Reviewer 1

Major analysis/presentation comments:

(C1) The authors focused on point-wise analysis of temperature data and ultimately found no significant local signal related to geomagnetic activity. I wonder, however, if such signal could be more clearly seen in data averaged over larger areas, i.e. obtained for individual sectors, latitudinal bands, or over the entire extratropical area. I base this(possibly unfounded) suspicion on the presence of uniformly positive anomalies across large segments of the analysis area, notable especially for the JJA and SON seasons at the 5 hPa level (Fig. 2). By averaging the temperature series from multiple grid points,signal-to-noise ratio can perhaps be improved; conditional averages considered by the t-test may then more clearly reflect the geomagnetism influence.

In order to answer to both the reviewers we added the section 3.2 to discuss the zonally averaged temperature differences and the new Figure 7

(C2) The autocorrelations seem to be substantial in some of the time series. The authors address their effect through a correction reducing the number of degrees of freedom considered in the t-test. Is there, however, any identifiable source of these autocorrelations (such as a long-term trend, or imprints of solar activity variations)? If so, removal of the respective components from the time series may potentially result in higher (and statistically more significant) contrast between temperatures pertaining to low/high geomagnetic activity periods.

There may be many causes for autocorrelation. We added a short discussion, see lines 110-115.

(C3) To quantify and visualize presence of autocorrelations in the temperature data,statistic of the Durbin-Watson (DW) test is shown in Fig. 1. Maybe presenting the lag-1autocorrelations instead of (or in addition to) the DW statistic would better illustrate the autocorrelation structures, as they are directly involved in calculation of the corrections applied in the paper (eq. (1)), and arguably more intuitively interpretable than the values of the DW statistic itself.

We added the lag-1 autocorrelation together the DW test result (see new Fig. 1)

(C4) A requirement of Gaussianity is mentioned with regard to the t-test (l. 88), but, un-like other test assumptions, it is not tackled any further. I assume that this assumptionis reasonably well satisfied, considering consistence of the data with AR(1) model (asdiscussed in the paragraph at l. 104+), but perhaps this could be mentioned explicitly?

We discussed this issue at lines 91-93 of the new manuscript.

It is however well known that the t-test, which assumes a statistical model where observations are statistically independent and it is widely, but incorrectly, believed that the t-test is valid only for normally distributed outcomes. Several authors (Efron, 1969; De Winter, 2013; Poncet et al., 2016) have shown that the t-test is suitable under symmetric, not necessarily normal, and asymmetric distributions.

(C5) For better comparability with topically close studies (especially Seppälä et al.(2009), by which much of the methodology in the current manuscript seems to beinspired), maybe results for lower atmospheric levels could also be shown/mentioned.

We discussed this point in the new section 3.2 and showing the results in Fig. 8.

(C6) Fig. 2: The positions of grid points with statistically significant negative temperature differences (and their corresponding purple outline) seem suspicious: instead of being located within the areas pertaining to negative differences, they appear near thel ine separating the + and – regions

CORRECTED

Minor/technical remarks:l. 19: **CORRECTED**
“and is thereby” to “and are thereby” **CORRECTED**
Table 1: “2001” misspelled as “20001” **CORRECTED**
l. 74-75: Did Seppälä et al. (2009) really use daily-step data in their analysis?

l. 85: maybe reference to Benjamini and Hochberg (1995) would be preferable here, as they are the original authors of the FDR method (as discussed later in the manuscript) **CORRECTED**

l. 144: Shouldn’t there be J rather than i in the numerator of the fraction? **CORRECTED It is not necessary to define J.**

l. 146: Welch’s variant of the t-test is mentioned (i.e., the form assuming unequal variances of the samples compared), yet t-test employing pooled variance is presented earlier in the text (eq. (2))

We discussed this issue in the interactive discussion. The Welch’s test is always applied even when there is no correction, we say that at line 72.

Fig. 3: The green outlines seem to be only partially drawn **CORRECTED. Now there is a gray area where differences are significant.**

l. 163: “point” to “points”l. 190: extra comma **CORRECTED**

Reviewer 2

Major comments:
-Statistical significance and physical link
This study showed that the stratospheric temperature response to the geomagnetica ctivity was not statistically significant. Although it may be due to no physical link between them, it may be due to insufficient data length or too large internal temperature variations. The authors should mention that statistical insignificance does not deny an existence of the physical link.

We stress this point in the new manuscript at pag. 7 – lines 197-199
It is natural to think that EEP would influence upper and mid-stratosphere temperatures through its impact on ozone. The results discussed in the previous sections suggest that the EPP influence on NH stratospheric temperatures is problematic to detect as it is much weaker than other causes of variability, among which the internal dynamical variability is paramount.

And at the end of conclusions, pag. 8 lines 231-232
I*t is clear that the absence or the presence of significance does not put an end to the research of a possible relationship between EEP and stratospheric temperature, that we suppose to be weak and consequently difficult to detect.*

-Zonal-mean temperatureAlthough this study is motivated by S09, the analyzed pressure levels were different(i.e., surface in S09 and stratosphere in this study). On the other hand, several previous studies examined geomagnetic activity impacts on stratosphere temperature, but only for zonal-mean temperature to my knowledge. In order to clarify whether this result canbe applied to zonal-mean fields or not, I recommend showing the result for zonal-mean temperature in addition to the horizontal distribution.

This suggestion was implemented adding the section 3.2 at pag. 7 and new figure 7.

-Ap index and F10.7
In this study (and S09), the Ap index was used to distinguish high and low geomagneticactivity years. Is there a potential that the correlation between Ap index and solar activity (i.e., F10.7) affects the result?-

Data length. In this study, the data between 1958-2006 was used to compare the result with S09. If the data period is extended to 2018 or 2019, does it affect the result?

As we already said in the interactive discussion we have a submitted paper based on sensitivity experiments where we face these questions. We can share here our conclusions that Solar activity alone has a significant impact on the ozone chemistry and on mesospheric and stratospheric temperature, whereas GA hasn't. However, there is a mutual interaction between SSI and GA and this happens when they are both high and act together.
In our opinion, observations do not allow to evaluate such complex interactions.

Minor comments:
-p.1, l.11-21Previous studies are not adequately cited. At least, references about energetic particle precipitation into the thermo/mesosphere and long lifetime of polar-night NOx should be added.
ADDED (see lines 13-22)
-p.2, l.29"Forecast" -> "Forecasts" **CORRECTED**
-p.2, l.55"2005" -> "2015" **CORRECTED**
-p.3, l.65"of S09" -> "as S09"-p.3, **CORRECTED**
l.82 and p.6, l.175"Wilks (2016)" -> "(Wilks, 2016)" **CORRECTED**
-p.4, l.91"use" -> "use of" **CORRECTED**

-p.4, l.95 Why were 10 and 5 hPa levels chosen? While 10 hPa is representative of middlestratosphere, it seems that 1 and 100 hPa levels are appropriate as representativelevels of upper and lower stratosphere, respectively.

