Answer to referee's comment :

Responses to Anonymous Referee #1

The authors thank the anonymous referee for the detailed review of the manuscript, for the meticulous pointing out of inconsistencies between tables and figures, as well as for all their comments and suggestions allowing a clear improvement of the paper.

Responses to specific comments:

This manuscript is well written and is an important contribution for more comprehensive trend analysis of atmospheric composition data. The work is robust with very good analysis and discussions of the different effects on the trend results using various prewhitening methods in addition to MK without prewhitening. It is very well appreciated the clear guidelines for choosing methods and approaches for assessing long term trends.
I will recommend the paper to be published as it is. I have only some small comments/questions which you may consider:

- Line 125. why is negative autocorrelation rare in atmospheric processes? Maybe explain a bit more the reasons and differences between negative and positive autocorrelation and/or give a reference.

  A negative autocorrelation changes the direction of the influence. In atmospheric processes, persistence is responsible for autocorrelation rather than "reaction" or "rebound" mechanisms. Persistence embodies the fact that atmospheric variables tend to change relatively slowly, and when changes occur, the autocorrelation tends to decrease toward zero rather than reach negative values. The latter would be the sign of some kind of rebound mechanisms where atmospheric parameters having particular values, for instance above average, would result in latter values being more likely below average. We cannot think of such examples in atmospheric processes, except for processes strongly correlated with natural cycles such as the circadian cycle. For instance, the difference of solar irradiance to the daily solar irradiance average would obviously exhibit a negative autocorrelation at 12h time lag, but it is just due to the high correlation of the solar irradiance with the solar zenith angle. Negative autocorrelation is a violation of independence but it is generally less worrisome because it appears less frequently than positive autocorrelation and it produces greater precision in the average than an independent series would.

- Line 272. Why is aerosol number concentration behaving different than the other components regarding the effect of granularity, i.e. the ss remain until the one-year aggregation?

The number concentration exhibits a less pronounced seasonal cycle than the other parameters, because its seasonal cycle has variable response to the temperature. For example, at JFJ during summer, higher temperatures lead to a larger influence of the planetary boundary layer and higher production/transport of primary aerosol. During winter, the colder temperatures can also lead to increase formation of new particles (secondary particles). The 3-month averaging corresponds approximately to a season, so that the small seasonal cycle is not able to mask the positive autocorrelation.

- Fig. 8 and paragraph 429-436. Here you compare the difference in granularity of monthly and seasonal data. Why use different data (scattering contra absorption)? To illustrate the difference in granularity it would have been more logic to use same dataset?

  The comparison between months and meteorological seasons would have been easiest with the same dataset. The authors however chose two different variables to show that the effect of the time granularities on the variability of the slope and the size of the confidence limits is similar for two different variables. This was an option and we try to give examples from all the time series along the paper to enhance that the results do not only concern a peculiar case of atmospheric parameter. The opposite choice was made for Fig. 10.

- Fig10 and paragraph 493-509. Not sure if I understand how the data selection has been done. Do all the periods contain the whole time series? I.e 10 years contain 3x10years data set if the time series is totally 30 years. I assume you somehow taken into account that the actual trend for the whole period will effect the results. Not homogeneous trend over a 30 year period. But why is it then so few data points for the 4 year trend, I,e N=360 and 120 for monthly and seasonal trends?

  For Fig. 10, only the period ending in 2018 with different lengths (4 years to 30 years) is presented, so that the 10 years correspond to the trend between 2009 and 2018 and contains only one 10 y data. If all potential x years trends were used, the mean of the numerous 4 y trends would potentially mask the increase of the absolute values of the slope and the larger difference between individual time segmentations for shorter period length.
  Since only one period of 4 years is used, the number of data in the time series is N=360 (=4 years*3 months*30 days) for a time segmentation into four meteorological seasons and whereas monthly trends for the same time series are computed with N=120 (=4 years*1 month*30 days) for monthly trends.
  The figure caption was modified in order to clarify the data selection: "*Figure 10: VCTFPW slopes and CL as a function of various period lengths ending in 2018 for the daily aerosol absorption coefficient for the division of the time series into a) 12 months and b) four meteorological seasons. Colors represent time period lengths and bigger symbols represent ss trends.*"

- The new algorithm applied. Is that made available? The scheme sketched in Figure 1 is not very easy to use for others to apply the method. It is recommended that the authors upload the scripts for others to use and adopt if possible.

  The new algorithm in Matlab, Python and R will be published in github and the doi will be given in the revised version of the manuscript. We have to finish to documentation of the code before releasing the doi in the next days. . This will happen soon (in conjunction with paper publication). The following section on code availability was added to the manuscript:" *We provide, in dedicated Github repositories hosted within the "mannkendall" organization (https://github.com/mannkendall), a Matlab (DOI: ; https://github.com/mannkendall/Matlab), Python (DOI: ; https://github.com/mannkendall/Python), and R (DOI: ; https://github.com/mannkendall/R) implementation of the algorithm presented in Sec. XX. In particular, these open-source codes, distributed under the BSD 3-Clause License, allow to compute the MK test and the Sen's slope with various prewhitening methods (3PW (default), PW, TFPW-Y, TFPW-WS and VCTFPW). The time granularity, period and temporal segmentation are chosen by the users during the preparation of the datasets. The level of the confidence limits for the MK test, the lag-1 autocorrelation, and the homogeneity between the temporal aggregation can also be defined by the user. The probability for the statistical significance, the statistical significance at the desired confidence level, the Sen's slope and its confidence limits are returned as results. A set a common tests is used to ensure that both the Python and R implementations are consistent with the (original) Matlab implementation of the code."*

Responses to Wenpeng Wang

The authors thank Wenpeng Wang for his detailed review of the manuscript, and for all the comments and suggestions allowing a clear improvement of the paper. The line numbers correspond to the manuscript submitted to AMTD.

Responses to specific comments:

General comments
The authors propose a new algorithm of trend analysis on autocorrelated meteorological data via incorporating the merit of three prewhitening techniques. The effect of time granularity, time segmentation and time series length on trend analysis are also evaluated on the basis of real meteorological observations.
The proposed algorithm is a good trial to purse the ideal goal of trend detection methods, that is high power with controllable Type I error, and accurate slope estimates. I think this algorithm is practically sound.

But still I have reservations about some statements in this paper. The manuscript and the quality of figures should be improved before it is formally published.

Specific comments:

1. Line 120-122. "*These approaches (variance correction approaches) appear not able to preserve the significance level and the power of the MK-test in the case of correlated time series with a trend*"

Comment: Both the variance correction approach and the prewhitening approach can preserve the pre-assigned significance level when there is no trend. Because detecting trends with known statistical confidence is the primary goal of trend analysis, either on independent data or autocorrelated data. The power of trend identification may be different for distinct methods.

> Your comment is completely right. The sentence was consequently improved: "*These approaches appear to preserve the pre-assigned significance level and the power of the MK-test in the absence of trend but not in the case of correlated time series and in the presence of a trend (Yue et al., 2002; Blain, 2013).*"

2. Line 139-140: *it (PW method) reduces the power of the test due to an over-/underestimation of ak1data in the case of a positive/negative trend.*

Comment: The existence of real trend, either positive or negative, can lead to an overestimation of lag-1 autocorrelation coefficient.

> The sentence was modified for clearity: "*This PW method results in a low amount of type 1 errors, but the existence of real trends, either positive or negative, can lead to an over-/underestimation of* ak_1^{data}*, which will reduces the power of the test.*"

3. Comment on Line 151-164: The brief introduction on the TFPW-WS method (Wang and Swail's 2001) includes some mistakes. I suggest rephrase this paragraph.

The original idea of Wang and Swail's (2001) was intended to implement the MK test on the prewhitened series, rather than on the prewhitened detrended series, as it was given by Eq.(8). If the prewhitened series are detrended, then we will never identify any trends. The critical value to stop iteration should be a tiny number, e.g. 0.0001, instead of 0.05. The primary consideration of iteration procedure was to mitigate the adverse effect of trend on the accuracy of lag-1 autocorrelation coefficient estimate.

> The authors do agree with the proposition of the reviewer and the manuscript was modified consequently:" *The original idea of Wang and Swail's (2001) was intended to implement the MK test on the prewhitened series, rather than on the prewhitened detrended series, as it was given by Eq.(8). If the prewhitened series are detrended, then trends will not be identified. Wang and Swail's (2001) propose an iterative TFPW method to mitigate the adverse effect of trend on the accuracy of the lag-1 autocorrelation estimate. This iterative procedure consists of: i) removing* ak_1^{data} *from the original time series and correcting the prewhitened data for the modified mean (eq. 5); ii) estimating the Sen's slope* β^{prew} *on the*

prewhitened data $A_{cor,t}^{prew}$; iii) removing the trend (β^{prew}) estimated on the PW data from the original data to obtain a prewhitened detrended time series $A_{cor,t}^{detr}$(eq. 6); and iv) applying iteratively i-iii until the ak_1 and slope differences become smaller than a proposed tiny threshold of 0.0001 (eq. 7)."

All the TFPW-WS trends have been recomputed with the threshold of 0.0001 without any marked differences in the results. The following sentence was then added at line 160: "*Note that the use of a higher threshold up to 0.05 does not significantly modify the results obtained on the considered time series*."

4. Line 164-167: *The PW-cor method refers to the preliminary step of the first iteration in the TFPW-WS method and consequently corrects the prewhitened data by the same factor. To the knowledge of the authors, this PW-cor method is not referenced in the literature but is a potential method tested in this study.*
Comment: After rephrasing the TFPW-WS method, please describe the PW-cor method more clearly.

The PW-cor is now described with more details: "*The preliminary step of the first iteration in the TFPW-WS method (removing ak_1^{data} from the original time series and correcting the prewhitened data for the modified mean eq. (5)) corresponds to the standard PW method but with the same correction factor ensuring a similar trend between the prewhitened and the original time series. This method called PW-cor is, to the knowledge of the authors, not referenced in the literature but is a potential method tested in this study."*

5. Line 184-185: *VCTFPW preserves to some extent the power of the test, but only mitigates the type 1 errors.*
Comment: Similar to other prewhitening methods, the VCTFPW method mitigates the inflationary type 1 errors raised by autocorrelation as its priority. Then the method preserves the power of the trend test to some extent.

The authors consider that the first "priority" of the VCTFPW method is to preserve the value of the slope and that this priority has also some effects on the type 1 and type 2 errors. The manuscript was modified: "*Statistical simulations by Wang (2015) showed that this new variance corrected prewhitening method (VCTFPW) leads to more accurate slope estimators, tends to mitigate the inflationary type 1 errors raised by autocorrelation and preserves to some extent the power of the test."*

6. Line 203-204: *If PW is ss but TFPW-Y is not, then the trend is considered as a false negative due to the lower test power of PW and the trend has to be considered as ss.*
Comment: If we consider the trend to be statistically significant, then we cannot say the detected trend is a false negative result. It is illogical to report a trend and meanwhile state this is an error.
Figure 1 should be revised accordingly.

