## Reviewer #1

This paper points out potential problems associated with use of the FDR procedure in the context of spatially correlated hypothesis tests. The authors show that the rule of thumb recommended in the 2016 Wilks paper, which was derived from a particular synthetic data setting with relatively few locally significant gridpoints, apparently behaves badly in the extreme restricted-domain example highlighted in the present paper, and so is evidently not optimal in general.

The paper is fairly short, and would be stronger if it were to include a spectrum of synthetic-data simulations aimed at quantifying optimized parameterization of the ratio of the FDR to the local test level, perhaps as a function of the proportion of local tests that are nominally significant (for example as suggested on line 196), or possibly in terms of the relationship between the domain size and the spatial autocorrelation length scale. Figure 2 seems to indicate that setting the FDR level closer to 0.05 might yield more consistent results in Figure 1d. At minimum, it would be interesting to see the counterpart of Figure 1d with equality of the FDR level and the local test level (i.e., 0.05).

We thank the reviewer for the constructive and very helpful comments. We followed the advice and ran synthetic-data simulations to test whether the observed overconfidence of FDR correction observed in the empirical example also arises in a controlled setting, and to assess how the choice of the FDR level $\boldsymbol{\alpha}$ affects the significance threshold p*. Details of the numerical experiment are described in the revised manuscript.

The main results are summarised in Fig. 1, which shows the sensitivity of p*, FDP (the realized false discovery proportion), and power to the signal fraction $f_s$ across different values of $\boldsymbol{\alpha}$, including the adaptive level $\boldsymbol{\alpha}_{eff} = 0.1 \cdot \exp(-N_{sig,0.05}/N)$ that adjusts the FDR level based on the proportion of locally significant tests. Consistent with the theoretical expectation, p* increases with $f_s$ and is higher for larger $\boldsymbol{\alpha}$ (Fig. 1a). FDP remains approximately constant for $f_s$ between 0.1 and 0.25, and decreases for larger signal fractions (Fig. 1b). 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 lager $\boldsymbol{\alpha}$ (Fig. 1c). The FDP–Power trade-off in panel (d) is consistent with these findings. Across most parameter combinations, $\boldsymbol{\alpha}_{eff}$ falls between $\boldsymbol{\alpha}$ = 0.05 and $\boldsymbol{\alpha}$ = 0.1, suggesting it provides a reasonable adaptive middle ground between FDP control and power retention. These results thus partially address the reviewer's suggestion of quantifying the optimal parameterization of the FDR level as a function of the proportion of nominally significant local tests. A fully optimized mapping would, however, require a more systematic investigation with some optimization strategy across a broader parameter space, and we consider this an important direction for future work, beyond the scope of the present paper. 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. 2). 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.

Performing FDR correction on the t2m example with a local test level of $\boldsymbol{\alpha}$ = 0.05 instead of 0.1 reduces the identified threshold to 3.85%, but yields no significance when considering the entire northern hemisphere (Fig. 3). For the northern hemisphere, this hints at the correction method being so strict with $\boldsymbol{\alpha}$ = 0.05, that no p-values meet the Benjamini-Hochberg criterion, as seen by the linear threshold not intersecting with the observed p-values (Fig. 4).

This result points towards the issue we identified with the FDR correction method also holding in the other direction: If one is zooming out enough, even a formerly robust signal may become non-significant due to the FDR correction becoming relatively strict.