This comment was implemented removing the 10 hPa and adding 1 and 100 hPa. We kept the 5 hPa level instead of the 10 hPa level because it teaches us that the application of the only FDR procedure sometimes is not enough and both the corrections have to be applied.

-p.4, l.109-110 Why the AR(1) process is suitable for explaining a cumulative impact is not clear to me. Please explain it in more detail.

We discussed this point in the interactive discussion.

-p.5, l.143"equatl" -> "equal" **CORRECTED**

-p.8Lu et al. and Long et al. should be reversed in order. **CORRECTED**
-p.10"20001" -> "2001" **CORRECTED**
-p. 12-14Units in temperature should be added. **ADDED**

A note on the statistical evidence for an influence of geomagnetic activity on JRA-55 northern hemisphere seasonal-mean stratospheric temperatures

Nazario Tartaglione[1,2], Thomas Toniazzo[1,2], Yvan Orsolini[2,3,4], and Odd Helge Otterå[1,2]

[1]NORCE Climate, Bergen, Norway
[2]Bjerkness Centre for Climate Research, Bergen, Norway
[3]Birkeland Centre for Space Science, University of Bergen, Bergen, Norway
[4]Norwegian Institute for Air Research, Kjeller, Norway

Correspondence: Nazario Tartaglione (nazario.tartaglione@norceresearch.no)

Abstract. We employ JRA-55, a recent second-generation global reanalysis providing data of high-quality in the stratosphere, to examine whether a distinguishable effect of geomagnetic activity on northern hemisphere stratospheric temperatures can be detected. We focus on how the statistical significance of stratospheric temperature differences may be robustly assessed during years with high and low geomagnetic activity. Two problems must be overcome. The first is the temporal autocorrelation of the data, which is addressed with a correction of the t-statistics by means of the estimate of the number of independent values in the series of correlated values. The second is the problem of multiplicity due to strong spatial autocorrelations, which is addressed by means of a false discovery rate (FDR) procedure. We find that the statistical tests fail to formally reject the null hypothesis, i.e. no significant response to geomagnetic activity can be found in the seasonal-mean northern-hemisphere stratospheric temperature record.

1 Introduction

There is a large interest in the potential climate impact of geomagnetic activity. One of the main mechanisms by which geomagnetic activity is thought to affect the middle atmosphere is through the production of nitrogen oxides (NO_x's), either by the continuous precipitation of auroral electrons penetrating into the lower thermosphere (Sinnhuber et al., 2012) or by the more episodic precipitation of higher energy electrons into the mesosphere (Andersson et al., 2014; Päivärinta et al., 2016). Downward transport from the mesosphere to the stratosphere in winter results in increased availability of NO_x in the dark polar stratosphere, where it is long-lived. NO_x can catalytically reduce ozone concentrations as the sun returns (Brasseur and Solomon, 1986; Callis et a and thus alter radiative heating rates, with potential observable impacts on stratospheric temperatures and possible implications also for surface air temperature (SAT). The amount of ~~NOx~~ NO_x in the middle atmosphere during late winter and spring depends on the ~~cumulated~~ cumulative effect of geomagnetic activity over the preceding months on the ~~NOx reservoir~~ NO_x reservoir (Jacob, 1999). Stratospheric NO_x concentrations however also depend on the magnitude of the downward transport from this reservoir, and are thereby affected by internal variability of the atmospheric circulation from year to year, especially in the Northern Hemisphere (NH) (Funke et al., 2005; Randall et al., 2006; Päivärinta et al., 2016).

The impact of energetic electron precipitation (EEP) driven by geomagnetic activity on NO_x and ozone concentrations has been well-documented after detailed satellite studies were carried out in the early 2000's ~~(Funke et al., 2005; Randall et al., 2005)~~(Funke et al., 2

Several recent studies ~~(Baumgaertner et al., 2010; Bucha, 2014; Lu et al., 2008; Seppälä et al., 2009, 2013)~~ (Baumgaertner et al., 2010; Bu a significant signal associated with geomagnetic activity in the observed climate. However, there remains considerable uncertainty regarding the precise attribution of such signal, and the existence of a direct link between EEP and stratospheric and tropospheric temperatures has remained controversial. Among others, the study of Seppälä et al. (2009)~~;~~ , henceforth S09, in particular, claims to find a significant, direct relationship between SAT and geomagnetic activity based on reanalysis data from the European Centre for Medium Range Weather Forecasts (ECMWF). In essence, S09 finds that the hypothesis that geomagnetic activity influences the SAT is supported by reanalysis data, whereas the null hypothesis that the SAT is not influenced by the geomagnetic activity at all is rejected. S09 compares seasonal SAT in years with high and low geomagnetic activity, and also considered the separate effect of the variation in solar irradiance associated with the 11-year heliomagnetic cycle.

The selection of years in S09 was based on two indices, Ap and f10.7. Ap (Rostoker, 1972) provides a measure for daily average level of geomagnetic activity. To account for cumulative effect of ~~NOx~~ NO_x production, transport and diffusion processes, Ap was commonly averaged over 4 months from late autumn to winter (Seppälä et al., 2009; Funke et al., 2014; Tomikawa, 2017). In particular, S09 used Ap averaged between October and January to define winters of high and low geomagnetic activity in the northern hemisphere. The second index, f10.7 (https://www.swpc.noaa.gov/phenomena/f107-cm-radio-emissions), is an indicator of the phase and intensity of the solar cycle. By compositing separately on the basis of Ap and f10.7, S09 obtained different samples of seasonal-mean data for years with high geomagnetic activity and for years with low geomagnetic activity. They then computed the SAT differences of the seasonal means (DJF, MAM, JJA, SON) between the two samples, and employed a t-test based on the set of daily-means (Seppälä, personal communication) used to compute the seasonal averages to discriminate against a null hypothesis of no effect. As a consequence of such procedure, S09's claimed of significance is marred by the presence of very strong temporal and spatial autocorrelation within the samples.

In this paper, we revisit the S09 hypothesis by adopting a rigorous methodology for significance testing on strongly autocorrelated data. We focus on wintertime stratospheric temperatures between 200 hPa and 1 hPa, a pre-requisite for possible surface impacts associated with EEP-related changes in ozone concentrations. Although our analysis is focused, in this paper on stratospheric temperature, we look at all the levels present in the dataset. We show that statistical testing appropriate to the data at hand is a crucial step in any analysis purporting to demonstrate an observed climate signal of geomagnetic activity.

Data and methods are described in section 2, including a discussion on the problem of autocorrelation in time and space. In section 3, the results obtained by applying the t-test to the stratospheric temperatures are shown. The analysis is applied to four different cases: with no correction at all, with the temporal and the spatial autocorrelation correction applied separately, and with both the corrections applied. In Section 4 conclusions are drawn.