The referee is right. If PW is ss but TFPW-Y is not, this is not a TFPW-Y false negative but a PW false positive and this has to be considered as not ss. Figure 1 was changed accordingly and the manuscript is revised:

- Lines 203-204: "*If TFPW-Y is ss but not PW, the trend is considered as a TFPW-Y false positive due to the too high type 1 errors of TFPW-Y and the trend has to be considered as not ss. If PW is ss but TFPW-Y is not, then the trend is considered as a PW false positive and the trend has to be considered as not ss.*"
- § 4.1: lines 322-342: "*To obtain a better view of the weakness of each MK-test, the percentage of false positives taking each of the prewhitening method as reference are reported in Table 3 for all the datasets. PW-cor has by definition the same ss as PW, so that their performances are given in the same column. PW has to be used as the best reference for false positives because it is the prewhitening method with the lowest type 1 error (Zhang and Zwiers, 2004, Yue et al., 2002, Blain, 2013, Wang et al. 2015[a]), whereas the consideration of the other prewhitening methods as reference allows for the evaluation of the discrepancy in ss among the methods. For the decadal trends, MK, TFPW-Y and VCTFPW have 32-47% of false positives taking PW as reference. This suggests that about two thirds and half of the trends determined using TFPW-Y and VCTFPW, respectively, are false positives. TFPW-WS has less than 2% of false positives, so that it can be considered to have equivalent performance as PW. For the trends on short periods, the lower amounts of false positive for MK and TFPW-Y are due to the overestimation of the slopes with these tests (see section 4.4) leading to trends that are more robust and enhanced ss. The unbiased estimate of the VCTFPW slope produces similar amounts of errors for the short-term trends as for the decadal trends. The percentage of false positives is similar if TFPW-WS is considered as the reference. If MK or TFPW-Y is taken as reference, PW and TFPW-WS have a very low number of false positive independent of the length of the period, leading to the conclusion that few cases remain uncertain. Note that 5-10% of cases have different ss at the 95% confidence level if MK or TFPW-Y is used, indicating that estimation of the ss using these two methods can have a slight impact on the results. Finally, all the prewhitening methods have a higher number of false positive if VCTFPW is considered as the reference because the added slope at the end of the VCTFPW procedure is smaller than the initial slope and leads to less detectable trends. Note also that the percentage of false positives of PW and TFPW-WS remains low (≤ 4%). For the time series considered in this study, the following conclusions can be made: 1) PW (and PW-cor) performs very well with a small (≤ 3.5%) number of false positives if other prewhitening methods are considered as reference; 2) TFPW-WS has a very low number of false positives (less than 2% if PW is taken as the reference); 3) VCTFPW exhibits high type 1 errors and should consequently not be used to determine the ss; and 4) The difference in ss between MK and TFPW-Y is related to only 5-10% of the trends."*
- Table 3:

Period	MK	TFPW-Y	TFPW-WS	PW/PW-cor	VCTFPW
≥ 10y N=2219	**32.5**	**37.1**	**1.7**	**reference**	**47.0**
	31.8	36.1	reference	0.7	46.4
	reference	*9.4*	*0.2*	*0.3*	*26.4*
	5.0	*reference*	*0.2*	*0.2*	*24.8*
	15.7	*18.4*	*4.0*	*3.5*	*reference*
< 8y N=1067	**16.0**	**14.1**	**0.7**	**reference**	**36.6**
	15.9	13.9	reference	0.5	36.7
	reference	*3.0*	*0.1*	*0.0*	*28.1*
	5.0	*reference*	*0*	*0.0*	*29.7*
	8.4	*8.1*	*1.3*	*1.1*	*reference*

7. Line 225-228: *Trend analyses were applied on several periods. For all the data sets, **a 10-year period** is considered first and then further possible **multi-decadal periods up to 60 y** for the radio-sounding time series. For the in-situ aerosol properties, **tests with 4 to 9 y periods** are also computed in order to illustrate the problems of trend analysis on very short time series.*
Line 781-782: *lag-1 autocorrelation of the observations (ak1data) and number of ss partial autocorrelations **for the 10y period (order)**, number of data in the **10y period** (N) and reference.*
Comment: I think in this section "3 Experimental", the authors should clarify how to analyze the measure data, in order to support the coming results. The meaning of "*a 10-year period*"or "*multi-decadal periods up to 60 y*" are unclear and obscure.
In table 2, the meaning of "*number of ss partial autocorrelations **for the 10y period (order)**, number of data in the **10y period** (N) and reference"* is unclear either.

- The expression "10 y period" corresponds to an analysis over 10 years of measurements. For most cases the more recent 10 y period is considered and corresponds to 2009-2018 for all parameters. The exception is the AOD where the 10 y period corresponds to 2006-2015 (no more recent AOD data were available). Sometimes all potential 10 y periods are considered, namely 2009-2018, 2008-2017, 2007-2016, etc.
- The expression " multi-decadal periods" correspond to periods of several decades, e.g. 20 y (1999-2018) or 30 y (1989-2018) and up to 60 y (1959-2018) for the tropopause and zero degree levels.
- Section 3 (Experimental) of the manuscript was modified to clarify this point: "*Trend analyses were applied on several periods. For all the data sets, the last 10-year period (e.g. 2009-2018 for the BND aerosol scattering coefficient) is considered first and then further possible multi-decadal periods (e.g. the last 20 y (1999-2018), 30 y (1989-2018)) up to 60 y for the radio-sounding time series."*

8. Line 275-276: *CL of MK, PW and TFPW-Y, which remove the lag-1 autocorrelation without compensation for the **mean values and the variances**…*

Comment: Does that mean “mean and variances of the slope estimate”?

> No, it means the mean value and the variance of the original time series. This is now explicitly written in the manuscript: “*CL of MK, PW and TFPW-Y, which remove the lag-1 autocorrelation without compensation for the mean values and the variances of the original time series, are smaller than for VCTFPW, PW-cor and TFPW-WS. PW-cor and TFPW-WS have the highest CL.”*

9. Line 278-280: *The **ss** often decreases for coarser time granularities occasionally leading to not **ss trends** for some of the prewhitening methods. PW, TFPW-WS and VCTFPW methods become **not ss** at finer time granularities than TFPW-Y and MK due to their lower number of false positives.*
Comment: It’s hard to identify the relationship between the significance of trend and the time granularity from Fig. 2.

> The authors agree that this relation is not that obvious from Fig. 2, where it can only be detected for some variables (e.g. TFPW-WS scattering coefficient at 1 and 3 months time granularity or PW, PW-cor, TFPW-WS and VCTFPW tropopause level at 1 month time granularity). This result is much more visible in Fig. 7, but it is an important result that the authors wish to already mention at this stage.

10. Line 281-282: *The discrepancies between prewhitening methods are larger than the discrepancies that occur when different temporal segmentations (months or meteorological seasons) are applied.*
Comment: Fig.2 does not support this finding.

- This statement is correct but not described with precision. The authors wanted to emphasize that the differences between the slopes computed from the various prewhitening methods are larger than between the different temporal segmentations for a defined prewhitening method. This is clearly visible in Fig. 2 a), b) and c) where, e.g., the slopes for all three temporal segmentations (different symbols) are very close but where the absolute values of PW and VCTFPW slopes are smaller than for the other prewhitening methods. The manuscript was modified to be more precise: “*The slope discrepancies between prewhitening methods are larger than the discrepancies that occur when different temporal segmentations (months or meteorological seasons) are applied for a defined prewhitening method.”*

11. Line 284-285: *the similarity of MK slopes with TFPW slopes*.
Line 350-352: *Due to the detrending procedure, the absolute values of the TFPW-Y slope are larger than the PW slopes and similar to the MK slope values (Fig. 2), even if a tendency to have larger TFPW-Y than MK slopes are observed.*
Line 367: *TFPW-Y slopes tend to be larger than MK slopes (Fig. 4b), with larger differences at high ak1 data leads.*

Comment: If my understanding is right, the MK and TFPW-Y should yield exactly the same slope of trend. The MK test does not estimate the slope of trend directly. It usually reports the magnitude of trend by the use of Sen's slope. The TFPW-Y also estimates Sen's slope as its first step. It will reinstall this trend to the prewhitened series without any modification. So these two slopes should be equal to each other.

> The TFPW-Y method reinstalls the Sen's slope (corresponding to the MK slope) to the detrended dataset after removal of the first-lag autocorrelation. The TFPW-Y slope is then estimated from the prewhitened time series (TFPW-Y data) and is not the same as the original slope. The Mk and TFPW-Y slopes are consequently somewhat different because they are computed from two different time series.

12. Line 285-287: *For example, the number of data points in the AOD time series (about 65 per year) induces higher CL for time granularities finer than the measurement frequency (about 10 days).*
Line 372-373: *Removing the lag-1 autocorrelation increases the variance, but decreases the mean.*
Line 391-395: *The spread of the slopes of the aerosol number concentration for the one-year aggregation on Fig. 2c shows that the yearly data still have a ss ak1 data for the longest periods of 20 and 24 years (see similar cases in Fig. 2). For shorter periods (5 to 9 years), the ak1 data decreases rapidly for averaging longer than 10 days and even becomes negative for yearly averages.*
Comment: These sentences are difficult to be understood. Please rephrase.

> The sentences were rephrased:
- Line 285-287:" *For example, the very low number of data points in the AOD time series (about 65 per year) corresponds to an average of one data per 5 days; there is consequently a very high amount of missing values for time granularities finer than this measurement frequency and this induces higher CL for time granularities of 1-3 days than granularity of 10 days."*
- Line 372-373: "*Removing the lag-1 autocorrelation leads to prewhitened data with a larger variance, but lower mean than the original time series."*
- Line 391-395: "*For the 10 y period represented on Fig. 6, none of the* ak_1^{data} *values are ss for a one-year time granularity. However, there are cases like the 24 y time series of the aerosol number concentration where* ak_1^{data} *is still ss for the one-year time granularity. In these cases, prewhitening methods have to be applied, which leads to the spread of the slopes for the various prewhitening methods visible on Fig. 2a."*

13. Line 294-296: *The yearly trend was computed for all periods (from 5y to 24y) at all considered time granularities (1 day to 1 month for the meteorological season temporal segmentation), leading to* ***40 trends****.*
Comment: Please clarify what is the 40 trends?

- The number 40 corresponds to trends computed for 8 different periods (5, 6, 7, 8, 9, 10, 20 and 24 years) and 5 time granularities (1, 2, 3, 10 and 30 days), so that 8*5= 40 trends.

14. Line 323-325: *PW is used as the reference for false positives because it is the prewhitening method with the lowest type 1 error, while TFPW-Y is the reference for false negatives because it is the most powerful test.*
Comment: It's inappropriate to state that the TFPW-Y is the most powerful test. The TFPW-Y tends to report significant trends at the expense of committing high type 1 error. This finding has been verified by many literatures. So we can say the TFPW-Y tends to identify significant trends more frequently than other methods, but we cannot say it is the most powerful test.

- The power of the test is defined (see § 21. Line 102 of new manuscript) as the potential to detect ss trend and correspond to low type-2 error. With this definition applied throughout the manuscript, this sentence is right.

15. Line 338-342: *For the time series considered in this study, the following conclusions can be made: 1) PW performs very well with an almost vanishingly small (≤0.3%) number of false negatives and the ss of PW-cor is similar to that for PW;3) VCTFPW has a very high type 1 and 2 errors and should consequently not be used to determine the ss; and 4) it is not possible to determine whether MK or TFPW-Y is the most powerful method.*
Line 786-788: *Table 3*
Comment: The three conclusions made here do not align with the consensus about the prewhitening method among the community. I suggest to recheck the results.
1) The PW tends to overestimate the lag-one autocorrelation coefficient without trend removal, see Hamed (2009).In addition, the PW reduces a portion of real trend, see Yue and Wang (2002).That's the reason why Yue et al. (2002) suggest to remove trend before whitening. So if theTFPW-Y is the reference for false negatives, the PW is less likely to miss only 0.2% significant trends.
3) As it was stated by the authors, e.g. Line 265-266,Table 1,Figure 4(a).The VCTFPW slopes lies between the TFPW and the PW slope values. So no matter one takes the PW or the TFPW-Y as the reference, the VCTFPW is less likely to commit the highest error among all the prewhitening methods.
4) For the autocorrelated data, the MK and TFPW-Y are not really powerful method. They only tend to report significant trends more frequently than other PW methods. However, both of them commit high type I error as a price.
I have to say, the above opinions are given by Monte-Carlo simulation results. They may not suitable to every real-world series. This study deals with measured data. So I suggest to recheck your results again.

The authors checked the scripts and recomputed all the results. The results presented in the submitted manuscript are correct and do not contradict the cited references. Here some further comments on the numbered remarks:

- Point 1): As stated in the answer to comment 6, it is not possible to detect false negatives without simulated time series with trends. As defined now in Fig. 1, what

was called “false negative” are in fact PW false positive if TFPW-Y is taken as reference. Fig. 1, Table 3 and the related descriptions were modified accordingly.

- Point 3) It is right that the VCTFPW lies between the TFPW-Y and the PW slope values and this is a sign that VCTFPW can be accepted as the best slope estimate. But slope estimate has nothing to do with the determination of the statistical significance, since the MK test is constructed to detect the ss but the slope estimate is performed via the Sen’s slope. The potential to commit error does not rely on the value of the slope.
- Point 4): the referee is right and the results of this study do completely agree with this statement. The discussion on the power of the method was discarded since the amount of false negative cannot be estimated with real time series.