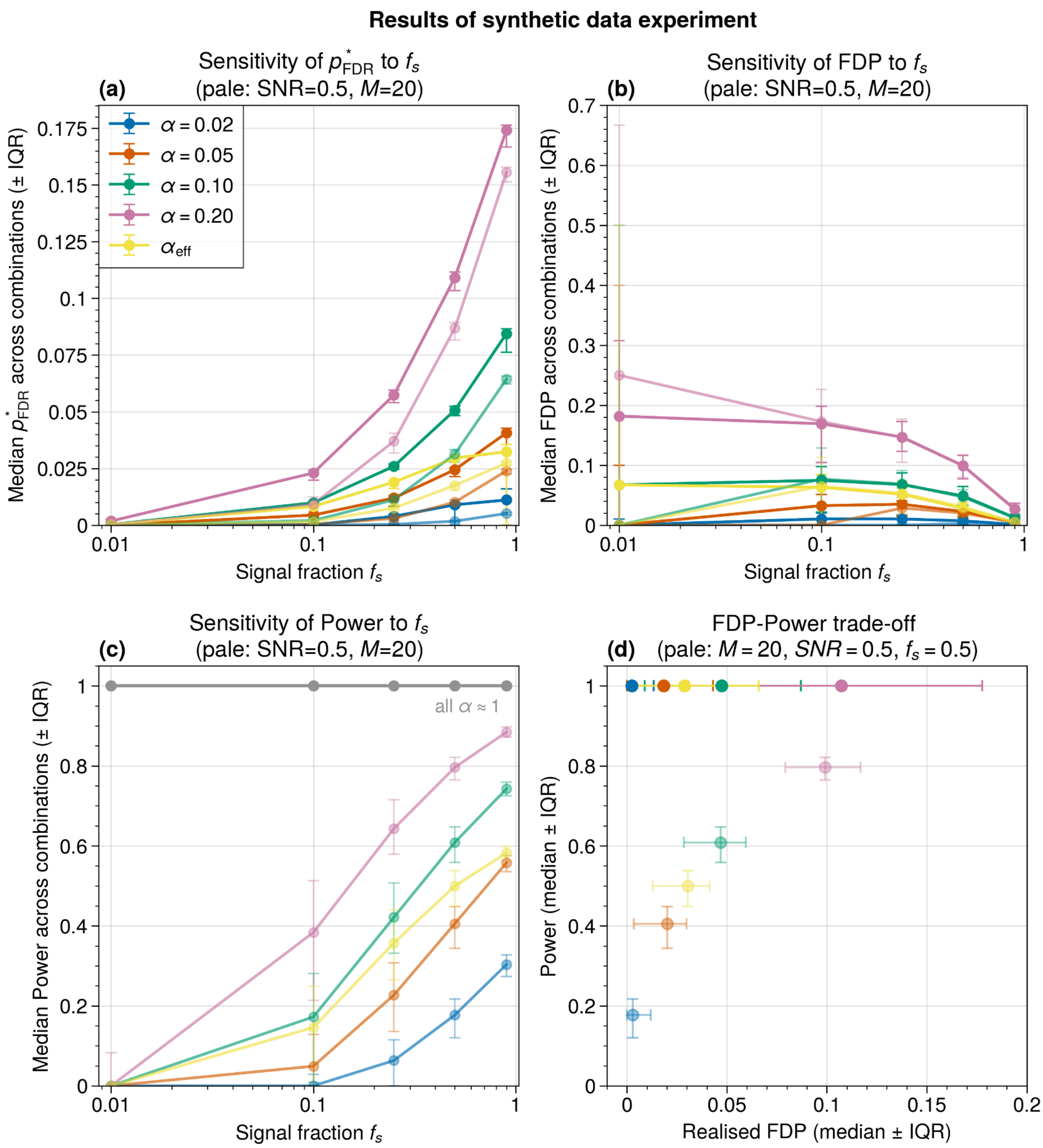


Figure 1: Results of the numerical experiment as a function of signal fraction $f_s$. Panels (a) - (c) show the median (markers) and interquartile range (error bars) of the FDR-adjusted threshold p*, the realized false discovery proportion (FDP), and power, respectively, aggregated across all 480 parameter combinations for M =1000 Monte Carlo replicates (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 $\boldsymbol{\alpha}$ 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 $\boldsymbol{\alpha}$ yield near-identical power close to unity.