2 Data and Methods

2.1 Data

To analyze the possible impact of geomagnetic activity in the stratosphere, we use the Japanese 55-year ~~Reanalysis~~ reanalysis (JRA-55) covering more than 55 years, extending from 1958 to the present (Kobayashi et al., 2015). Due to the selection of cases of high and low geomagnetic activity as in S09, only data up to 2006 is used here. In JRA-55 reanalysis ozone is used interactively in the radiation code, although it is treated differently in the pre- and post-1979 satellite era. This is an important asset for JRA-55 since the EEP will primarily affect ~~NOx~~ NO_x and ozone, and this feature is not commonly found in other reanalysis systems, such as the ECMWF reanalyses. Older generation reanalyses tend to suffer from temporal inhomogeneities because of the sequential introduction of new satellite data during the assimilation period, especially in the SH as shown recently by Long et al. (2017). For these various reasons, we restricted our analysis to the recent JRA-55 reanalysis. Tomikawa (2017) also used the JRA-55 reanalyses to investigate the signature of geomagnetic activity, but focused exclusively on the SH. He found a temperature signal in the upper stratosphere, but only in July. The S09 selection shown in Table ~~1~~ 1 is used to compute the significance of the seasonal differences. The criteria used to select the different years are based on the Ap and f10.7 values, and are the same as used by S09. The definition of high and low geomagnetic activity is the same as S09. We hence investigate the potential signatures on stratospheric temperature during the same winters and in the following seasons of the same calendar year as S09 did for SAT. The set of data is denominated N1 as in S09 (Table 1).

2.2 Data autocorrelation and statistical significance

S09 computed the SAT differences of the seasonal means (DJF, MAM, JJA, SON) between those selected high Ap and low Ap years, and employed the Welch's t-test (hereafter only t-test) to assess the likelihood of the differences given a null hypothesis of no effect. Such a test assumes a statistical model in which observations are normally distributed and statistically independent. In particular, the t-test is sensitive to the temporal autocorrelation or serial correlation within the samples. When serial correlation is not taken into account in the data, statistically significant differences in two means, which may not be different at all, are found more frequently than expected (Zwiers and von Storch, 1995). S09's analysis is affected by this problem, because ~~they~~ the authors used daily-mean data in their t-test~~(personal communication)~~, which are highly autocorrelated in time. As seasonal averages can still suffer from temporal autocorrelation, the serial dependence is checked by means of the Durbin-Watson test (Durbin and Watson, 1950). While the serial correlation, in general, is reduced from seasonal averaging, it can still persist, especially in summer. To deal with such serial correlation a correction is applied as suggested by Zwiers and von Storch (1995). The temporal autocorrelation is not the only potential caveat that needs to be considered when testing a hypothesis. When performing a significance test simultaneously on many samples one will at some point find statistically significant ~~points~~ temperature differences simply by accident. Unfortunately, the dominant approach to the multiplicity problem is generally to test the single grid points and then to report them as "significant" when the null hypothesis is locally rejected ~~Wilks (2016)~~(Wilks, 2016). Sometimes temporal and spatial autocorrelation is not addressed at all~~. However~~, but , there are some exceptions. Maliniemi et al. (2014), for instance, trying to find a relationship ~~between~~ between solar activity

and surface air temperature dealt with temporal and spatial autocorrelation using a Monte Carlo approach. To overcome this multiplicity problem ~~, Wilks (2016) suggests to use~~ in our analysis, we apply the false discovery rate controlling ~~procedure (Benjamini and Hochberg, 1995)~~introduced by Benjamini and Hochberg (1995) and proposed in the atmospheric sciences by (Wilks, 2006, 2016).

2.3 Accounting for temporal autocorrelation

The t-test is a widely used method for hypothesis testing within the climate community. It is however well known that the t-test, which assumes a statistical model where observations are statistically independent and ~~Gaussian,~~ it is widely, but incorrectly, believed that the t-test is valid only for normally distributed outcomes. Several authors (Efron, 1969; De Winter, 2013; Poncet et al., 2016) h shown that the t-test is suitable under symmetric, not necessarily normal, and asymmetric distributions. The t-test is sensitive to time autocorrelation or serial correlation within the samples. The effect of serial correlation is, usually, to make comparisons of means too liberal. The null hypothesis assuming equal means is hence rejected more frequently than expected. Two separate reasons favor the use of seasonal-mean data instead of daily-mean data. The first reason is that any influence of EEP on temperature is expected to accumulate over seasonal time scales. The second reason is that daily temperatures are strongly serially correlated, whereas seasonal data have less correlation between two consecutive years ~~,~~ for instance. In fact, one of the causes of the serial correlation is that the variable of interest varies seasonally. Nevertheless, even for seasonal means it is important to account for serial correlations, as there may be other causes leading temporal autocorrelation, including persistence. Fig. 1a that shows the results of the Durbin-Watson test (Durbin and Watson, 1950) applied at the seasonal temperatures at 5 ~~and 10 hPa. The~~ hPa. Similar pictures can be obtained by plotting the lag-1 autocorrelation (Fig. 1b), but the Durbin-Watson test, which is a classical test to check whether data are serially correlated~~. In this case~~, is better, compared to including the lagged response, as it tests for autocorrelation in the residuals and it is suitable when in time series there are trends or seasonal patterns. When data are serially correlated, the test gives values close to zero~~;~~ , whereas when data are not correlated ~~at all,~~ the test statistic values, as a rule of thumbs, are in the range of 1.5 to 2.5. There is also the possibility of serial anti-correlation: in such a case, the value would be above 2.5, but this situation was not found in our study.

During the winter and spring seasons, the data generally do not have a very strong temporal autocorrelation, and the t-test can be applied with a lower risk of obtaining false positive outcomes. ~~However, there~~ There are some regions where the temporal autocorrelation still persists, such as over North America. ~~The data are very auto-correlated during~~ Local local higher autocorrelation values during other seasons can also be a result due also to low frequency variance caused by large scale teleconnections (Madden, 1977). During the summer season ~~,~~ data are very autocorrelated - and to a large extent also in autumn, but they will be analyzed in any case as it is worthwhile as well to show how the procedure used to assess the possible impact of the geomagnetic activity responds to serially correlated data. In general, autocorrelation is mainly due to persistence of temperature patterns year by year. For instance, this is the case for example of the large value of temperature autocorrelation found during the summer season. However, we cannot exclude other causes, including a possible impact of the solar activity.

Serial correlation can be corrected for by adopting, for example, the strategy suggested by Zwiers and von Storch (1995). This procedure is valid under the assumption that the time series, from which the data are sampled can be ~~modeled~~ modelled as

an autoregressive process of order 1 or AR(1). Vyushin et al. (2012) have shown that the AR(1) representation fits modeled stratospheric temperature data very well according to standard goodness of fit tests. Seidel and Lanzante (2004) found a similar result with temperature observed by radiosondes and satellites.