16. Line 345: *The slope of the trend is always enhanced by the positive ak1data*.
Comment: I think it should be “the slope estimates of the trend is influenced by the positive lag-one autocorrelation”. The autocorrelation increases the difficulty of an accurate slope estimation. But it does not increase or decrease the real slope of the trend.

- The referee is right, this sentence is problematic. The slope of the trend is not modified by the autocorrelation in the time series, but it is the slope estimate performed on the original dataset that is influenced. However, it remains correct that the slope estimate performed on the original dataset is enhanced by positive lag-one autocorrelation. The manuscript was modified: “*The slope estimated on the original data is always enhanced by the positive* ak_1^{data}”

17. Line 628-629: *Consistent with the literature, the use of MK, TFPW-Y and VCTFPW results in a large amount of false positive results while TFPW-WS results in less than 2% of false positives.*
Comment: After recheck your results, e.g. table 3, this conclusion should be revised accordingly.

- The results were checked and this statement is still right. The labels “false negative” was however incorrect since false negative cannot be determined on real measurement (see answer to comment 6), the real value of the slope staying unknown. The number of false positives depends on the prewhitening method chosen as reference. Table 3 was consequently modified and now included the percentage of false positive with each prewhitening method taken as reference. Since PW is commonly accepted to be the method with the least amount of false positive, it is now given in bold, whereas the prewhitening methods known to have a much higher amount of type-1 errors are displayed in italic.

18. Line 637: *The confidence limits are much broader for coarser time granularities and the ss is lower.*
Comment: Fig. 8 supports this conclusion but Fig. 10 does not. As the time granularity becomes coarser, the confidence limits are much narrower in Fig. 10.

- These conclusions are supported by Figures 2, 7, 8 and 9 where one of the variable is the time granularity. Figure 10 presents the slope, the confidence limits and the ss as a function of the length of the period considered to compute the trend but the time granularity was not considered. These results were computed for a common time granularity of one day.

19. Comment on Figure 2: it is hard to distinguish the time segmentation.

The authors do agree that the density of information on Fig. 2 requires better clarity about the main results suggested by this figure. The slopes computed from the two temporal segmentations (12 months and 4 meteorological seasons) were removed and, instead, boxplots were inserted allow estimation of the discrepancy between the temporal segmentation into four meteorological seasons (considered as the best use for all the used time series) and the 12 months temporal segmentation or no use of segmentation.

20. Comment on Figure 7:it is not easy to identify different PW methods.

The authors changed the symbols and increased their size. Ss trends are consequently no longer given by bigger symbols, but instead are indicated by the red and black lines describing the 95% and 90% confidence level.

21. Comment on Figure 8 and 10: it is unclear how to analyze the slope of trend as well as the confidence limit within each time segmentation. It should be well explained.

In the revised manuscript, the figure caption of Figs 8 and 10 specifies that the slopes correspond to dots and the CL to vertical lines. I hope that I have well understood the referee's requirement.

22. I suggest to improve the quality of the figures, to make them self-explaining.

First the figure captions were all revised in order to increase the clarity and to homogenize the descriptions. Second used symbols are now all described in legends on the figures. All the figures were also verified and modified:
- Fig. 1: To correspond closely to the published code in github, the statistical significance was symbolized with $P_{\text{prewhitening method}}$ instead of S and the choice of P_{3PW} as equal to the min of the $P_{TFPW\text{-}Y}$ and P_{PW} (or p-value(3PW)=max(p-value(TFPW-Y), p-value(PW))) is explicitly given.
- Fig. 2: The size of the symbols for the ss are now specified in the legend. The results for the temporal segmentations of 12 months and 4 meteorological seasons are no more displayed but are replaced by boxplots allowing the comparison with the displayed results without temporal aggregation.
- Fig. 3: the used granularities and periods are now specified in the figure caption and the color scale of Fig. 3b is labelled. The titles precise that all periods, all granularities and meteorological seasons were used for a) panel and that all time

series, decadal period, granularities and time segmentations were used for b) panel.
- Fig. 4 : the symbols were added in the legend of panel a.
- Fig. 5: the used granularities and periods are specified in the figure caption. The term “all trends” was replaced by “all periods, all granularities” in the figure title.
- Fig. 6: A title was added specifying that 10 y period of all times series were used for this figure. The sizes of the symbols for the ss are now specified in the legend.
- Fig. 7 the symbols were modified to allow the distinction between the prewhitening methods and the figure caption specifies that no temporal segmentation was used. Ss trends are no longer displayed with bigger symbols, but the ss at 90s and 95% confidence levels is given by the black and red lines.
- Fig. 8: the use of 10 y period is specified in the title. The figure caption now attributes the slope to dots and the CL to vertical lines. The y-axe label also mentions the confidence limits. The size of the symbols for the ss is now specified in the legend. Vertical lines are also added to separate the results for temporal segment. A title is added above the legend to specify that the colors correspond to various granularities.
- Fig. 9: Legends describing the time segmentations and periods are added to the figure.
- Fig. 10: The figure caption now attributes the slope to dots and the CL to vertical lines. The y-axe label also mentions the confidence limits. The sized of the symbols for the ss are now specified in the legend. Vertical lines are also added to separate the results for temporal segment. A title is added above the legend to specify that the colors correspond to various periods. The title of the figure specifies that a granularity of one day was used.
- Fig. 11: it is now specified in the figure caption that the slope were normalized by the median of the data. The color scales have now clear legends. The titles were modified to mention that all the time series, granularities and temporal segmentations as well as periods of at least 10 y were used for both panels a and b.

Technical corrections

Line 109. Zwang and Zwiers(2004)does not given in the reference list.

Line 168. I think the correct citation about the VCTFPW method should be “Wang, W., et al., 2015. Variance correction pre-whitening method for trend detection in auto-correlated data. Journal of Hydrologic Engineering, 04015033. doi:10.1061/(ASCE)HE.1943-5584.0001234.”

Line 272: “*aerosol absorption coefficient*” should be “aerosol scattering coefficient”.

Thanks for this very detailed review, the technical corrections were applied.

New manuscript with track-changes:

Effects of the prewhitening method, the time granularity, and the time segmentation on the Mann-Kendall trend detection and the associated Sen's slope

Martine Collaud Coen[1], Elisabeth Andrews[2,3], Alessandro Bigi[4], Giovanni Martucci[1], Gonzague Romanens[1], ~~Giovanni Martucci[1]~~Frédéric P.A. Vogt[1] and Laurent Vuilleumier[1]

[1] Federal Office of Meteorology and Climatology, MeteoSwiss, Payerne, Switzerland
[2] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA
[3] NOAA/Earth Systems Research Laboratory Boulder, CO, USA
[4] Università di Modena e Reggio Emilia, Department of Engineering "Enzo Ferrari", Modena, Italy

Correspondence: Martine Collaud Coen (martine.collaudcoen@meteoswiss.ch)

Abstract

~~The most~~The Mann-Kendall test associated with the Sen's slope is a very widely used non-parametric method for trend analysis ~~is the Mann-Kendall test associated with the Sen's slope. The Mann-Kendall test~~. It requires serially uncorrelated time series, ~~whereas~~yet most of the atmospheric processes exhibit positive autocorrelation. Several prewhitening methods have therefore been designed to overcome the presence of lag-1 autocorrelation. These include a prewhitening, a detrending and/or a correction ~~for~~of the detrended slope and the original variance of the time series. The choice of which prewhitening method and temporal segmentation to apply has consequences for the statistical significance, the value of the slope and of the confidence limits. Here, the effects of various prewhitening methods are analyzed for seven time series comprising in-situ aerosol measurements (scattering coefficient, absorption coefficient, number concentration and aerosol optical depth), Raman Lidar water vapor mixing ratio, and ~~the~~ tropopause ~~and~~ zero degree temperature levels measured by radio-sounding. These time series are characterized by a broad variety of distributions, ranges and lag-1 autocorrelation values and vary in length between 10 and 60 years. A common way to work around the autocorrelation problem is to decrease it by averaging the data over longer time intervals than in the original time series. Thus, the second focus of this study ~~is evaluation of~~evaluates the effect of time granularity on long-term trend analysis. Finally, a new algorithm involving three prewhitening methods is proposed in order to maximize the power of the test, to minimize the amount of erroneous detected trends in the absence of a real trend and to ensure the best slope estimate for the considered length of the time series.

Keywords: Seasonal Mann-Kendall test, Theil-Sen's slope, prewhiten, detrend, autocorrelation

1. Introduction

To estimate climate changes and to validate climatic models, long-term time series associated with statistically adapted trend analysis tools are necessary. The basic requirements needed to apply specific statistical tools are usually well described, but end-users often do not systematically test if the properties of their time series fulfill these requirements. An inappropriate usage of the statistical tools may lead to misleading conclusions. It may also happen that a time series does not meet the complete criteria of any of the statistical tools. In that case, the statistical tool ~~can~~must be adapted or ~~the use of~~ different methods with complementary strengths and weaknesses ~~should~~must be ~~applied~~used.

The time series properties that can cause misuse of statistical tools for trend analysis primarily concern the statistical distribution, the autocorrelation, missing data or periods without measurements, the presence of seasonality, irregular sampling, the presence of negatives and the rules applied in the case of data below-detection limits. A large number of trend analysis tools such as the whole family of least mean square and generalized least squares methods are parametric methods and, consequently, require normally distributed residues. Unfortunately, many atmospheric measurements, which strongly depart from the normal distribution, do not meet this requirement so that non-parametric methods have to be used. Non-parametric techniques are commonly based on rank and assume continuous monotonic increasing or decreasing trends. The Mann-Kendall (MK) test associated with the Sen's slope is the most widely applied non-parametric trend analysis method in atmospheric and hydrologic research (Gilbert, 1987; Sirois, 1998). While it has no requirement on data distribution, it must be applied on serially independent and identically distributed variables. The second condition of homogeneity of distribution is not met if a seasonality is present, but it can be solved by using the seasonal Mann-Kendall test developed by Hirsch et al. (1982). The first condition of independence is not met if the data are autocorrelated, which is often the case ~~where~~for atmospheric variables that are controlled by autocorrelated physical or chemical processes. To correctly analyze ~~properly~~ autocorrelated and not normally distributed errors, two different strategies are usually applied ~~as described below~~.

The first strategy tends to decrease the amount of autocorrelation by aggregating time series into monthly, seasonally, yearly ~~data~~bins or even in longer periods. However, coarser time granularities (e.g., due to longer averaging periods) do not ensure that autocorrelation is removed. Moreover, the aggregation implies a decrease of the information density in the time series, such as the diurnal or seasonal cycles, the variance of the data and to some extent the data distribution. The aggregation conditions (length of the time unit, making the time unit consistent with the observed seasonality, starting phase of the time series and the averaging method) may influence the trend results (de Jong and de Bruin, 2012; Maurya, 2013) in what is called the Modifiable Temporal Unit Problem (MTUP).

The second strategy focuses on the development of algorithms to reduce the impact of the autocorrelation artifacts on the statistical significance of the MK test and on the Sen's slope. Two kinds of algorithms are usually used: (i) the prewhitening of the data to remove the autocorrelation and (ii) inflation of the variance of the trend test statistic to take into account the number of independent measurements instead of the number of data points (the autocorrelation reduces the number of degrees of freedom in tests).

In this study, the effects of various prewhitening methods on the MK statistical significance and on the slope are analyzed for time series of in situ aerosol properties, aerosol optical depth, temperature levels (tropopause and zero degree levels) and remote sensing water vapor mixing ratio. This study also analyses the effect of the time granularity on the MK statistical significance, on the strength of the slope and on

the confidence limits of various atmospheric compounds for the atmospheric time series listed above. Additionally, a new methodology combining three prewhitening methods and called 3PW is proposed in order to handle correctly the autocorrelation without decreasing the power of the test, while still computing the correct slope value.