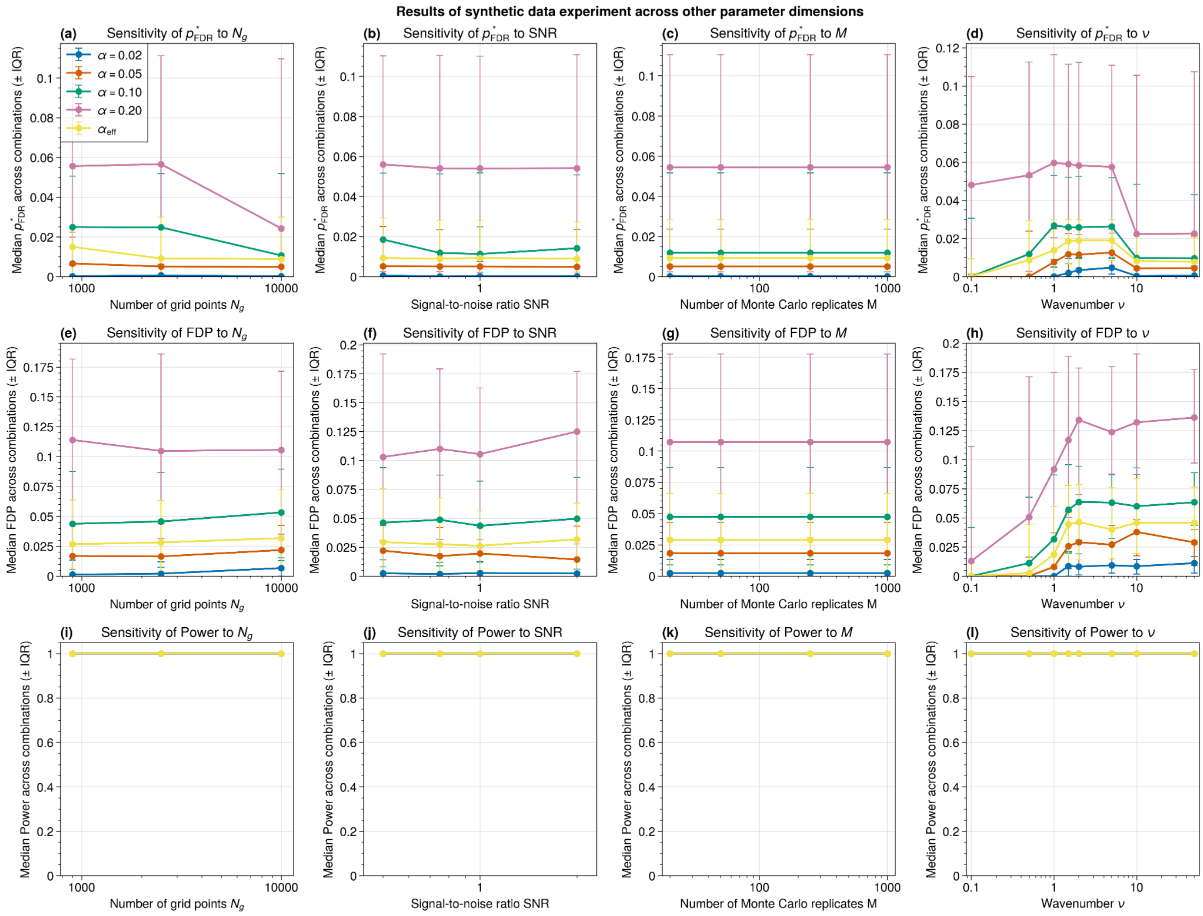


Figure 2: Sensitivity of the FDR-adjusted threshold p* (top row), realized false discovery proportion FDP (middle row), and power (bottom row) to domain size $N_g$ (first column), signal-to-noise ratio SNR (second column), number of Monte Carlo replicates M (third column), and noise wavenumber v (fourth 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 $\boldsymbol{\alpha}$ as given in the legend in panel (a). In the bottom row, the median power across all parameter combinations is shown as a single yellow line, as all choices of $\boldsymbol{\alpha}$ yield near-identical power close to unity.