If EEP has a cumulative impact during the different seasons, it has to be shown that the means of two subsets with high (H) and low (L) Ap values from the set N1 must be different.

To test the null hypothesis ~~of equal means~~ $H_0 : \mu_H = \mu_L$ with the t-statistics at the 5% significance level one let's apply the t-test under the condition that the standard deviation is scaled by the equivalent sample sizes m_e and n_e that can be computed, by:

$$n_e = n\left(\frac{1-\rho_1}{1+\rho_1}\right) \tag{1}$$

where n is the original size of one out of two samples and ρ_1 is the parameter of the AR(1) process representing the autocorrelation at lag 1; and similar for m_e. The t-test is then corrected in the following way:

$$t = \frac{\bar{H} - \bar{L}}{s\left(\frac{1}{\sqrt{m_e}} + \frac{1}{\sqrt{n_e}}\right)} \tag{2}$$

where $\bar{H}$ and $\bar{L}$ are the sample averages and s^2 is the pooled variance

$$s^2 = \frac{\sum_{i=1}^{m}\left(H_i - \bar{H}\right)^2 + \sum_{i=1}^{m}\left(L_i - \bar{L}\right)^2}{m+n-2} \tag{3}$$

2.4 Accounting for spatial autocorrelation

Spatial autocorrelation produces the so-called multiplicity problem, which arises when testing a statistical hypothesis on many samples (the domain's grid points, in our case) simultaneously. A single hypothesis test allows for a null hypothesis ~~against~~ and an alternative hypothesis~~, which~~. The alternative hypothesis will be favored when an extreme value, usually with a probability (called value) that is less than 5% is found (Wilks, 2016). Making a statistical test on multiple points, for example within a spatial domain, means that more realizations will be available and there will be many grid points where one is more likely to reject the null hypothesis. In an ideal situation, where the value is set to 0.05 and each point is statistically independent of the others, it is expected to find that 5% of the points will be statistically significant by accident. The situation is worse when the grid points are correlated, as is often the case when analyzing meteorological and climate data. This problem, known in the literature as the multiplicity problem, has been encountered in several studies, although most of the studies in the atmospheric science have not properly addressed the issue yet (Wilks, 2016). Some solutions have been proposed, each having their own advantages and disadvantages. Wilks (2016) gives a brief historical outline and shows different solutions to this problem. One technique to address this issue is by using the false discovery rate (Benjamini and Hochberg, 1995). According to Wilks (2006, 2016) the false discovery rate is the expectation of the fraction of true null hypothesis rejections among all the rejections and it

is the best available approach to analyze multiple hypothesis test results, even when those results are mutually correlated. ~~According to~~

As stated by Wilks (2016) the FDR procedure requires smaller values to reject the local null hypothesis arising the standard of the test. For the sake of the reader we will describe the FDR algorithm as described in Wilks (2016). The algorithm operates on the collection of $H_0 : \mu_H = \mu_L$ values from m_e (number of grid points) local hypothesis tests p_i, with $i = 1, ..., N$, which are sorted in ascending order. Rejection of the test happens when the p_i values are not larger than a threshold level p_{FDR} that is a function of the distribution of the sorted p_i values. More specifically to define which values pass the test the following formula is used:

$$\left[p_i : p_i \leqslant \alpha_{FDR}\left(\frac{i}{N}\right)\right]$$

where α_{FDR} is the chosen FDR control level that here is taken equal to 0.05. For a given value of α_{FDR} , the largest value of i, ~~let's say J, such that $p_J \leqslant \alpha_{FDR}\left(\frac{i}{N}\right)$~~ such that $p_i \leqslant \alpha_{FDR}\left(\frac{i}{N}\right)$ defines the threshold below which the local null hypotheses are rejected. the largest value of i, such that

3 Results

~~As can be seen in Fig. 2, the Welch's~~

3.1 Stratospheric levels

We start with the application of the t-test on ~~the stratospheric (~~5 ~~and 10 hPa) temperature shows that there~~ hPa temperature (Fig. 2), which represents the level where the statistically significant area is the largest among all the examined pressure levels. There are large areas with ~~significant points,~~ statistically significant temperature difference at 5% level ~~, considering a distribution with two tails, during the winter and the~~ especially during winter and summer.

At 5 hPa, the area with significant ~~points~~ differences covers most of the hemisphere in JJA, but, as can be seen from the analysis of the Durbin-Watson test, the ~~summertime~~ summer season exhibits a large temporal autocorrelation. Hence, the ~~significant points~~ statistically significant areas observed in JJA should originate from this autocorrelation. In winter, the area with significant points cover North America, another region where the Durbin-Watson test suggests serial correlation. ~~At 5 hPa, the area with significant points cover most of the hemisphere in JJA.~~ It is clear from Fig. 2 that a possible impact, of the geomagnetic activity, if it exists, would be limited at higher latitudes, from 40° to 90°.~~Hence, the corrections described earlier concerning the spatial and temporal autocorrelations were applied using the p-values of those latitudes since low latitudes are dominated by large statistically insignificant areas.~~

~~Thus~~Because of the strong temporal autocorrelation, it is expected that at least in summer ~~, these significant points~~ these significant differences should be false positive outcomes and they should be reduced or completely removed when applying the serial correlation correction~~(Fig. 3)~~. In fact, by applying the correction of serial dependence to the 5 hPa temperature

differences the t-test results change dramatically ~~, with almost all the significant points~~ as Fig. 3 shows. The statistically significant differences are removed everywhere in JJA. However, in DJF ~~a few significant points~~ small areas with significant differences are still present at ~~the 5 and 10 hPa levels~~that level. The Durbin-Watson test somehow predicted that there will be not significant points after applying the Zweirs and von Storch algorithm in the areas where the Durbin-Watson test value was close to zero.

On the other hand, the problem of multiplicity is solved here by means of the FDR procedure described in section 2. When applying such a procedure without correcting the serial dependence ~~all the significant points at 10 hPa disappear, while some significant points~~ e some significant temperature differences still persist at 5 hPa during summertime (Fig. 4). However, the only application of FDR remove all significant differences when it is applied to other pressure levels (e.g. 10 hPa). This result is important as the FDR procedure is quite powerful in removing most of the false positive ~~points but, clearly, it cannot be sufficient as the~~ differences but, how Fig. 2 shows it is not sufficient in presence of a strong temporal correlation that can still leave ~~grid points~~ regions where the t-test rejects the null hypothesis when, in fact, it would be true. This result is particularly important and it recommends the application of both the corrections strongly.