2. The Mann-Kendall methodology (prewhitening methods)

The MK-test for trends is a non-parametric method based on rank. The calculated S statistic is normally distributed for a number of observation N>10 and the significance of the trends is tested by comparing the standardized test statistic $Z=S/[var(S)]^{0.5}$ with the standard normal variate at the desired significance level. For N≤10, an exact S distribution has to be applied (see e.g., Gilbert, 1987). Hirsch et al. (1982) extend the Mann-Kendall test to take seasonality in the data into account as well as the existence of multiple observations for each season. A global or ~~annual~~yearly trend can be considered only if the seasonal trends are homogeneous at the desired confidence level (Gilbert, 1987). Confidence limits (CL) are defined as the ~~100~~100x(1-p) percentiles of the standard normal distribution of all the pairwise slopes computed during the Sen's slope estimator, where p is the chosen confidence limit.

2.1 The problem of the autocorrelation in the MK-test

The MK-test determines the validity of the null hypothesis H_0 of the absence of a trend against the alternative hypothesis H_1 of the existence of a monotonic continuous trend. While no assumptions are needed about the data distribution (i.e., the definition of a non-parametric test), the MK-test does require that the data are serially independent, namely the absence of autocorrelation in the time series. Statistical tests are prone to two types of error. The first is an incorrect rejection of the null hypothesis H_0 ~~that is called~~ (a "type 1 error~~".~~"). This error is related to ~~a too~~an erroneously high statistical significance leading to false positive cases. The second is an incorrect acceptance of the null hypothesis H_0 ~~that is called~~(a "type 2 error~~".~~"). This error can be understood as the power of the test being too low leading to false negative cases.

The adverse effect of ~~the~~ positive autocorrelation in time series on the number of type 1 errors was suggested by Tiao et al. (1990) and Hamed and Rao (1998) and later simulated (Kulkarni & von Storch, 1995, Zwang and Zwiers, 2004, Blain, 2013, Wang et al., ~~2015~~2015[a], Wang et al., 2015[b], Hardison et al., 2019). All these studies clearly showed that positive autocorrelation in time series ~~largely~~significantly increases the number of type 1 errors, whereas prewhitening procedures ~~increased~~increase the number of type 2 errors. Larger lag-1 autocorrelation (ak_1) leads to higher percentage of type 1 errors and to a larger bias in the Sen's slope. Zwang and Zwiers (2004) also show that the occurrence of both types of error largely depends on the length of the time series, with longer periods leading to a strong reduction of errors and to a lower bias in the trend slope estimation.

A popular solution to get rid of the autocorrelation problem in the MK-test is to aggregate the time series in order to decrease ak_1. While the use of coarse time granularity effectively decreases the autocorrelation, the suppression of autocorrelation is not guaranteed, even in monthly or yearly

aggregations. Moreover, aggregation greatly decreases the number of observations N and can potentially affect the MK-test errors, the slope biases and the ~~CL~~CLs.

Two kinds of statistical procedures were developed to correct the MK-test for autocorrelation in the data. The variance correction approaches (Hamed and Rao, 1998; Yue and Wang, 2004; Hamed 2009; Blain, 2013) consider inflating the variance of the S statistic so that the number of independent observations instead of the total number of observations is taken into account. These approaches appear ~~not able~~ to preserve the pre-assigned significance level and the power of the MK-test in the absence of trend but not in the case of correlated time series ~~with~~and in the presence of a trend (Yue et al., 2002; Blain, 2013). The prewhitening approaches consider removing the lag-1 autoregressive (AR(1)) process in the time series prior to applying the MK-test. Several algorithms with various strengths and ~~defaults~~disadvantages have been published, and are described in the next section. Since negative autocorrelations are rare in atmospheric processes, only positive autocorrelations are taken into account in this study. Several studies have shown that the prewhitening methods are also applicable in case of negative serial correlations but with dissimilar consequences (Rivard and Vigneault, 200, Yue and Wang, 2002, Bayazit et al., 2004, Wang et al., 2015[b]).

2.2 The prewhitening methods

This section describes all the prewhitening methods known to the authors. The advantages and disadvantages of each method are summarized in Table 1. It has to be noted that, for all the methods proposed, the prewhitening can be applied only if ak_1 is statistically significant (ss) following a normal distribution at the two-sided 95% confidence interval. The first implemented prewhitening method (hereafter called PW) simply removes the lag-1 autocorrelation ak_1^{data} from the original data X at the time t:

$$X_t^{PW} = X_t - ak_1^{data} X_{t-1} \quad (1)$$

This PW method results in a low amount of type 1 errors, but ~~it reduces the power of the test due~~the existence of real trends, either positive or negative, can lead to an over-/underestimation of ak_1^{data} ~~in~~, which will reduces the ~~case~~power of ~~a positive/negative trend~~the test. A further procedure called trend-free prewhitening (TFPW) consists of removing the autocorrelation on detrended data. Yue et al. (2002) published the most commonly used method that consists of: i) estimating the Sen's slope β^{data} on the original data; ii) removing the trend to obtain a detrended time series A^{detr} (eq. 2); iii) removing the lag-1 autocorrelation ak_1^{detr} on A^{detr} to generate a detrended prewhitened time series $A^{\text{detr-prew}}$ (eq. 3); and iv) adding the trend back in to generate the processed time series to evaluate (i.e., X_t^{TFPW-Y}) (eq. 4):

$$A_t^{detr} = X_t - \beta^{data} t \quad (2)$$

$$A_t^{detr-prew} = A_t^{detr} - ak_1^{detr} A_{t-1}^{detr} \quad (3)$$

$$X_t^{TFPW-Y} = A_t^{detr-prew} + \beta^{data} t \quad (4)$$

This approach is called trend-free prewhitening (TFPW-Y) and restores the power of the test, albeit at the expense of an increase of type 1 errors. ~~Wang and Swail (2001) propose an iterative TFPW method that~~The original idea of Wang and Swail's (2001) was intended to implement the MK test on the prewhitened series, rather than on the prewhitened detrended series, as it was given by Eq.(8). If the prewhitened series are detrended, then trends will not be identified. Wang and Swail's (2001) propose an iterative TFPW method to mitigate the adverse effect of trend on the accuracy of the lag-1 autocorrelation estimate. This iterative procedure consists of: i) removing ak_1^{data} from the original time series and correcting the prewhitened data for the modified mean (eq. 5); ii) estimating the Sen's slope β^{prew} on the prewhitened data $A_{cor,t}^{prew}$; iii) removing the trend (β^{prew}) estimated on the PW data from the original data to obtain a prewhitened detrended time series $A_{cor,t}^{detr}$(eq. 6); and iv) applying iteratively i-iii until the ak_1 and slope differences become smaller than a proposed tiny threshold of 0.~~05~~0001 (eq. 7).

$$A_{cor,t}^{prew} = X_t^{PW-cor} = (X_t - ak_1^{data} X_{t-1})/(1 - ak_1^{data}) \quad (5)$$

$$A_{cor,t}^{detr} = (X_t - \beta^{prew} t) \quad (6)$$

$$A_{cor,t}^{detr-prew} = (A_{cor,t}^{detr} - ak_1^{detr-prew} A_{cor,t-1}^{detr})/(1 - ak_1^{detr-prew}) \quad (7)$$

$$X_t^{TFPW-WS} = A_{cor,t}^{detr-prew} \quad (8)$$

After n iterations until $ak_1^{detr-prew,n-1} - ak_1^{detr-prew,n} < 0.~~05~~0001$ *and* $\beta^{prew,n-1} - \beta^{prew,n} < 0.~~05~~0001$. Note that the use of a higher threshold up to 0.05 does not significantly modify the results obtained on the considered time series.

Wang and Swail's (2001) TFPW method (TFPW-WS) restores the low number of type 1 errors without decreasing the power of the test (Zhang and Zwiers, 2004). The factor $(1-ak_1^{detr-prew})^{-1}$ is needed to ensure that the prewhitened time series possesses the same trend as the original time series. ~~The PW-cor method refers to~~ The preliminary step of the first iteration in the TFPW-WS method ~~and consequently corrects~~(removing ak_1^{data} from the original time series and correcting the prewhitened data ~~by~~for the modified mean eq. (5)) corresponds to the standard PW method but with the same correction factor~~.~~ ensuring a similar trend between the prewhitened and the original time series. This method called PW-cor is, to the knowledge of the authors, ~~this PW-cor method is~~ not referenced in the literature but is a potential method tested in this study.

Finally, Wang et al. (~~2015~~2015[a]) proposed a further approach in order to correct TFPW-Y for both the elevated variance of slope estimators and for the decreased slope caused by the prewhitening. Practically, the variance of $A^{detr-prew}$ (i.e., σ_A^2) is restored to the variance of X (i.e., σ_X^2) to generate the $A_{VC}^{detr-prew}$ time series:

$$A_{VC,t}^{detr-prew} = A_t^{detr-prew} * \frac{\sigma_X^2}{\sigma_A^2} \quad (9)$$

The slope estimator β^{data} is decreased in the case of positive autocorrelation by the square root of the variance inflation factor (VIF) to obtain the corrected slope β_{VC}^{detr}(eq. 11). Matalas and Sankarasubramanian (2003) provided a simple way to compute the limiting values of VIF for a sufficiently large sample size and a first order autocorrelation:

$$VIF \approx (1 + ak_1^{detr})/(1 - ak_1^{detr}) \quad (10)$$

So that

$$\beta_{VC}^{detr} = \beta^{data} / \sqrt{(1 + ak_1^{detr})/(1 - ak_1^{detr})} \quad (11)$$

and

$$X_t^{VCTFPW} = A_{VC,t}^{detr-prew} + \beta_{VC}^{detr} t \quad (12)$$

Statistical simulations by Wang (~~2015~~2015[a]) showed that this new variance corrected prewhitening method (VCTFPW) leads to more accurate slope estimators, tends to mitigate the inflationary type 1 errors raised by autocorrelation and preserves to some extent the power of the test~~, but only mitigates the type 1 errors~~.

2.3 A new algorithm (3PW) involving three prewhitening methods

As described in sect. 2.2 and Table 1, each of the presented prewhitening methods has a ~~primary~~specific advantage: the low sensitivity to type 1 ~~error~~errors for PW, the high-test power for TFPW-Y, and the unbiased slope estimate for VCTFPW. TFPW-WS has both a low type 1 error and a high test power, but requires more computing time due to the iteration process. Here, we propose a new algorithm~~,~~ (3PW), described in Fig. 1, which combines the advantages of each prewhitening method:

- The ak_1^{data} of the original time series is calculated. If it is not ss, the MK test is applied on the original time series. If ak_1^{data} is ss, PW, TFPW-Y and VCTFPW are applied in order to obtain three prewhitened time series that are thereafter named after the specific prewhitening method for purposes of clarity.
- The MK-test that defines the statistical significance is applied on the PW and TFPW-Y data. If both tests are ss or not ss, the trend is considered as ss or not ss, respectively. If TFPW-Y is ss but not PW, the trend is considered as a TFPW-Y false positive (due to the ~~too high~~higher sensitivity to type 1 errors of TFPW-Y) and the trend has to be considered as not ss. If PW is ss but TFPW-Y is not, then the trend is considered as a PW false ~~negative due to the lower test power of PW~~positive and the trend has to be considered as ~~ss~~not ss. The probability P for the statistical significance is given by the higher probability between PW and TFPW-Y.
- The Sen's slope is then computed on the VCTFPW data in order to have an unbiased slope estimate.

3. Experimental

In order to have a broader view of the effects of the various PW methods, several very different time series (Table 2) were used: three surface in-situ aerosol properties (absorption coefficient, scattering coefficient and number concentration) measured at Bondville (BND), a remote, rural station in Illinois, USA; the aerosol optical depth (AOD) measured at Payerne (PAY) on the Swiss plateau; the tropopause

and the zero-degree temperature levels measured by radio-sounding launched at PAY; and the water vapor mixing ratio at 1015 m measured by remote sensing at PAY. The shortest time series (AOD and water vapor mixing ratio) cover only 10 years (y) of measurements while the longest time series cover 60 y. The three in-situ aerosol properties are Johnson-distributed and diverge strongly from a normal distribution. The other time series exhibit distributions that also diverge from a normal distribution but to a lower extent ~~so~~, such that some of them have residuals of a least mean square fit, which are normally distributed. The values of some of the time series span over several orders of magnitude and the scattering and absorption coefficients time series contains negative values due to detection limit issues in very clean conditions. The time series of the zero-degree temperature level ~~time series~~ also includes negative altitudes, since it is interpolated to altitudes lower than sea level in the case of negative ground temperature at PAY (S. Bader et al., 2019)). All the data have high ak_1^{data} at the daily time granularity and exhibit clear seasonal cycles with maxima in summer.