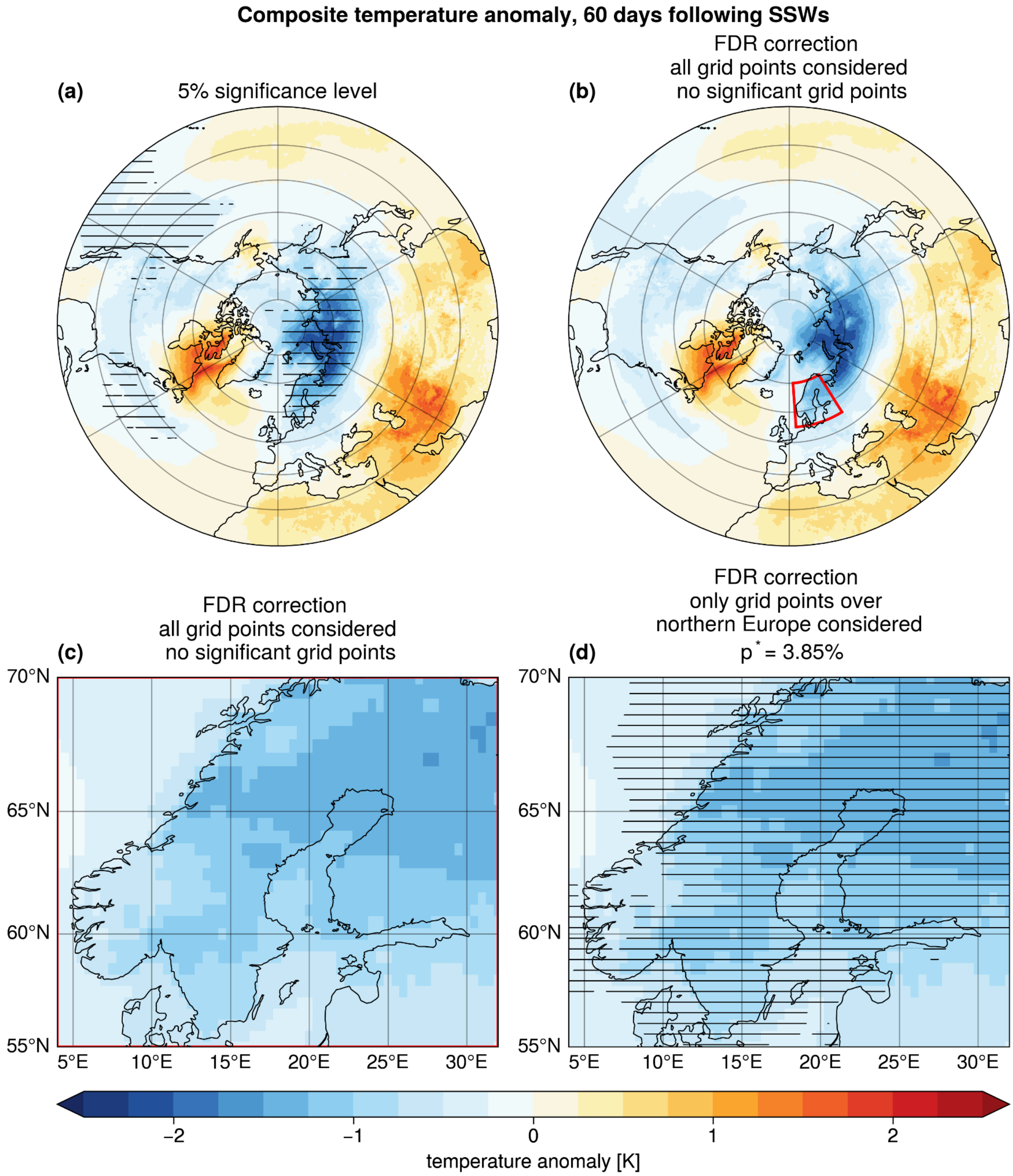


Figure 3: As Fig. 1 in the manuscript, but for ***α*** = 0.05.