The application of such corrections dealing both with temporal and spatial autocorrelation removes all the ~~significant points~~ statistically significant differences in the domain (~~not shown as it would be the same figure as~~ Fig. 2~~, but with no significant points~~b) and the ~~combined tests fail~~ t-test with the combined correction fails to reject the null hypothesis.~~A similar conclusion is obtained with temperature differences at other stratospheric levels (not shown), ranging from 100 to~~

A similar result is obtained for all the other levels in the dataset, temperature differences at 1 ~~hPa. At those levels, the areas with significant points are even smaller than those at 5 or 10 hPa~~and 100 hPa temperatures are shown in Fig. 5 and Fig. 6 without and with both the corrections. The application of the false discovery rate on those fields eliminates all the significant ~~points~~temperature differences, showing that also at those levels there is no detectable impact of geomagnetic activity on the ~~stratospheric temperature~~atmospheric temperature.

3.2 Zonally averaged temperature and 2 m temperatures.

Several studies have shown the possible impact of EEP or energetic particle precipitation on the observations using zonal mean temperatures (Tomikawa, 2017; Seppälä et al., 2013). Thus, we show how without any correction even the zonal mean temperature difference has areas that are statistically significant at 5% level (Fig. 7a). In particular, there are statistically significant areas in all the seasons, but spring, between 10 and 1 hPa. There are no statically significant area (Fig. 7b) after applying the two corrections that account for spatial and temporal autocorrelations.

It is natural to think that EEP would influence upper and mid-stratosphere temperatures through its impact on ozone. The results discussed in the previous sections suggest that the EPP influence on NH stratospheric temperatures is problematic to detect as it is much weaker than other causes of variability, among which the internal dynamical variability is paramount. As this work is motived by S09 that analyzed the 2 m temperature, Fig. 8a shows the 2m temperature difference (High Ap – Low Ap) without any correction. There are large areas where the seasonal temperature differences are statistically significant at 5%

level.

The application of both the spatial and temporal autocorrelation corrections remove almost all these areas. However, some small areas of statistically significant temperature differences are still present. They are in the polar region and over Russia during the winter season and over the Scandinavia during the Spring (Fig. 8b). As it is not easy to explain these statistically significant surface temperature differences with a causal relationship with EEP, given the lack of signal aloft, there may be some other reasons that can justify this significance with other causes, among which a positive outcome obtained by chance.

4 Conclusions

Climate data often exhibit temporal and spatial autocorrelations which should be taken into account when testing ~~an~~ a hypothesis, a task that is often neglected ~~Wilks (2016)~~(Wilks, 2016). The effect of temporal autocorrelation was addressed with an appropriate procedure described in Zwiers and von Storch (1995) . The problem of evaluating results of multiple hypothesis tests in a spatial domain was further addressed by means of the false discovery rate procedure. In this paper, the possible impact of geomagnetic activity on the seasonal-mean stratospheric temperature in the JRA-55 reanalysis was evaluated by means a Welch's t-test under four different cases: 1) with no correction of temporal and spatial autocorrelation, 2) with correction on temporal autocorrelation only, 3) with correction on spatial autocorrelation only, and finally 4) with both the corrections. Most of the cases examined show significant points when temporal and spatial autocorrelations are not corrected, while not showing any significant point when including just one out of the two corrections. In other words, in most cases, there is not even a need to apply both corrections to infer that there is no impact of geomagnetic activity. However, the statistically significant temperature differences at 5 hPa show that it strongly recommended the application of both the corrections for the spatial and temporal autocorrelation. In some cases, like for the JJA temperature difference at ~~10 hPa(Fig. 4), there were~~ 5 hPa, there are a few significant ~~points~~ areas remaining when applying one out of the two corrections (Figs 3 and Fig. 4), but those significant ~~points~~ areas disappeared when both corrections were applied. ~~Thus, when applying~~ Finally, the procedures to take into account these autocorrelations, the significance test typically fails to reject the null hypothesis. ~~We~~ This result is found for all the pressure levels analyzed and for zonally averaged temperature. The only temperature field that has still statistically significant differences after applying both the corrections is the 2m temperature. There are two seasons, DJF and MAM, where small statistically significant areas are present in the polar region. In absence of a signature aloft, we therefore conclude that, based on the JRA-55 reanalyses, not enough evidence is available at present to suggest that the null hypothesis of no impact of geomagnetic activity on NH stratospheric temperatures is false. A remaining caveat concerns the definition of seasons of high or low geomagnetic activity, which is here the same as in S09 and is based on a lagged 4-month averaged Ap index, (i.e., from October to January for wintertime geomagnetic activity). Some sensitivity studies to this definition, e.g., to treat more intense shorter episodes of EEP or to treat differently the seasonal lag or accumulation of EEP, is certainly warranted for future studies. It is clear that the absence or the presence of significance does not put an end to the research of a possible relationship between EEP and stratospheric temperature, that we suppose to be weak and consequently difficult to detect.

Data availability. Data can be downloaded from the Meteorological Research Institute/Japan Meteorological Agency/Japan or from Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory

Author contributions. NT performed the statistical tests, all the authors contributed in the interpretation of the data and wrote the paper.

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

Acknowledgements. This work was funded by the project SOLENA - Solar effects on natural climate variability in the North Atlantic and Arctic - Research Council of Norway, Program for Space Research Project: 255276/E10. The authors acknowledge the Meteorological Research Institute/Japan Meteorological Agency/Japan for the JRA-55 reanalysis. We are extremely grateful to two reviewers for their valuable comments, corrections and suggestions.

References

Andersson, M. E., Verronen, P. T., Rodger, C. J., Clilverd, M. A., and Seppälä A: Missing driver in the connection from energetic electron precipitation impacts mesospheric ozone, Nat. Commun., 5, 5197, https://doi.org/10.1038/ncomms6197, 2014.

Baumgaertner, A. J. G., Jöckel, P., Dameris, M., and Crutzen, P. J.: Will climate change increase ozone depletion from low-energy-electron precipitation?, Atmos. Chem. Phys., 10, 9647-9656, https://doi.org/10.5194/acp-10-9647-2010, 2010.

Benjamini, Y., and Hochberg,Y.: Controlling the false ~~discoveryrate~~discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal ~~StatisticalSociety, SeriesB~~Statistical Society, Series B, 57, 289-300, 1995.

Brasseur, G. P. and Solomon, S.: Aeronomy of the Middle Atmosphere, D. Reidel Publishing Company, 2nd revised edn., 1986.

Bucha, V.: Geomagnetic activity and North Atlantic Oscillation. Stud. Geophys. Geod., 58, 461-472. https://doi.org/10.1007/s11200-014-0508-z, 2014.

~~Durbin~~

Callis, L. B., Natarajan, M., Lambeth, J. D., and Baker, D. N.: Solar-atmospheric coupling by electrons (SOLACE). 2. Calculated stratospheric effects of precipitating electrons, 1979–1988,J. Geophys. Res., 103, 28421–28438, https://doi.org/10.1029/98JD02407, 1998.

de Winter, J. C. F.: "Using the Student's t-test with extremely small sample sizes", Practical Assessment, Research, and Evaluation: Vol. 18, Article 10, 2013.