Trend analyses were applied on several periods. For all the data sets, ~~a~~the last 10-year period (e.g. 2009-2018 for the BND aerosol scattering coefficient) is considered first and then further possible multi-decadal periods (e.g. the last 20 y (1999-2018), 30 y (1989-2018)) up to 60 y for the radio-sounding time series. For the in-situ aerosol properties, tests with 4 to 9 y periods are also computed in order to illustrate the problems of trend analysis on very short time series. The number of data points in the time series (N) depends on the length of the period and on the time granularity. The choice of temporal segmentation to address seasonality for the seasonal MK-tests can also affect N and was evaluated by segmenting the time series into months and meteorological seasons for time granularities up to one month. The MK-test was also applied on the complete time series without considering seasonality (no temporal segmentation) for comparison purposes, even though, properly, seasonal MK-tests must be used when seasonal cycles are present.

To assess the statistical significance, the two-tailed p-values are computed. For a more comprehensive presentation of the results, the statistical significance is presented here as 1 minus p-value so that the ss at a 95% confidence level is effectively given by ss=0.95. If not further specified, the ss of the trend and of ak_1^{data} is given at the 95% confidence level, whereas CL and X_{homo} are given at the 90% confidence level. The slopes (in percent) are normalized by the median of the data. Periods of at least 10 y and trends on these periods are further called decadal periods and decadal trends.

4. Results and discussion

As explained in the methodology section (Sect. 2), the trend results (e.g., the ss, the slopes and the CL) depend on a number of factors, the most important ~~factors~~ones being the prewhitening method, the number of data points in the time series and the presence of autocorrelation. The choice of the prewhitening method clearly affects the ss, the slope and the CL. Analysis choices such as the time granularity, the length of the analyzed period and the temporal segmentation to address seasonality affect ak_1^{data}, N and the variance of the time series. There is a pronounced interdependency among these variables involving critical choices in the presentation of the results. Some general plots are first presented to provide insights into the primary results for some of the time series. They are followed by a more detailed analysis of the effects of the prewhitening method, the time granularity, the temporal segmentation, the length of the data series and the number of data points in the time series.

MK trend results (Fig. 2) of the aerosol number concentration, the aerosol ~~absorption~~scattering coefficient, the tropopause level and the AOD are plotted as a function of the time granularity for the MK-test and for all the prewhitening methods. The discrepancy between the results ~~are shown for~~computed with no temporal segmentation ~~(circle)~~ and for two different temporal segmentations to address seasonality (four meteorological seasons ~~(square)~~ and 12 months ~~(triangle)).~~) can be estimated from the inserted boxplots. The three aerosol properties exhibit decreasing trends while the results of the tropopause level time series indicate a positive trend. The negative aerosol slopes are related to the decreasing aerosol load in Western Europe and North America (Collaud Coen et al., 2020, Yoon et al., 2016). The increasing tropopause level trend is related to global warming (Xian and Homeyer, 2019). The results of the trends will not be further described and discussed, since this study is only focused on the methodology of the trend analysis.

The common features for all the time series considered here are:

- The MK, TFPW-Y, TFPW-WS and PW-cor methods result in similar slopes.
- As described in Wang et al. (~~2015~~2015[a]), the absolute value of the VCTFPW slopes lies between the TFPW and the PW slope values. The absolute value of the PW slopes is always smaller than the TFPW slope values.
- ~~The MK, TFPW-Y, TFPW-WS and PW-cor methods result in similar slopes.~~
- Large time aggregations usually lead to not ss ak_1^{data} and, consequently, prewhitening methods do not need to be applied~~.~~ to those cases. The ak_1^{data} of all prewhitening methods is not ss for three-month aggregations of the tropopause level and AOD datasets and for one-year aggregation of the aerosol ~~absorption~~scattering coefficient and AOD. The ak_1^{data} of the aerosol number concentration remains ss until the one-year aggregation.
- CL are smaller for finer time granularities in the presence of ss ak_1^{data}.
- CL of MK, PW and TFPW-Y, which remove the lag-1 autocorrelation without compensation for the mean values and the variances of the original time series, are smaller than for VCTFPW, PW-cor and TFPW-WS. PW-cor and TFPW-WS have the highest CL.
- The ss often decreases for coarser time granularities occasionally leading to not ss trends for some of the prewhitening methods. PW, TFPW-WS and VCTFPW methods become not ss at finer time granularities than TFPW-Y and MK due to their lower number of false positives.
- The slope discrepancies between prewhitening methods are larger than the discrepancies that occur when different temporal segmentations (months or meteorological seasons) are applied for a defined prewhitening method.

Apart from these ~~common results~~general observations, there are features that depend on the time series, such as the effects of the applied temporal segmentation to address seasonality, the similarity of MK slopes with TFPW slopes, and the time granularity leading to not ss ak_1^{data}. For example, the very low number of data points in the AOD time series (about 65 per year) corresponds to an average of one data per 5 days; there is consequently a very high amount of missing values for time granularities finer than this measurement frequency and this induces higher CL for time granularities ~~finer~~of 1-3 days than ~~the measurement frequency (about~~granularity of 10 days~~).~~.

4.1 Effects of the prewhitening methods

As predicted theoretically, the ss depends on the prewhitening method, with higher ss for the MK and TFPW-Y methods that are related to higher type 1 errors (false positives), while PW and VCTFPW have a lower ss and a lower test power. This is verified on the individual time series, e.g., for the aerosol number concentration results presented in Fig. 3a. The yearly trend was computed for all periods (from 5y to 24y) at all considered time granularities (1 day to 1 month for the meteorological season temporal segmentation), leading to 40 trends. The results show that:

- The MK-test ss without prewhitening has a median of 1, with the ss for the upper quartile and upper whisker also equal to 1 and thus within the 95% confidence level so that only 5 trends out of 40 evaluated (i.e., 12.5%) are not ss.
- The TFPW-Y ss has a median slightly lower than 1 and only 3 trends (7.5%) outside the 95% confidence level.
- The TFPW-WS ss has a median of 0.996 which is lower than MK and TFPW-Y. The lower quartile for TFPW-WS, is 0.89, which is outside the 95% confidence level and indicates that 32.5% of the trends are not ss.
- The results of both PW and PW-cor are similar to the TFPW-WS with median ss of 0.995, a lower quartile of 0.84 and 32.5% of the trends are not ss.
- The VCTFPW ss has the lowest median (0.98), first quartile (0.83) and lower whisker (0.63) leading to 37.5% of trends being not ss.

Similar results are found for all time series, but with less difference amongst the methods when the trends are obviously present or absent and more differences for weak trends.

According to Monte-Carlo simulations presented in the literature (e.g. Yue et al., 2002, Wang et al., ~~2015~~2015[a], Hardison et al., 2019), TFPW-Y leads to a high number of false positives. Since this study deals with measured data, the rate of false positives is defined as trends that are ss with TFPW-Y but not ss with PW, since the latter is the method with the lowest rate of type 1 error. Figure 3b shows that the number of false positives depends, as expected, on the strength of the slope and on ak_1^{data}. Weaker trends (smaller slopes in percent) are usually associated with lower ss and consequently lead to a larger number of false positives. The impact of the PW and TFPW-Y depends largely on ak_1^{data} absolute values, i.e., higher ak_1^{data} leads to stronger modification of the original time series with lower means (e.g., the mean of X_t^{PW} is less than the mean of X_t) and reduced variances for positive ak_1^{data}. The highest ak_1^{data} values (between 0.85 and 0.9) found in the time series studied lead to 60% to 100% false positives while ak_1^{data} values between 0.8 and 0.85 lead to at least 40% false positives.

To obtain a better view of the weakness of each MK-test, the percentage of false positives ~~and false negatives~~ taking each of the prewhitening method as reference are reported in Table 3 for all the datasets. ~~PW is~~ PW-cor has by definition the same ss as PW, so that their performances are given in the same column. PW has to be used as the best reference for false positives because it is the prewhitening method with the lowest sensitivity< to type 1 errors (Zhang and Zwiers, 2004, Yue et al., 2002, Blain, 2013, Wang et al. ~~error, while TFPW-Y is the reference for false negatives because it is the most powerful test.~~2015[a]), whereas the consideration of the other prewhitening methods as reference allows for the evaluation of the discrepancy in ss among the methods. For the decadal trends, MK, TFPW-Y and VCTFPW have ~~33-49~~32-47% of false positives~~.~~ taking PW as reference. This suggests that about two thirds and half of the trends determined using TFPW-Y and VCTFPW, respectively, are false positives. TFPW-WS has less than 2% of false positives ~~whereas PW-cor has similar false positives as PW. While PW,~~, so that it can be

considered to have equivalent performance as PW~~-cor and TFPW-WS have a low percentage of false negatives, false negatives make up ~5% of the trends for MK and up to one third for VCTFPW.~~. For the trends on short periods, the lower amounts of ~~type 1 and 2 errors~~false positive for MK and TFPW-Y are due to the overestimation of the slopes with these tests (see section 4.4) leading to trends that are more robust ~~trends~~ and enhanced ss. The unbiased estimate of the VCTFPW slope produces similar amounts of errors for the short-term trends as for the decadal trends. ~~While the choice of PW as reference to compute the~~ The percentage of false positives is similar if TFPW-WS is considered as the reference. If MK or TFPW-Y is taken as reference, PW and TFPW-WS have a very low number of ~~type 1 errors is obvious~~ false positive independent of the length of the period, leading to the conclusion that few cases remain uncertain. Note that 5-10% of cases have different ss at the 95% confidence level if MK or TFPW-Y is used, indicating that estimation of the ss using these two methods can have a slight impact on the results. Finally, all the prewhitening methods have a higher number of false positive if VCTFPW is considered as the reference because the added slope at the end of the VCTFPW procedure is smaller than the initial slope and leads to less detectable trends. Note~~(Zhang and Zwiers, 2004, Yue et al., 2002, Blain, 2013, Wang et al. 2015), MK could~~ also ~~be considered as an alternative reference for the power of the test instead of TFPW-Y. If MK is the power of test reference, then the TFPW-Y~~ that the percentage of false ~~negatives is 9.4% for the decadal trends~~ positives of PW and ~~3.5% for the short-term trends. MK and TFPW-Y then each result in 3-10% of false negatives, however the false-negatives are for different cases for the two tests.~~ TFPW-WS remains low (≤ 4%). For the time series considered in this study, the following conclusions can be made: 1) PW (and PW-cor) performs very well with ~~an almost vanishingly~~ a small (≤ ~~0.~~3.5%) number of false ~~negatives and the ss of PW-cor is similar to that for PW~~positives if other prewhitening methods are considered as reference; 2) TFPW-WS has a very low ~~amount of both type 1 and~~number of false positives (less than 2 ~~errors;~~% if PW is taken as the reference); 3) VCTFPW ~~has a very~~exhibits high rate of type 1 ~~and 2~~ errors and should consequently not be used to determine the ss; and 4) ~~it is not possible to determine whether~~The difference in ss between MK ~~or~~and TFPW-Y is related to only 5-10% of the ~~most powerful method~~trends.

The effects of the prewhitening method on the slope (Fig. 2 and 4) also follow the theoretically deduced assumptions:

- The slope ~~of~~estimated on the ~~trend~~original data is always enhanced by the positive ak_1^{data}, which adds a multiple of the t-1 value to the t value (e.g., Eqn 1 and 3). By removing the autocorrelation, PW leads to a strong decrease in the absolute value of the slope that becomes smaller than the MK slope. The CL_{PW} are also somewhat decreased (Fig. 5) due to the decreased mean and variance of the prewhitened time series, relative to the original dataset.
- Due to the detrending procedure, the absolute values of the TFPW-Y slope are larger than the PW slopes and similar to the MK slope values (Fig. 2), even if a tendency to have larger TFPW-Y than MK slopes are observed (Fig. 4b). The CL_{TFPW-Y} are similar to the CL_{PW} because the variance and mean are similar for both the PW and TFPW-Y prewhitened time series.
- Due to the corrected slope and variance, the absolute values of the VCTFPW slopes are much smaller than the TFPW-Y slopes but larger than the PW slopes.