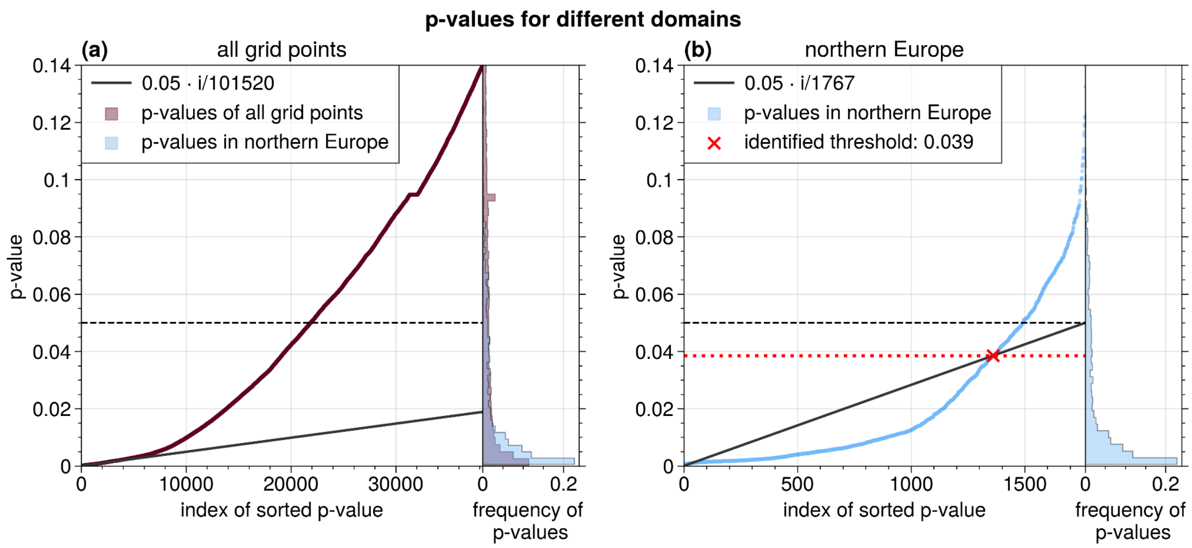


Figure 4: As Fig. 2 in the manuscript, but for $\alpha$ = 0.05.

A few more specific comments:

1. para beginning line 65. Not quite accurate: FDR is actual, not expected number of incorrectly rejected nulls. The Benjamini-Hochberg procedure limits (“controls”) the proportion of such rejections, in expectation. So alpha-FDR = 0.1 implies that, on average, no more than 10% of rejected nulls are false positives, and indeed there may be fewer than this. (The characterization on line 91 is correct).

   We agree and fixed the wording of the sentence in the manuscript.

2. line 110. The test setup here appears to assume implicitly that SSW events are uniformly distributed in the November-March data window, which is substantially longer than 60 days. Is this the case in the observations? Also, is there a physical justification for use of the 60-day period, external to the test data? What is the effect of temperature nonstationarity during November-March?

   While SSWs themselves are not uniformly distributed within the extended winter season, occurring preferentially in mid- to late winter (Birner and Albers, 2017), the t2m data used here have been de-seasonalised, removing the climatological annual cycle; we therefore treat the anomaly statistics as approximately stationary across the November–March period.

   The resampling approach assumes that the de-seasonalised t2m anomaly statistics are stationary across the Nov–March window, i.e., that a randomly drawn 60-day anomaly window looks statistically the same regardless of which winter month it's drawn from. Whether SSWs themselves cluster in January or are spread evenly across the season does not affect that assumption, since we are testing with respect to the background distribution of temperature anomalies that any 60-day window would show.

We nonetheless acknowledge that temperature variance is not fully stationary across November–March (see Fig. 5 for examples of some spatial domains), , which introduces a minor anti-conservative bias in the pointwise p-values. This does not, however, affect the domain-dependence of the FDR-adjusted threshold, which is a consequence of the FDR correction method. We have updated the manuscript to mention this limitation.

The 60-day time window is motivated by evidence that SSWs influence tropospheric circulation for up to 60 days following their onset (Baldwin and Dunkerton, 2001; Oehrlein et al., 2021), corresponding to the period over which the disrupted lower-stratospheric and upper-tropospheric circulation propagates downward. We have added a brief justification of this choice to the manuscript.