Durbin, J., and Watson, G. S.: Testing for Serial Correlation in Least Squares Regression, I. Biometrika. 37 (3–4): 409–428. ~~doi:~~https://doi.org/10.1093/biomet/37.3-4.409, 1950.

Efron, B.: Student's t-Test Under Symmetry Conditions. Journal of the American Statistical Association.Vol. 64, 1278-1302, 1969.

Funke, B., López-Puertas, M., Gil-López, S., von Clarmann, T., Stiller, G. P., Fischer, H., and Kellmann, S.: Downward transport of upper atmospheric NO_x into the polar strato sphere and lower mesosphere during the Antarctic 2003 and Arctic 2002/2003 winters, J. Geophys. Res., 110, D24308, ~~doi:~~https://doi.org/10.1029/2005JD006463, 2005.

Funke, B., Puertas, M. L., Holt, L., Randall, C. E., Stiller, G. P., and von Clarmann, T.: Hemispheric distributions and interannual variability of NOy produced by energetic particle precipitation in 2002–2012, J. Geophys. Res., 119, 13565–13582, doi:10.1002/2014JD022423, 2014.

Jacob, D. J.: Introduction to Atmospheric Chemistry, Princeton Uni-versity Press, Princeton New Jersey, 1999.

Kobayashi, S., and Coauthors: The JRA-55 reanalysis: General specifications and basic characteristics. J. Meteor. Soc. Japan, 93, 5–48, doi:https://doi.org/10.2151/jmsj.2015-001, 2015.

Long, C. S., Fujiwara, M., Davis, S., Mitchell, D. M., and Wright, C. J.: Climatology and interannual variability of dynamic variables in multiple reanalyses evaluated by the SPARC Reanalysis Intercomparison Project (S-RIP), Atmos. Chem. Phys., 17, 14593-14629, https://doi.org/10.5194/acp-17-14593-2017, 2017.

Lu, H., Jarvis, M. J., and Hibbins, R.E.: Possible solar wind effect on the northern annular mode and northern hemispheric circulation during winter and spring, J. Geophys. Res., 113, D23104, doi:10.1029/2008JD010848, 2008.

Madden, R. A.: Estimates of the Autocorrelations and Spectra of Seasonal Mean Temperatures over North America. Mon. Wea. Rev., 105, 9–18, https://doi.org/10.1175/1520-0493(1977)105<0009:EOTAAS>2.0.CO;2, 1977.

Maliniemi, V., Asikainen, T. and Mursula K.: Spatial distribution of Northern Hemisphere winter temperatures during different phases of the solar cycle, J. Geophys. Res. Atmos., 119, 9752–9764, doi:10.1002/2013JD021343, 2014.

P¨aiv¨arinta, S.-M., Verronen, P. T., Funke, B.,Gardini, A. Seppälä, A., and Andersson, M.E. : Transport versus energetic particle precipitation: Northern polar stratospheric NO_y and ozone in January–March 2012, J. Geophys. Res. Atmos., 121, 6085–6100, doi:10.1002/2015JD024217, 2016.

Poncet, A., Courvoisier, D. S., Combescure, C., and Perneger, T. V.: Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12(2), 61–71. https://doi.org/10.1027/1614-2241/a000110, 2016.

Randall, C.E., Harvey, V.L., Manney, G.L., Orsolini, Y., Codrescu, M., Sioris, C., Brohede, S., Haley, C.S., Gordley, L.L., Zawodny, J.M., and Russell, J.M.: Stratospheric effects of energetic particle precipitation in 2003–2004, Geophys. Res. Lett.,32, L05802, doi:https://doi.org/10.1029/2004GL022003, 2005.

Randall, C. E., Harvey, V. L., Singleton, C. S., Bernath, P. F., Boone,C. D., and Kozyra, J. U.: Enhanced NO_x in 2006 linked to strong upper stratospheric Arctic vortex, Geophys. Res. Lett., 33,L18811, https://doi.org/10.1029/2006GL027160, 2006.

Rostoker G.: Geomagnetic indices, Rev. Geophys. Space Phys., 10, 935-950, 1972.

Seidel, D. J., and Lanzante, J. R.: An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes, J. Geophys. Res., 109, D14108, doi:10.1029/2003JD004414, 2004.

Seppälä, A., Randall, C.ELu, H., Clilverd, M.A., Rozanov, E., and Rodgerand Rodger, C.J.: Geomagnetic activity and polar surface air temperaturevariability, signatures in wintertime stratosphere wind, temperature, and wave response J. Geophys. Res. , 114, A10312, doi:10.1029Atmos., 118, pp. 2169-2183, 10.1002/2008JA014029, 2009. jgrd.50236, 2013.

Seppälä, A., Lu, HRandall, C.E., Clilverd, M.A., and Rodger. Rozanov, E., and Rodger C.J.: Geomagnetic activity signatures in wintertime stratosphere wind, temperature , and wave response and polar surface air temperature variability, J. Geophys. Res.Atmos., 118, pp.2169-2183, 10.1002, 114, A10312, doi:10.1029/jgrd.50236, 2013. 2008JA014029, 2009.

Sinnhuber, M., Nieder, H., and Wieters, N.: Energetic Particle Precipitation and the Chemistry of the Mesosphere/Lower Thermosphere. Surv. Geophys 33, 1281–1334, https://doi.org/10.1007/s10712-012-9201-3, 2012.

Tomikawa, Y.: Response of the Middle Atmosphere in the Southern Hemisphere to Energetic Particle Precipitation in the Latest Reanalysis Data, SOLA, 13A, 1-7, 2017.

Vyushin, D. I., Kushner, P. J., and Zwiers, F. (2012), Modeling and understanding persistence of climate variability, J. Geophys. Res., 117, D21106, doi:10.1029/2012JD018240, 2012.

Wilks, D.S.: Statistical Methods in the Atmospheric Sciences. Elsevier Science: Burlington, MA, 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, doi:10.1175/bams-d-15-00267.1, 2016.

Zwiers, F.W. and von Storch, H.: Taking Serial Correlation into Account in Tests of the Mean. J. Climate, 8, 336–351, https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2, 1995.

Table 1. Years used to define the N1 set, following S09.

