

Reviewer #3

The manuscript discusses the use of false discovery rate (FDR) multiple testing control for assessing statistical significance in spatially distributed data in geosciences. The authors highlight important issues to consider, such as considering multiple testing, avoiding post-hoc domain changes, and reporting the assumptions of statistical tests done. The main issue they discuss is that the use of the FDR control for multiple testing can be confusing for spatially coherent/dependent data. They suggest cautious interpretation of adjusted thresholds. The topic is highly relevant for analyses in geosciences, but some issues need clarification and enhanced description. Below I give some specific comments as well as a few technical details.

We thank the reviewer for the helpful comments. Below, we respond to each comment individually.

The FDR is defined as the proportion of false discoveries among the discoveries (rejections of the null hypothesis). To best of my understanding, the authors do **not** claim that this control would not hold in some cases, but this is not very clearly stated. Instead the authors want to highlight that if one applies/restricts the analysis to a region where many true discoveries are present, the number of false discoveries can be surprisingly high - even though the proportion of false discoveries remains controlled. It would be good to describe the FDR control idea more precisely, to emphasize that the FDR control works as supposed, and then discuss that there are situations where one might like to introduce a stricter control instead of FDR. Hereby, it would be good to also define/explain more precisely what is meant by "relatively permissive".

We agree, the aim of this manuscript is not to show that FDR control would not hold in some cases, but that the interpretation of FDR-adjusted pointwise statistical significance may be misleading. We describe the idea of the FDR and BH correction in Sect. 2, where we highlight in line with the reviewer's comment that the BH correction permits identifying more false rejections when the analysis contains many rejections (relative to all points considered), and fewer false rejections when only a small fraction of tests are rejected. This becomes particularly problematic when combined with post-hoc selection, as in the example provided in relation to the choice a posteriori of the Scandinavian domain. We now further state explicitly that the latter is only possible if two different nominal levels are chosen. Under identical nominal significance levels ($\alpha_{\text{FDR}} = \alpha_{\text{pointwise}}$), the FDR threshold p^* can never exceed $\alpha_{\text{pointwise}}$.

The performance of the methods is illustrated with a data example. This example shows that the test significances are dependent on the domain choice. It is, however, not possible to show how the number of false/true positives is changing, as the truth is not known. To illustrate differences in false/true positives along domain change and highlight

the points that the authors wish to do, simulated experiments where the true null and alternative hypotheses are known could be prepared.

In line with the first reviewer's comments, we have conducted a numerical experiment with a known number of true positives and updated the manuscript by an additional section. These results of the numerical experiment also cover this comment and are discussed below, as well as in the new Sect. 5 in the revised manuscript.

Fig. R1 shows the sensitivity of p^* , FDP (the realized false discovery proportion), and power to the signal fraction f_s across different values of α . Consistent with the theoretical expectation, p^* increases with f_s and is higher for larger α (Fig. R1a). FDP remains approximately constant for f_s between 0.1 and 0.25, and decreases for larger signal fractions (Fig. R1b). Power is near 1 across all parameter combinations, but in the low-SNR, low sample regime (SNR=0.5, $M=20$) it rises from near zero toward 0.5-0.8 with increasing f_s and is systematically lower for larger α (Fig. R1c). The FDP–Power trade-off in panel (d) is consistent with these findings. A discussion of the synthetic data simulations has been added to the manuscript.

Regarding the relationship between domain size and spatial autocorrelation length scale, we have added a supplementary figure showing the sensitivity of p^* , FDP, and power to other parameters, such as the number of grid points N_g and wavenumber v , which controls the spatial autocorrelation length scale λ (Fig. R2). We find that p^* tends to decrease with larger N_g and, to a lesser extent, with increasing v beyond $v=10$, i.e., for noise with relatively small horizontal scales. FDP remains relatively stable across N_g but increases at higher wavenumbers, while power remains near 1 throughout. The interquartile ranges across parameter combinations are, however, substantial, suggesting that neither trend is robust in isolation and that the behaviour may depend strongly on the signal fraction f_s and its interaction with other parameters.