These theoretical assumptions are validated in all cases with the ss trends analyzed in this study. The water vapor mixing ratio and the zero degree level both have a very high autocorrelation (about 0.9 at

one-day time granularity). In such cases, the removal of the autocorrelation can lead to not ss trends and the absolute values of the VCTFPW slope are not always larger than PW slope values.

The slope difference among the methods depends directly on ak_1^{data}. A more nuanced estimate of the slope dependence is shown in Fig. 4 where the differences among the prewhitening methods are plotted. As already mentioned, the VCTFPW method largely mitigates the slope overestimate of the TFPW-Y method at large ak_1^{data} so that the increase of the slope absolute value for increasing ak_1^{data} does not exceed a factor of two (100% difference in Fig. 4a). The difference between VCTFPW and TFPW-Y slopes can reach 200-1000% for the largest ak_1^{data}. The overestimation of the slope by TFPW-Y is much larger than the underestimation by PW if VCTFPW is taken as a reference for slope estimation. TFPW-Y slopes tend to be larger than MK slopes (Fig. 4b), with larger differences at high ak_1^{data} leads. Finally, the slope difference between MK and both TFPW-WS and PW-cor does not depend on ak_1^{data} and the TFPW-WS and PW-cor slopes are usually nearly identical as suggested by their similar relationship to the MK slope (Fig. 4c-d).

The effects of the prewhitening method on CL (Fig. 5) are explained by their modification of the mean and the variance of the data. Removing the lag-1 autocorrelation ~~increases the~~ leads to prewhitened data with a larger variance, but ~~decreases the~~lower mean than the original time series. The correcting factor of $(1-ak_1)^{-1}$ used in the TFPW-WS and PW-cor methods restores the mean (eq. 5), whereas the VCPWTF method restores the initial variance (eq. 9). All increases of the variance make the CL interval wider, whereas the decrease of the mean decreases the CL interval. CL_{TFPW-Y} and CL_{PW} are the narrowest due to lower mean and variance values while $CL_{TFPW-WS}$ and CL_{PW-cor} are the widest due to larger variance induced by the prewhitening and a mean identical to the original data. CL_{VCTFPW} are intermediate with a variance similar to the original data but a lower mean.

4.2 Effects of the time granularity

Averaging is often used to decrease ak_1^{data} in the time series. To investigate this, the ak_1^{data} values are plotted as a function of the time granularity for the last 10 y of all the time series (Fig. 6a). The decrease of ak_1^{data} with aggregation does not have a large impact until granularity is coarser than one-month. For one-month time granularity and less, aggregation leads to an ak_1^{data} difference smaller than 0.2 in 5 of the time series. Three-month and one-year aggregation involve a sharper reduction of ak_1^{data}. Additionally ak_1^{data} for one-year aggregation is, for most of the time series, no longer ss and, sometimes, even negative. The decrease in ak_1^{data} is not continuous with time granularity, with ak_1^{data} often larger for 10 days or one month than for 3 days aggregation. These local minima can be explained by a competitive effect between the ak_1^{data} decrease and a reduction of the measurement variance. ~~The spread of the slopes of the aerosol number concentration for the one-year aggregation on Fig. 2c shows that the yearly data still have a ss ak_1^{data} for the longest periods of 20 and 24 years (see similar cases in Fig. 2). For shorter periods (5 to 9 years), the ak_1^{data} decreases rapidly for averaging longer than 10 days and even becomes negative for yearly averages.~~For the 10 y period represented on Fig. 6, none of the ak_1^{data} values are ss for a one-year time granularity. However, there are cases like the 24 y time series of the aerosol number concentration where ak_1^{data} is still ss for the one-year time granularity. In these cases, prewhitening methods have to be applied, which leads to the spread of the slopes for the various prewhitening methods visible on Fig. 2a.

TFPW-Y and TFPW-WS remove the autocorrelation computed from the detrended data. Fig. 6b and 6c show the difference in ak_1 between the original and the detrended time series as a function of the time

granularity. The ak_1^{detr} continuously increases with aggregation whereas $ak_1^{detr\text{-}prew,n}$ sometimes decreases (e.g., for one-month or three-months aggregations for scattering coefficient and number concentration, respectively). While the differences in ak_1 from the original time series are larger for TFPW-WS than for TFPW-Y, they remain relatively small and exceed 0.05 only in few cases.

Figure 7 presents the effect of the time granularity on ss of the trends for the zero degree temperature level data set for different periods (identified by colors) and various prewhitening methods (identified by symbols). MK and PW-cor are not included since their ss values are nearly identical to the TFPW-Y and PW ss values, respectively. As expected, TFPW-Y exhibits the highest ss, followed by TFPW-WS, while PW and VCTFPW exhibit the lowest ss. The ss always decreases at coarser time granularities for all prewhitening methods until ak_1^{data} becomes not ss, usually at an average of 3 months. This decrease in ss is larger for the PW, TFPW-WS and VCTFPW than for TFPW-Y. For robust trends analyzed (e.g., the period of 40 y in Figure 7), the trend remains ss at the 95% or 90% confidence level for the finest time granularity (3 days for PW and TFPW-WS and 1 month for TFPW-Y), but this is often not the case for weak trends.

When ak_1^{data} is not ss at high time granularity, the prewhitening methods can no longer be applied and the ss is similar for all methods. Without prewhitening, the ss is inversely proportional to the variance reduction caused by the aggregation. For TFPW-Y, the removal of the prewhitening due to not ss ak_1^{data} at three months aggregation corresponds however to a decrease of the ss of the trend. The $ak_1^{detr\text{-}prew,n}$ of the 40 y period is ss for the one-year time granularity as can be seen by the TFPW-WS ss that is different than the ss of the other prewhitening methods (Fig. 7), leading to lower ss than without prewhitening. The increase of the ss with the period length is also obvious, with smaller differences between TFPW-Y and PW for longer periods. The longest period (40 y) and the finest time granularities (1-3 days) lead to no false positives for TFPW-Y, which is not the case for shorter periods or coarser time granularities.

The effect of the time granularity on the slope strongly correlates with the ak_1 time granularity dependence. A decrease of the autocorrelation with aggregation induces a reduction of the prewhitening effects on the slopes leading to a decrease in the differences between slopes (see Figs. 2 and 4).

The loss of ss with coarser time granularities is even more pronounced when evaluated for each month or meteorological season (Fig. 8). This is due to the lower N per season (1/4 for meteorological season and 1/12 for months). Similarly, the decrease in the difference in slopes due to aggregation and the reduction of the prewhitening effects is more pronounced when temporal segmentation is applied due to the reduction of the number of data points in each temporal segment.

Fig. 8 clearly shows that the coarsest time granularities enhance the variability for the different temporal segmentation choices. For example, the interval between the minimum and maximum slopes is 2.3 larger for the monthly average than for the daily average for the scattering coefficient temporally segmented into 12 months (Fig. 8a) and 3.7 times larger for the absorption coefficient with meteorological seasons (Fig. 8b), respectively. In some cases, the sign of the slope changes with the time granularity when the trends are not ss. As already observed in Fig. 2, the CL also increase with time granularity due to the decrease in N. The effects of the time granularity on the ss, the slope and the CL are more pronounced for a monthly than for meteorological seasons temporal segmentation due N being three times lower for the months than it is for the seasons.

4.3 Effects of temporal segmentation to address seasonality

The division of the year into temporal segments is a necessary condition of the MK-test if the data exhibit a clear seasonality. Statistically, it is important to have equivalent segments with similar lengths to obtain similar N per segment. The time series presented in this study are all dependent on phenomena related to the temperature (e.g., atmospheric circulation, boundary layer height, source changes, etc.), and thus change with the meteorological seasons. The seasonality of time series primarily affected by other meteorological phenomena (e.g., the Asian monsoon, which is better characterized by dry and humid seasons, rather than the standard 4 meteorological seasons) have to be carefully studied in order to choose both the appropriate temporal segmentation and the appropriate time granularity. For example, a time granularity that does not respect the seasonal variation of a time series can lead to erratic results (de Jong and de Bruin, 2012).

The effects of the chosen temporal segmentation to address seasonality are presented here for the VCTFPW slope and CL, but they are similar for the other methods as well. The effect of including temporal segmentation on the ss of the yearly trend is rather small with a difference of only 2-3% in the number of ss trends (not shown). The division into four meteorological seasons always results in the largest number of ss trends, while the division into 12 months is less powerful for short periods due to the low number of points for each month ($N \leq 10$) for a 10 y period. The application of no temporal segmentation, which does not met the MK-test requirements in the presence of a seasonality, is less powerful for decadal trends. No systematic effects due to the choice of temporal segmentation on the slope were found. Different temporal segmentation choices lead, most of the time, to comparable slopes. The effect of the prewhitening method is always much more pronounced than the effect of the choice of temporal segmentation.

Figure 9 presents the CL intervals normalized by the trend slope as a function of the time granularity for the aerosol scattering coefficient without temporal segmentation (blue) or divided into monthly (green) or meteorological seasons (red) for several periods between 5 y and 24 y. Due to the decrease of N, finer temporal segments induce an increase of the CL. In the case presented in Fig. 9, monthly segments have CL intervals four times larger than when seasonality is not considered and 2 times larger than meteorological seasons for the longest periods. It should be recalled, however, that not considering seasonality for time granularity finer than one-year is not allowed due to the observed seasonal variation in the aerosol scattering coefficient time series.

In the case of a seasonal MK-test, yearly trend results can be considered only if the trends are homogeneous among the temporal segments (see Sect. 2.1). The division of the time series into four meteorological seasons leads to more homogeneous trends (three times and 25 times for decadal and short periods, respectively) at the 90% confidence level than the division into 12 months (Table 4). Thus, if meteorological seasons correspond to the observed temporal cycle of the studied time series then those seasons should be the preferred temporal division to consider rather than monthly divisions. Monthly segmentation could be considered when the observed variability of time series is shorter or longer than the 3 months length of a meteorological season.

4.4 Effects of length of the time series

As already stipulated under sect. 2.1, a special statistic that deviates from the normal statistic has to be applied to compute the statistical significance for $N \leq 10$. Shorter periods involve smaller N, and N is further

affected by the choice of granularity. The special statistic has to be applied for trends computed on one-year averages and period < 11 years (i.e., N≤10). Note: the effect of the natural variability of a data set on trends computed on short periods will not be directly discussed here, but only the statistical effect on the trends determined for the various time series studied here.

Fig. 10 shows the effect of the reduction of the period length on the slope, the CL and the ss for the aerosol absorption coefficient dataset. The first obvious effect is that the absolute values of the slope are larger for shorter periods and there are large differences both for the individual months and meteorological seasons. Further, these large slopes for short time periods are associated with high CL and low ss. They are due to the cumulative effects of the predominant importance of the first and last years for short periods and to the low N in the time series. For the shortest period considered here (4y), the division of a daily time series into four meteorological seasons involves trends computed with N=360 (=4 years*3 months*30 days) whereas monthly trends for the same time series are computed with N=120 (=4 years*1 month*30 days). The reduction of N by a factor of three explains the larger and more variable slope values, the higher CL and the lower ss of the monthly trends compared to the meteorological season's trends. The effects due to the reduction of N are minimized by the use of daily time granularity, but they are maximized by the use of larger aggregations leading for example to N=12 and 4, respectively, for monthly aggregation (hence the tendency for increases in CL with larger aggregation in Fig. 9). It should be noted that the influence of the length of the time series is usually more important than the choice of time granularity. ~~Also,~~ For short time series, the yearly slopes can differ depending on the chosen temporal segmentation (see, e.g., the yearly slopes of 5y, 6y and 7y on Fig. 10). These results, then, support the standard recommendation of only computing long-term trends on time series of at least 10y.

