00087041.2018.1534358, 2019.

Schutte, M. K., Olivetti, L., Krouma, M., and Messori, G.: Stratospheric and tropospheric contributions to North American cold temperatures and subsequent North Atlantic jet stream anomalies, SSRN preprint, https://doi.org/10.2139/ssrn.5614336, available at SSRN: https://ssrn.com/abstract=5614336 or http://dx.doi.org/10.2139/ssrn.5614336, 2025.

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., A. Di Luca, S. G., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., and Zhou, B.: Weather and Climate Extreme Events in a Changing Climate, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)], pp. 1513–1766, Cambridge University Press, https://doi.org/10.1017/9781009157896.013, 2021.

Sun, W., Reich, B. J., Tony Cai, T., Guindani, M., and Schwartzman, A.: False discovery control in large-scale spatial multiple testing, J. R. Stat. Soc. B: Statistical Methodology, 77, 59–83, https://doi.org/10.1111/rssb.12064, 2015.

Taggart, R. J. and Wilke, D. J.: Warnings based on risk matrices: a coherent framework with consistent evaluation, Nat. Hazards Earth Syst. Sci., 25, 2657–2677, https://doi.org/10.5194/nhess-25-2657-2025, 2025.

Walker, S. G. T.: Correlation in Seasonal Variations of Weather, III: On the Criterion for the Reality of Relationships Or Periodicities, Meteorological Office, google-Books-ID: cuGtuAAACAAJ, 1914.

Wilks, D. S.: On "Field Significance" and the False Discovery Rate, Journal of Applied Meteorology and Climatology, 45, 1181–1189, https://doi.org/10.1175/JAM2404.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2006.

Wilks, D. S.: "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, B. Am. Meteorol. Soc., 97, 2263 – 2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.