Results of synthetic data experiment

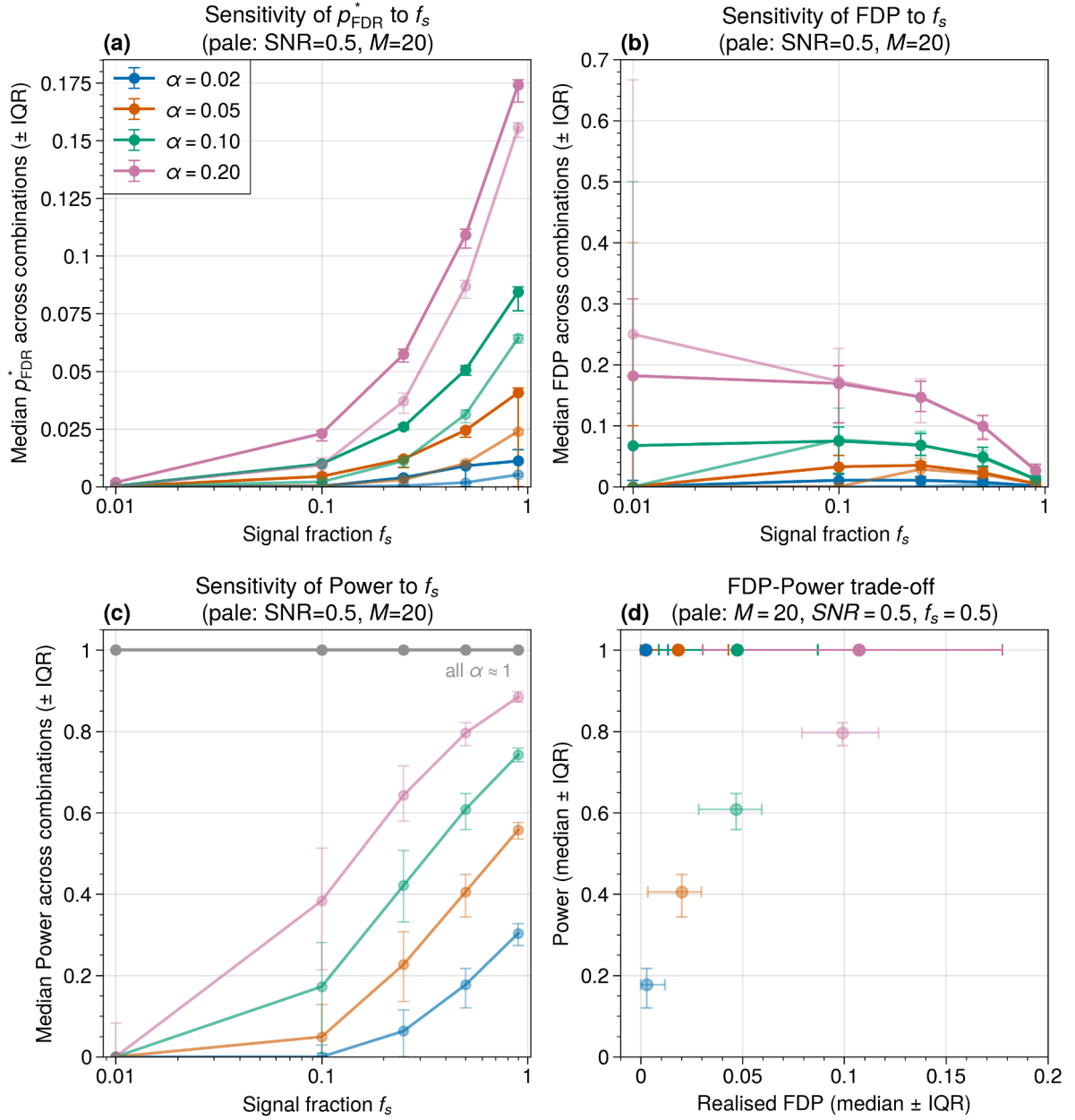


Figure R1: Results of the numerical experiment as a function of signal fraction f_s . Panels (a) - (c) show the median (markers) and interquartile range (error bars) of the FDR-adjusted threshold p^* , the realized false discovery proportion (FDP), and power, respectively, aggregated across all 480 parameter combinations for $M=1000$ (solid) and across the low-power regime (SNR=0.5, $M=20$; pale). Panel (d) shows the FDP-Power trade-off as a scatter plot, with markers indicating the median and error bars the interquartile range across all combinations (solid) and the low-power regime (pale). Colors indicate the choice of FDR level α as given in the legend in panel (a). In panel (c), the median power across all parameter combinations is shown as a single gray line, as all choices of α yield near-identical power close to unity.