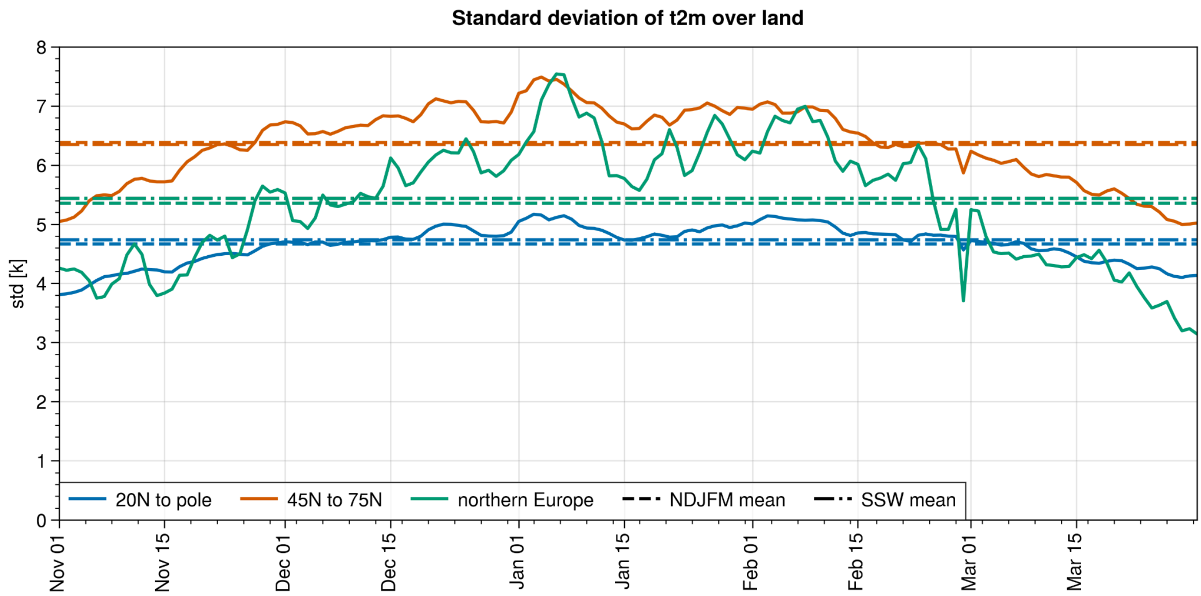


Figure 5: Standard deviation of detrended, deseasonalized t2m over land for each calendar day between November and March for three different spatial domains. Horizontal dashed lines show the extended winter (NDJFM: November -March) mean, horizontal dash-dotted lines show the mean for SSWs, based on each calendar day during the 60-day window following the SSW onset date.

3. line 125. The choice of the small Northern European gridbox appears to have been made a posteriori, after calculation of the initial hemispheric analysis. The several papers cited, apparently to justify this choice, presumably were based on the same or substantially overlapping historical data. The possible impact on the second, spatially restricted, analysis should be discussed more fully.

As noted by the reviewer, the focus on northern Europe was motivated by previous studies documenting a robust SSW surface response in this region (e.g., King et al., 2019; Kolstad et al., 2010; Monnin et al., 2022), which are indeed based on very similar data. However, we would like to highlight that the specific domain boundaries were also informed by inspection of the hemispheric t2m composites, which is precisely the kind of a posteriori domain selection our manuscript cautions against. We consider this example self-illustrating: it demonstrates that such choices arise naturally in practice, even when the underlying signal is physically well-motivated and supported by other literature.

We further note that the key result, i.e., the substantial increase in the FDR-adjusted threshold upon domain restriction, is not sensitive to the precise domain boundaries. Figures 6 and 7 show qualitatively consistent behaviour for alternative domain definitions, confirming that the effect follows from the FDR correction method rather than from this specific regional choice. We have made this explicit in the manuscript and have also added a sentence acknowledging the a posteriori nature of the domain selection.