Case	Hemisphere	High Ap years	Low Ap years
N1	NH	1958, 1960, 1961, 1975, 1982, 1984, 1985, 1989, 1990, 1993, 1994, 1995, 2003, 2004, 2005	1962, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1977, 1978, 1980, 1981, 1987, 1988, 1991, 1996, 1997, 1998, 1999, 2001, 2002, 2006

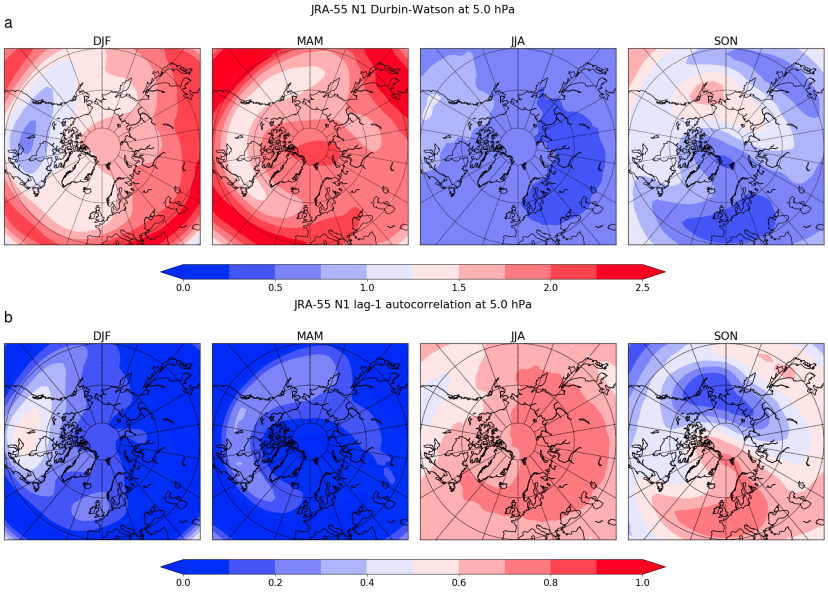

Figure 1. Results of Durbin-Watson test (a) and lag-1 autocorrelation (b) for JRA-55 stratospheric temperature at 5 ~~and 10 hPa~~ for the period between 1958-2006.

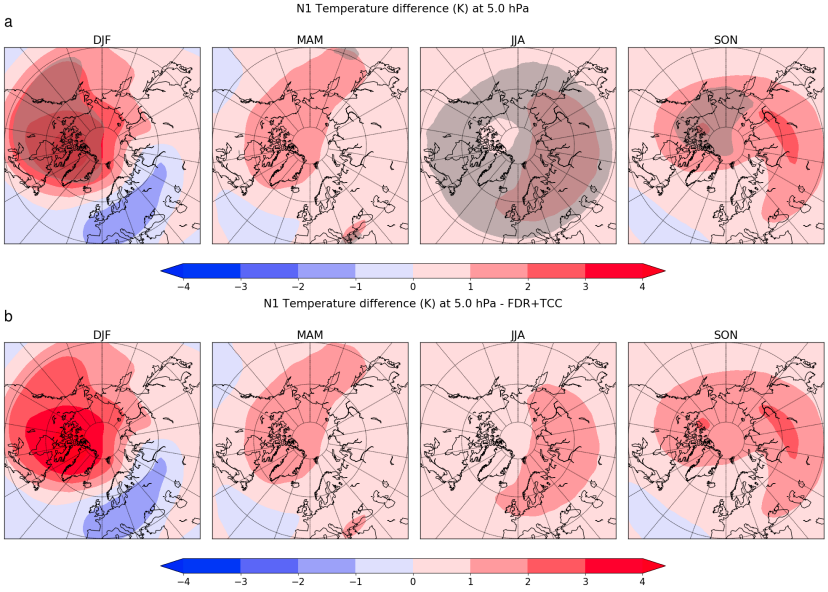

Figure 2. Northern hemisphere seasonal differences in stratospheric temperature (High Ap - Low Ap) at ~~10~~ 5 hPa without (~~bottom~~a) and ~~5 hPa~~ with (~~top~~b) temporal and spatial autocorrelation correction. ~~Dots~~ Gray areas represent statistically significant ~~grid points with~~ temperature differences at the 5% confidence levels~~. Green and violet lines encompass significant positive and negative areas.~~

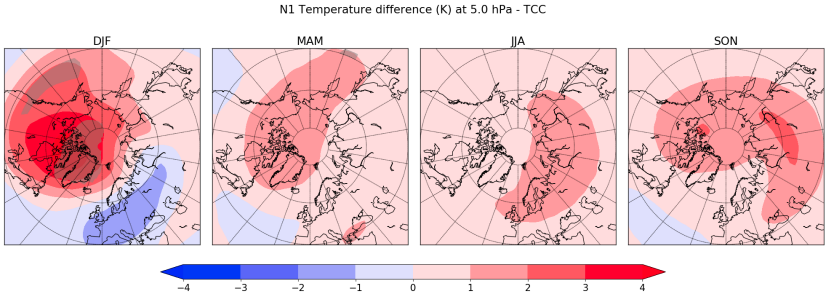

Figure 3. ~~As Figure 2 but~~ Northern hemisphere seasonal differences in stratospheric temperature (High Ap - Low Ap) at 5 hPa after applying the correction for serial dependence. Gray areas indicates statistically significant areas at the 5% confidence level.

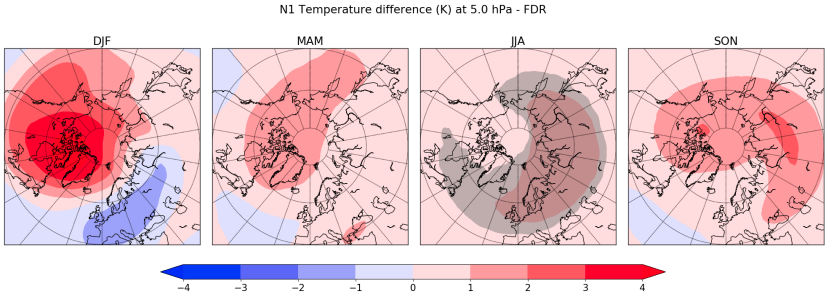

Figure 4. ~~As Figure 2, but~~ Northern hemisphere seasonal differences in stratospheric temperature (High Ap - Low Ap) at 5 hPa after applying the FDR correction. Gray areas indicates statistically significant areas at the 5% confidence level – before FDR correction.

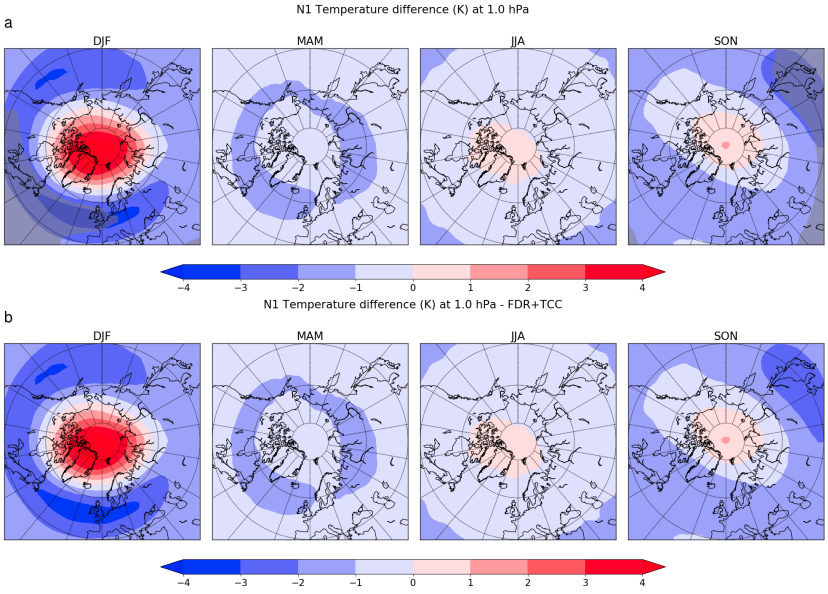

Figure 5. As Fig. 2 but for the 1 hPa level.