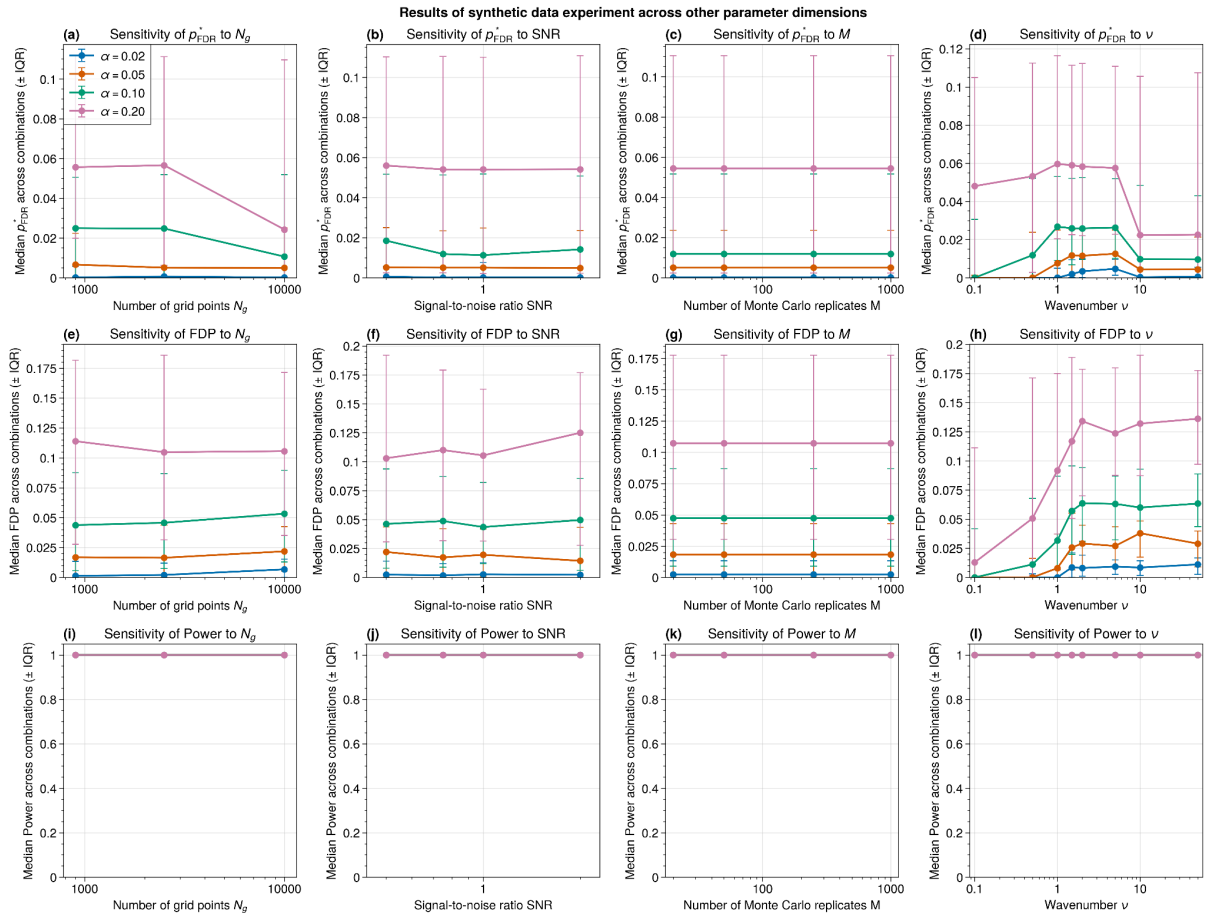


Figure R2: Sensitivity of the FDR-adjusted threshold p^* (top row), realized false discovery proportion FDP (middle row), and power (bottom row) to domain size N_g (first column), signal-to-noise ratio SNR (second column), number of Monte Carlo replicates M (third column), and noise wavenumber ν (fourth column). In each panel, markers show the median and error bars the interquartile range across all remaining parameter combinations. Colors indicate the choice of FDR level α as given in the legend in panel (a). In the bottom row, the median power across all parameter combinations is shown as a single purple line, as all choices of α yield near-identical power close to unity.