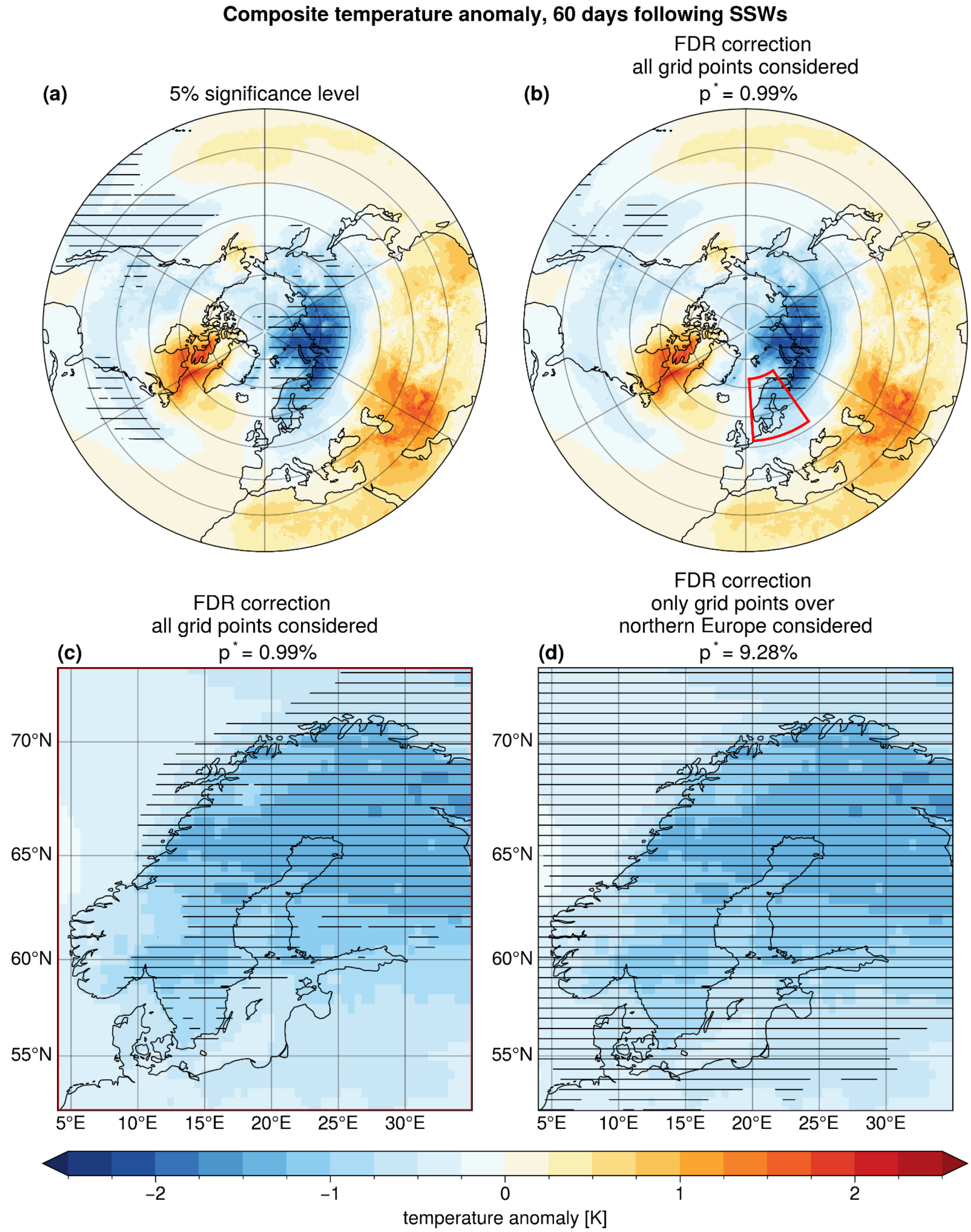


Figure 6: As Fig. 1 in the manuscript, but for a slightly larger northern European domain ranging between 4°E - 35°E, and 52°N - 73°N.