4.5 Effects of the number of data points

The number of data points N in the time series is a key variable underlying the effects of the time granularity, the temporal segmentation to address seasonality and the period discussed in the previous sections. Because a long-term trend analysis is statistically sound only for time series of at least a decade in length, only decadal and multi-decadal trends are considered in this section. Figure 11 is computed using ~~the new algorithm~~3PW (e.g., Fig. 1) for all decadal trends for all time series, temporal segmentation choices and time granularities and represents the percentage of ss trend as a function of slope and N categories. Fig. 11a shows that time series with robust trends, identified by high normalized slopes, need fewer data points to reach the 95% confidence level significance than time series with less robust trends. In contrast, weaker trends, identified by low normalized slopes, need at least several hundreds or even thousands of data points to become ss. In consequence, the smallest slopes need longer periods and finer time granularities to be identified as statistically significant.

Figure 11 also clearly shows that small N leads statistically to larger normalized slopes and thus demonstrates that trends computed on short periods and with a long averaging time are usually greatly overestimated. The use of prewhitening methods with a large type 1 error will, in addition, falsely indicate ss trends (see Sect. 4.1 and Table 3). The use of MK or TFPW-Y tests on short, highly autocorrelated and highly aggregated time series will definitely produce false positive trends with high absolute slopes.

The effects of the temporal segmentation to address seasonality and the time granularity on the confidence limits are primarily caused by the modification of N. The direct impact of N on CL as a function of slope robustness is plotted on Fig. 11b. As expected, weaker slopes and lower N lead to the largest CL with values of thousands percent of the slope for the worst cases. These high CL are not obviously related to a low ss if a prewhitening method with high type 1 error was used.

5. Discussion

The main effects of the various prewhitening methods on ak_1, the slope, the ss and the CL can be summarized as follow:

- ak_1 depends mostly on the intrinsic characteristics of the time series and on the choice of time granularity
- The CL intervals depend primarily on the number of data points and, thus, the length of the time series, choice of time granularity and of temporal segmentation to address seasonality.
- The ss depends mostly on the robustness of the slope, on the number of data points and on the prewhitening method.
- The slope depends mostly on the prewhitening method, with PW leading to too low slopes and MK, TFPW-Y, TFPW-WS and PW-cor resulting in absolute values of the slope that are too high, considering VCTFPW as an unbiased slope estimate.

The prewhitening methods presented here consider only the lag-1 autocorrelation. Atmospheric processes can, sometimes, be better represented by a higher order of autoregressive models with ss partial correlations at lags>1 (Table 2). These higher order lag correlations could be considered by prewhitening with the appropriate number of lags, but this was not tested during this study. Klaus et al. (2014) applied higher order autoregressive prewhitening to stable oxygen and hydrogen isotopes measured in precipitation and concluded that the ss is mostly decreased by higher order lags correlations whereas the slope is less affected. The effect of AR(2) (auto-regressive process of order 2) autocorrelation with ak_2= 0.2 on the type 1 and 2 errors of MK and TFPW-Y was found to be similar to strong AR(1) autocorrelation (Hardison et al., 2019) in Monte Carlo simulations, for slopes and residual variances derived from 124 ecosystem time series.

Time series with a pronounced seasonality can also exhibit an ak_1 seasonality. Tests were performed in order to compute ak_1 for the various choices of the temporal segmentation instead of on the entire time series. This variant was not further pursued due to the difficulty in applying seasonal ak_1, which were not always ss, leading to the application of the prewhitening method to only some of the temporal segments. These differences in the treatment of each segment yielded erratic results that could not be considered as homogeneous for a yearly trend.

The slopes computed from the various prewhitening methods for the real atmospheric data sets considered here exhibit a large spread and only studies with simulated time series are able to provide insight into the slope bias of the methods. Yue et al. (2002) shows that TFPW-Y leads to a better estimate of the slope than PW, which systematically underestimated the real slope. Zhang and Zwiers (2004) compared the MK, PW and TFPW-WS methods for various slope and ak_1 strengths as well as for various periods (30-200 years). They show that PW underestimates the slope for all slope strengths and periods

for positive ak_1, with the biases being larger for higher autocorrelation. They also note that the biases did not decrease with the length of the time series. In contrast, they find that MK and TFPW-WS overestimate the slope for period < 200 y and high ak_1. In this case they showed that, while the biases are also larger for higher autocorrelation, they are significantly lower for long periods (200y), allowing calculations of almost unbiased slope estimates. These Monte Carlo simulations used yearly time granularity so that their N corresponds to the length of the period. Their evaluation of the importance of N is not as nuanced as presented in our study in which N could be larger than the number of years in the time series for time granularities < 1 y.

The results of our study should be compared to the shortest periods (30 y) of the Zhang and Zwiers (2004) results, where they found an underestimation of the slope by PW and an overestimation by MK and TFPW-WS. Wang et al. (~~2015~~2015[a]) showed that the VCTFPW method leads to root mean square errors (RMSE) of the slope lower than the RMSE for TFPW-Y slopes for all slopes and ak_1 values for a time series period of 30 y. A longer period of 60 y results in lower VCTFPW RMSE only for small slopes. Finally, a recent study (Hardison et al., 2019) shows that both generalized least squares model and the Sen's slope of MK-tests (MK and TFPW-Y) consistently overestimate the trend slope with strong ak_1 and short periods (up to 80% for 10 y and 21% for 20 y). The spread of the estimated slopes increases with ak_1 and is mediated by the length of the period. This suggests that the choice of the VCTFPW method as an unbiased estimator for time series shorter than 100 years is probably a better choice than TFPW-Y, but has to be considered in the context of the CL size in order to obtain a better estimate of the real long-term trend.

All the simulation studies described above report slope per year based on yearly aggregated time series. Their number of data points corresponds then to the time series length. In contrast, N as defined in this study, could be much larger for an equivalent time series length as we considered data aggregations between 1d to 1y. The shortest simulated periods were 10 y (Hardison et al., 2019, Yue and Wang, 2004, Hamed, 2009), 20 y (Yue et al., 2002), 25 y (Bayazit and Önöz, 2007) and 30 y (Zhang and Zwiers, 2004, Wang et al., ~~2015~~2015[a]). All the recommendations of these authors about erratic results for "short periods" always concern decadal or even multi decadal trends and are, consequently, even more relevant for trend results for periods shorter than 10 y.

Based on the results presented in this study as well as the findings from the literature referenced above, the following recommendations can be made:

- A prewhitening method must be used on time series when ak_1^{data} is ss.
- The seasonal MK-test must be used on time series with a clear seasonal cycle. The chosen temporal segmentation to address seasonality for the MK-test has to be compatible with the observed seasonality of the time series.
- Finer time granularities should be used in order to maximize the number of data points and will yield smaller confidence limits and larger ss. The choice of the time granularity must also be compatible with the observed seasonality of the time series.
- Periods shorter than 10 y must be handled with great caution and periods shorter than 8 y should not be used for long-term trend analysis.
- When describing trend results the sign of the slope should not be mentioned if it is not ss, because not ss trends cannot, by definition, be distinguished from zero trends. Moreover, not ss trends have a larger dependency on how the trends are computed (time granularity, period, prewhitening method, temporal segmentation to address seasonality,…).

- In the presence of ss lag-1 autocorrelation, either PW and TFPW-Y together or TFPW-WS should be used to assess statistical significance. MK, TFPW-Y alone and VCTFPW lead to a high number of false positives.
- The slope should be corrected in order to take into account the effect of the prewhitening on the mean and the variance of the time series. We recommend the VCTFPW method to eliminate slope biases, at least for time series shorter than 30 y.
- In presence of ss trends, the confidence limits must also be considered in order to assess the uncertainty in the slope.

6. Conclusion

Several prewhitening methods including solely prewhitening, the trend-free prewhitening from Yue et al. (2002) and from Wang and Swail (2001) as well as the variance-corrected trend-free prewhitening method of Wang et al. (~~2015~~2015[a]) were tested on seven time series of various in-situ and remote sensing atmospheric measurements. Consistent with the literature, the use of MK, TFPW-Y and VCTFPW results in a large amount of false positive results while TFPW-WS results in less than 2% of false positives. The power of the test is good for all the applied MK-tests for the time series considered here.

The effect of ~~the~~ choosing time granularities ranging from ~~1~~one day to one year was also evaluated since a common way to overcome the autocorrelation problem is to average time series to a coarser time granularity. It was found that the ak_1^{data} could remain ss up to at least monthly granularity and was sometimes still ss for yearly averages. Finer time granularities exhibit higher ak_1^{data} leading to a larger difference of the estimated slope by the various prewhitening methods. MK, TFPW-Y, TFPW-WS and PW-cor result in the largest absolute values of the slope and PW the smallest. VCTFPW slopes are found between these two extremes. The confidence limits are much broader for coarser time granularities and the ss is lower, so that ss at the 95% confidence level is rarely achieved. The main impact of keeping a fine time granularity is that it allows computation of the trends on a high number of data points, which improves the power of the test and decreases the uncertainties in the slope.

Since all the time series studied exhibited clear seasonal cycles, two temporal segmentations (12 months and 4 meteorological seasons) were tested for the seasonal MK-test. The segmentation into four meteorological seasons resulted in more homogeneous trends among the segments, a necessary condition to compute yearly trends. The division into meteorological seasons also resulted in a higher number of data points available in each temporal segment relative to division into monthly segments. No systematic effect of the choice of temporal segment on the slope was observed and the difference between temporal segment choices was always much lower than the differences among the prewhitening methods.

Finally, a new 3PW algorithm was proposed combining several prewhitening methods to obtain a better estimate of trend and statistical significance than would be achieved with any individual prewhitening method. PW and TFPW-Y were used to compute the statistical significance of the trend and VCTFPW was applied to estimate the slope. This approach takes advantage of the low sensitivity of type 1 errors of PW, the high test power of TFPW-Y and the less biased slope estimated by VCTFPW.

Code availability

We provide, in dedicated Github repositories hosted within the "mannkendall" organization (https://github.com/mannkendall), a Matlab (DOI: ; https://github.com/mannkendall/Matlab), Python (DOI: ; https://github.com/mannkendall/Python), and R (DOI: ; https://github.com/mannkendall/R) implementation of the algorithm presented in Sec. XX. In particular, these open-source codes, distributed under the BSD 3-Clause License, allow to compute the MK test and the Sen's slope with various prewhitening methods (3PW (default), PW, TFPW-Y, TFPW-WS and VCTFPW). The time granularity, period and temporal segmentation are chosen by the users during the preparation of the datasets. The level of the confidence limits for the MK test, the lag-1 autocorrelation, and the homogeneity between the temporal aggregation can also be defined by the user. The probability for the statistical significance, the statistical significance at the desired confidence level, the Sen's slope and its confidence limits are returned as results. A set a common tests is used to ensure that both the Python and R implementations are consistent with the (original) Matlab implementation of the code.

Author contribution

EA, GM, GR and LV did the measurements, QC and data bank transfer of the time series. MCC developed the new 3PW algorithm, wrote the matlab routines, computed the long-term trends and write the manuscript. FV and AB translated the matlab code into Python and R, respectively. All the co-authors revised the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to thank Patrick Sheridan (NOAA) for mentoring and providing the Bondville data, Derek Hageman (University of Colorado) for programming efforts in data acquisition and archiving, and the on-site technical staff from the Illinois State Water Survey for their long-term support and care for the instrumentation.

References

Andrews, E., Sheridan, P., Ogren, J.A., Hageman, D., Jefferson, A., Wendell, J., Alastuey, A., Alados-Arboledas, L., Bergin, M., Ealo, M., Hallar, A.G., Hoffer, A., Kalapov, I., Keywood, M., Kim, J., Kim, S.-W., Kolonjari, F., Labuschagne, C., Lin, N.-H., Macdonald, A., Mayol-Bracero, O.L., McCubbin, I.B., Pandolfi, M., Reisen, F., Sharma, S., Sherman, J. P., Sorribas, M. and Sun, J.: Overview of the NOAA/ESRL Federated Aerosol Network*, Bull. Amer. Meteor. Soc.*, 100, 123-135, doi:10.1175.BAMS-D-17-0175.1, 2019 .

Bader, S., Collaud Coen, M., Duguay-Tezlaff, A., Frei, C., Fukutome, S., Gehrig, R., Maillard Barras, E., Martucci, G., Romanens, G., Scherrer, S., Schlegel, T., Spirig, C., Stübi, R., Vuilleumier, L., and Zubler, E.; Klimareport 2018, Ed. Bundespublikationen BBL, Nr. 313.001.d, ISSN: 2296-1488, https://www.meteoswiss.admin.ch/content/dam/meteoswiss/de/service-und-publikationen/Publikationen/doc/klimareport_2018_de.pdf , 2019.