The nominal local significance levels exceedances that are discussed apparently occur only if α_{FDR} is set two times as large as the target and local field significance level. It would be good to state this clearly whenever discussed.

In line with the second reviewer's comments (major point 1, and some specific comments), we have specified the context in several places throughout the manuscript. These adjustments also cover this comment. We also refer the reviewer to our reply to their first comment.

Regarding the discussion on the alternatives to the FDR control. 1) If FDR is considered "too permissive", an alternative would also be to use the stricter family-wise error rate (FWER) control. 2) It is generally not a good strategy to adjust the significance levels after seeing the data and its preliminary significances, so I would be careful on suggesting such a strategy.

We thank the reviewer for highlighting these two aspects. In line with our answers to the second reviewer on the same aspect, we would like to highlight that FWER-controlling methods derived under an independence assumption tend to become overly conservative under positive spatial correlation, achieving a true error rate below the nominal level, and an under-detection of signals (Livezey and Chen, 1983). This makes employing FWER-controlling methods less desirable when analyzing geospatial data where positive spatial correlation is common among neighbouring grid points. While we suggest a modified expression of the Walker criterion (Walker 1914) that takes into account spatial correlation, the primary aim of our manuscript is not to derive a new method for FDR control. Nevertheless, some type of FWER control may be a potential alternative, so we now mention FWER in Sect. 6 when discussing practical recommendations.

Regarding the second point, we have removed the suggestion of a specific α_{eff} , and refer instead to adaptive FDR correction methods in the literature as possible refinements. Furthermore, we clarified that capping the FDR-adjusted level at 5% should be decided upon a priori.

The data example uses a resampling approach to test the statistical significance of the t2m anomalies. Here, the 60-day mean t2m anomaly following the Sudden Stratospheric Warming (SSW) dates is computed. Also, 10000 randomly selected 60-day periods from extended winter (November–March) are selected and the mean t2m computed from these. The task is then to compare the t2m anomaly after SSW to the random t2m means. As resampling was done, I would expect here a resampling test being computed for the mean t2m anomaly after SSW to be colder than the random mean t2m. However, the authors used a Welch's t-test. The basic form of this test is for comparing two samples based on the sample means and their standard errors (not available here). Consequently, it should be clarified what test exactly is used.

We thank the reviewer for highlighting this contradiction. We mis-specified the test as Welch's t-test in the analysis and in the text of our manuscript. The original idea was to bootstrap the distribution of the test statistics under the null hypothesis, and use it to compute the p-values. However, while reviewing this part of the manuscript, we realized that there were some mistakes in the procedure. We replaced the procedure with a standard one-sample t-test, since we judge that our sample reasonably meets the assumptions; this slightly changes the shown results (but not the conclusions) compared to the previous version of the manuscript. The specification of the statistical test has been corrected in the updated manuscript and figures have been updated accordingly.

Some details:

page 3, lines 78-79: This sentence needs rephrasing on what exactly is meant. To what is the FDR method compared to here?

We mean to say here (in line with our answer to your first comment) that we do not aim to show that FDR control fails in any case, but that the local interpretation of FDR-corrected thresholds can be problematic if the users do not fully understand the properties of the FDR threshold adjustment procedure, or if post-hoc domain selection is performed. We elaborate on this aspect in the following sentences of that and the following paragraph.

page 3, lines 84-85: Please explicitly state the different significance levels that you either set or wish to have.

We thank the reviewer for this suggestion and have provided examples for numerical values of the local significance level and α_{FDR} .

page 4, line 113: What does 'composite' mean in this connection?

We mean in this context the SSW-composite. The explanation of statistical testing in this paragraph has been adjusted.

page 5, line 127-128: 'anomalies...appear'

We thank the reviewer for highlighting the grammatical error which we corrected.

References

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](https://doi.org/10.1175/1520-0493(1983)111<0046:SFSaid>2.0.CO;2), publisher: American Meteorological Society Section: Monthly Weather Review, 1983.

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.