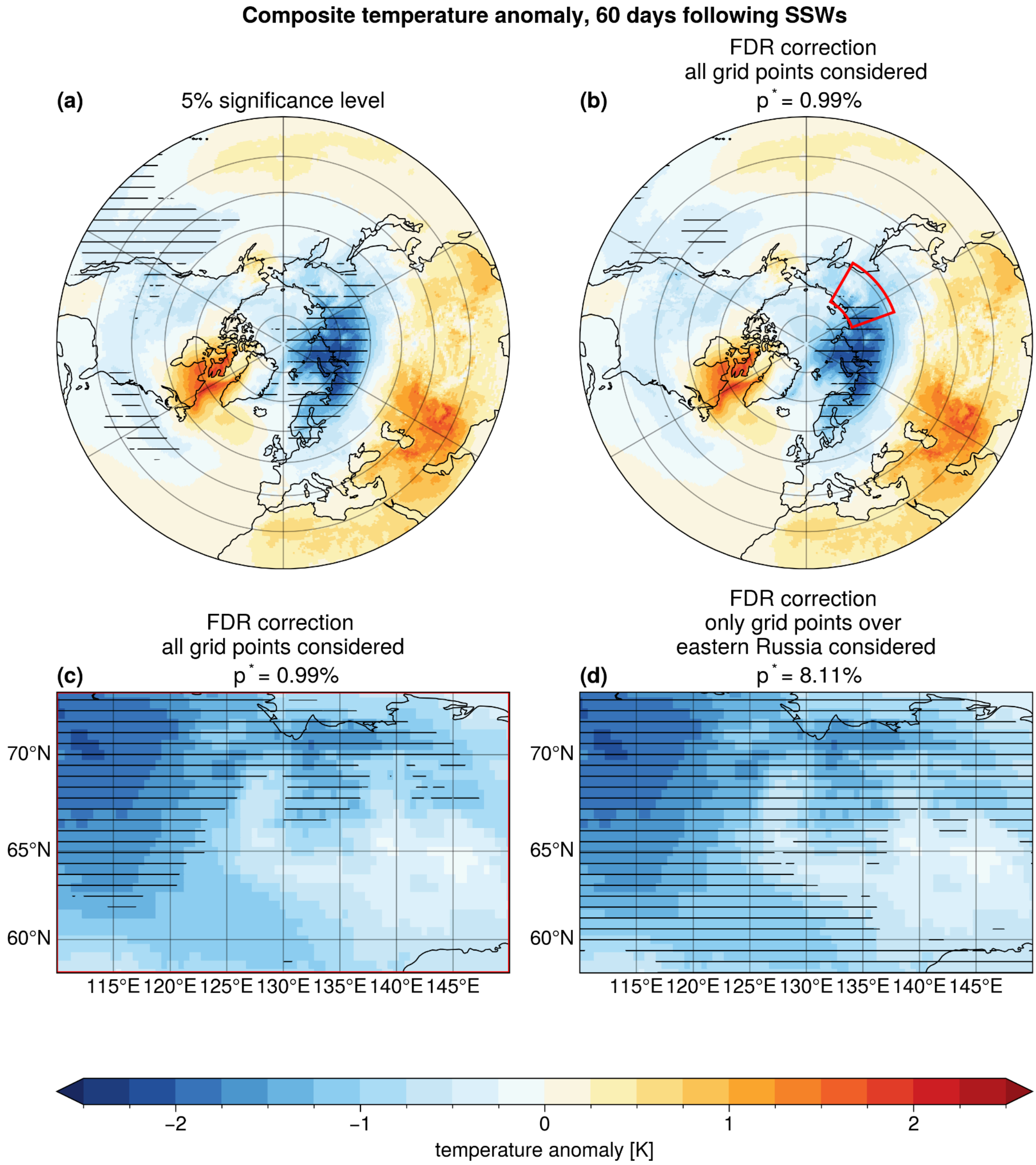


Figure 7: As Fig. 1 in the manuscript, but for a domain across eastern Russia ranging between 110°E - 150°E, and 58°N - 73°N.

## References

Baldwin, M. P. and Dunkerton, T. J.: Stratospheric Harbingers of Anomalous Weather Regimes, Science, 294, 581–584, https://doi.org/10.1126/science.1063315, 2001.

Birner, T. and Albers, J. R.: Sudden Stratospheric Warmings and Anomalous Upward Wave Activity Flux, SOLA,, 13A, 8-12, https://doi.org/10.2151/sola.13A-002, 2017.

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.

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.

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