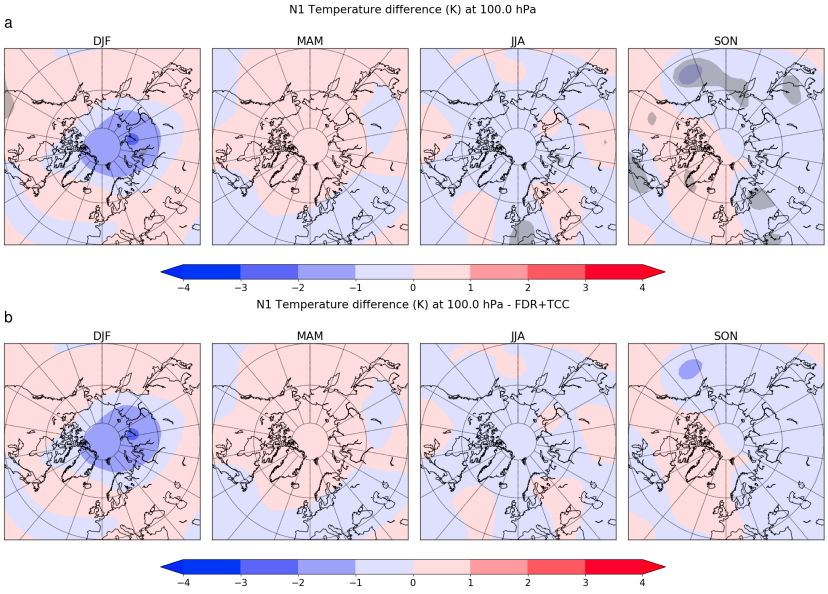

Figure 6. As Fig. 2 but for the 100 hPa level.

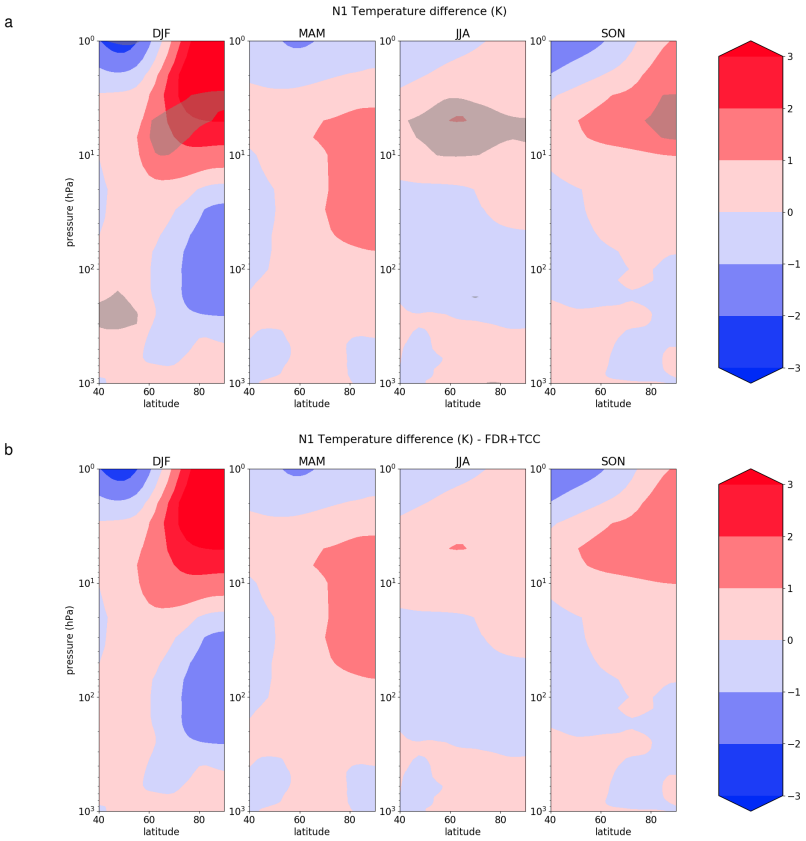

Figure 7. Zonal mean temperature differences (High Ap – Low Ap) without (a) and with (b) temporal and spatial autocorrelation corrections. The gray areas indicate statistically significant temperature differences at the 5% confidence level.

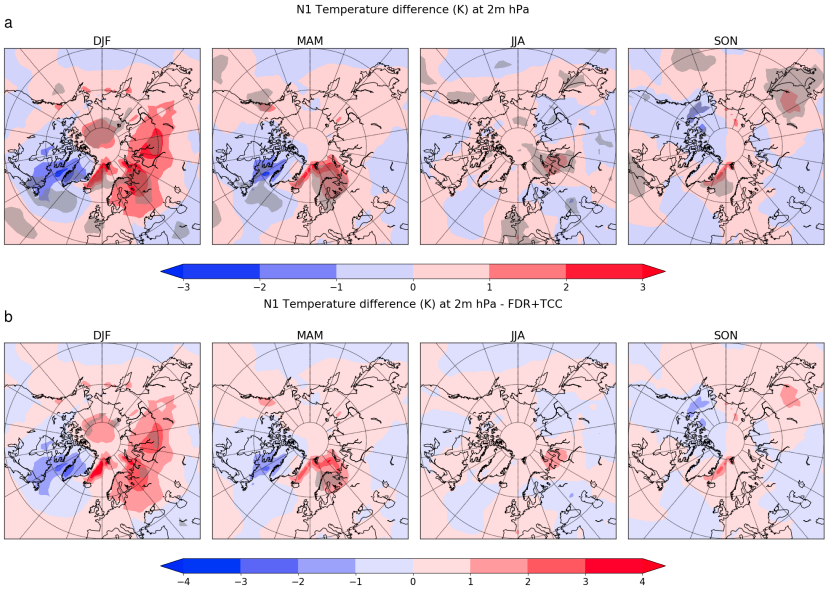

Figure 8. 2m temperature differences (High Ap – Low Ap) without (a) and with (b) temporal and spatial autocorrelation corrections. The gray areas indicate statistically significant temperature differences at the 5% level.