Bayazit, M., Önöz, B., Yue, S. and Wang, C.: Comment on "Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test" by Sheng Yue and Chun Yuan Wang, *Water Resources Research*, *40*(8), 1–7, https://doi.org/10.1029/2002WR001925, 2004.

Bayazit, M., and Önöz, B.: To prewhiten or not to prewhiten in trend analysis?, Hydro. Sci., *52*(4), 611–624, https://doi.org/10.1623/hysj.53.3.669, 2007.

Blain, G. C.: The modified Mann-Kendall test: on the performance of three variance correction approaches, Bragantia, Campinas, *72*(4), 416–425, https://doi.org/10.1590/brag.2013.045, 2013.

Brocard, E., Jeannet, P., Begert, M., Levrat, G., Philipona, R., Romanens, G., and Scherrer, S. C.: Upper air temperature trends above Switzerland 1959–2011, J. Geophys. Res. Atmos., 118, 4303–4317, doi:10.1002/jgrd.50438, 2013.

Collaud Coen, M., Andrews, E., Alastuey, A., Arsov, T. P., Backman, J., Brem, B. T., Bukowiecki, N., Couret, C., Eleftheriadis, K., Flentje, H., Fiebig, M., Gysel-Beer, M., Hand, J. L., Hoffer, A., Hooda, R., Hueglin, C., Joubert, W., Keywood, M., Kim, J. E., Kim, S.-W., Labuschagne, C., Lin, N.-H., Lin, Y., Lund Myhre, C., Luoma, K., Lyamani, H., Marinoni, A., Mayol-Bracero, O. L., Mihalopoulos, N., Pandolfi, M., Prats, N., Prenni, A. J., Putaud, J.-P., Ries, L., Reisen, F., Sellegri, K., Sharma, S., Sheridan, P., Sherman, J. P., Sun, J., Titos, G., Torres, E., Tuch, T., Weller, R., Wiedensohler, A., Zieger, P., and Laj, P.**: Multidecadal trend analysis of aerosol radiative properties at a global scale,** Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-1174, in review, 2020.

De Jong, R., and De Bruin, S.: Linear trends in seasonal vegetation time series and the modifiable temporal unit problem, Biogeosciences, 9(1), 71–77, https://doi.org/10.5194/bg-9-71-2012, 2012.

Gilbert, R. O.: Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold Company, New York, 1987.

Hamed, K. H.: Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data, J. Hydrol., 368(1–4), 143–155. https://doi.org/10.1016/j.jhydrol.2009.01.040, 2009.

Hamed, K.H., and Rao, A.R.: A modified Mann–Kendall trend test for autocorrelated data, J. Hydrol., 204, 182–196, 1998.

Hardison, S., Perretti, C. T., Depiper, G. S., and Beet, A.: A simulation study of trend detection methods for integrated ecosystem assessment, ICES J. Marine Sci., https://doi.org/10.1093/icesjms/fsz097, 2019.

Hicks-Jalali, S., Sica, R. J., Haefele, A., and Martucci, G.: Calibration of a water vapour Raman lidar using GRUAN-certified radiosondes and a new trajectory method, Atmos. Meas. Tech., 12, 3699-3716, https://doi.org/10.5194/amt-12-3699-2019, 2019.

Hirsch, R. M., Slack, J. R., and Smith, R. A. : Techniques of trend analysis for monthly water quality data, Water Resour. Res., 18,107–121, 1982.

Klaus, J., Chun, K. P., and Stumpp, C.: Temporal trends in δ 18 O composition of precipitation in Germany : insights from time series modelling and trend analysis, Hydrol. Process., https://doi.org/10.1002/hyp.10395, 2014.

Kulkarni, A., and von Storch, H.: Monte Carlo Experiments on the Effect of Serial Correlation on the Mann-Kendall Test of Trend Monte Carlo experiments on the effect, Meteorologische Zeitschrift, 82–85, 1995.

Laj, P., Bigi, A., Rose, C., Andrews, E., Lund Myhre, C., Collaud Coen, M., Wiedensohler, A., Schultz, M., Ogren, J. A., Fiebig, M., Gliß, J., Mortier, A., Pandolfi, M., Petäjä, T., Kim, S.-W., Aas, W., Putaud, J.-P., Mayol-Bracero, O., Keywood, M., Labrador, L., Aalto, P., Ahlberg, E., Alados Arboledas, L., Alastuey, A., Andrade, M., Artíñano, B., Ausmeel, S., Arsov, T., Asmi, E., Backman, J., Baltensperger, U., Bastian, S., Bath, O., Beukes, J. P., Brem, B. T., Bukowiecki, N., Conil, S., Couret, C., Day, D., Dayantolis, W., Degorska, A., Dos Santos, S. M., Eleftheriadis, K., Fetfatzis, P., Favez, O., Flentje, H., Gini, M. I., Gregorič, A., Gysel-Beer, M., Hallar, G. A., Hand, J., Hoffer, A., Hueglin, C., Hooda, R. K., Hyvärinen, A., Kalapov, I., Kalivitis, N., Kasper-Giebl, A., Kim, J. E., Kouvarakis, G., Kranjc, I., Krejci, R., Kulmala, M., Labuschagne, C., Lee, H.-J., Lihavainen, H., Lin, N.-H., Löschau, G., Luoma, K., Marinoni, A., Meinhardt, F., Merkel, M.,

Metzger, J.-M., Mihalopoulos, N., Nguyen, N. A., Ondracek, J., Peréz, N., Perrone, M. R., Petit, J.-E., Picard, D., Pichon, J.-M., Pont, V., Prats, N., Prenni, A., Reisen, F., Romano, S., Sellegri, K., Sharma, S., Schauer, G., Sheridan, P., Sherman, J. P., Schütze, M., Schwerin, A., Sohmer, R., Sorribas, M., Steinbacher, M., Sun, J., Titos, G., Tokzko, B., Tuch, T., Tulet, P., Tunved, P., Vakkari, V., Velarde, F., Velasquez, P., Villani, P., Vratolis, S., Wang, S.-H., Weinhold, K., Weller, R., Yela, M., Yus-Diez, J., Zdimal, V., Zieger, P., and Zikova, N.: **A global analysis of climate-relevant aerosol properties retrieved from the network of GAW near-surface observatories,** Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2019-499, in review, 2020.

Matalas, N. C., and Sankarasubramanian, A.: Effect of persistence on trend detection via regression, Water Resources Research, 39(12), https://doi.org/10.1029/2003WR002292, 2003.

Maurya, R.: Effect of the Modifiable Temporal Unit Problem on the Trends of Climatic Forcing and NDVI data over India, phD Thesis, https://webapps.itc.utwente.nl/librarywww/papers_2013/msc/gfm/maurya.pdf, 2013.

Nyeki, S., Wacker, S., Aebi, C., Gröbner, J., Martucci, G., and Vuilleumier, L.: Trends in surface radiation and cloud radiative effect at four Swiss sites for the 1996–2015 period. Atmos. Chem. Phys., Hydrol. Process., 19, 13227–13241, doi:10.5194/acp-19-13227-2019, 2019.

Rivard, C., and Vigneault, H.: Trend detection in hydrological series : when series are negatively correlated, Hydrol. Process. *2743*, 2737–2743, https://doi.org/10.1002/hyp, 2009.

Sherman, J. P., Sheridan, P. J., Ogren, J. A., Andrews, E., Hageman, D., Schmeisser, L., Jefferson, A., and Sharma, S.: A multi-year study of lower tropospheric aerosol variability and systematic relationships from four North American regions, Atmos. Chem. Phys., 15, 12487–12517, doi:10.5194/acp-15-12487-2015, 2015.

Sirois, A.: A brief and biased overview of time-series analysis of how to find that evasive trend, WMO/EMEP Workshop on Advanced Statistical Methods and Their Application to Air Quality Data Sets, Annex E., Global Atmosphere Watch No. 133, TD- No. 956, World Meteorological Organization, Geneva, Switzerland, 1998.

Tiao, G. C., Reinsel, G. C., Xu, D., Pedrick, J. H., Zhu, X., Miller, A. J., Dluisi, J. J., Mateer, C. L., and Wuebbles, D.J.: Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation, J. Geophys. Res., 95, 20,507–20,517, 1990.

Wang, W., Chen, Y., Becker, S., and Liu, B.: ~~Linear trend detection in serially dependent hydrometeorological data based on a~~ Variance correction ~~Spearman rho~~pre-whitening method~~, Water, 7(12), 7045–7065., 2015~~ for trend detection in auto-correlated data, J. Hydrol. Eng., 04015033.1-10, doi:10.1061/(ASCE)HE.1943-5584.0001234, 2015[a].

Wang, W., Chen, Y., Becker, S., and Liu, B.: Linear trend detection in serially dependent hydrometeorological data based on a variance correction Spearman rho method, Water, 7(12), 7045–7065. https://doi.org/10.3390/w7126673, 2015[b].

Wang, X. L. and Swail, V. R: Changes of extreme wave heights in Northern Hemisphere oceans and related atmospheric circulation regimes, J. Climate, 14, 2204–2221, https://doi.org/10.1175/1520-0442(2001)014, 2001.

Yoon, J., Pozzer, A., Chang, D.Y., Lelieveld, J., Kim, J., Kim, M., Lee, Y.G., Koo, J-H., Lee, J., and Moon, K.J., Trend estimates of AERONET-observed and model-simulated AOTs between 1993 and 2013, Atmos. Environ., 125, 33–47, http://dx.doi.org/10.1016/j.atmosenv.2015.10.058, 2016.

Xian, T., and Homeyer, C. R.: Global tropopause altitudes in radiosondes and reanalyses, Atmos. Chem. Phys., 19, 5661–5678, https://doi.org/10.5194/acp-19-5661-2019, 2019.

Yue, S., Pilon, P., Phinney, B., and Cavadias, G.: The influence of autocorrelation on the ability to detect trend in hydrological series, Hydrol. Process., 16(9), 1807–1829. https://doi.org/10.1002/hyp.1095, 2002.

Yue, S. and Wang, C. Y.: The applicability of pre-whitening to eliminate the influence of serial correlation on the Mann-Kendall test, Water Res. Res. 38(6), 10.1029/2001WR000861, 4–1–7, 2002.

Yue, S., and Wang, C.: The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series, Water Resources Management, *18*(3), 201–218. https://doi.org/10.1023/B:WARM.0000043140.61082.60, 2004.

Zhang, X. and Zwiers, F. W.: Comment on "Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test" by Sheng Yue and Chun Yuan Wang, Water Resour. Res., 40, W03805, doi:10.1029/2003WR002073, 2004.

Zhang, X., Zwiers, F. W., and Li, G.: Monte Carlo experiments on the detection of trends in extreme values, J. Climate, 17(10), 1945–1952, https://doi.org/10.1175/1520-0442(2004)017, 2004.

Tables

Table 1: Advantages and disadvantages of the MK-test and of the various prewhitening methods.

Method	How it works	Advantages/Disadvantages
MK	• Applied on the data without modification	• High type I error • High test power • slope increased by ak_1^{data}
PW (Kulkarni & von Storch, 1995)	• Remove the autocorrelation	• Low type I error • Low test power • Smaller absolute slope
PW-cor	• Remove the autocorrelation • Preserve the slope	• Low type I error • Low test power • Similar slope as MK
TFPW-Y (Yue et al., 2002)	• Remove the slope • Remove the autocorrelation • Add the trend	• High type I error • High test power • Larger absolute slope
TFPW-WS (Wang & Swail, 2001)	• Apply TFPW iteratively until $ak_{1}^{detr\text{-}prew}$ and the slope stay constant: ➢ Remove the autocorrelation ➢ Compute the slope ➢ Remove the trend from the original data • Remove the final $ak_{1}^{detr\text{-}prew}$	• Low type 1 error • High test power • Similar slope as MK
VCTFPW (Wang, 2015)	• Remove the trend • Remove the autocorrelation • Correct the variance similar to initial variance • Add the trend with corrected slope	• Middle type I error • Medium test power • Unbiased slope estimate

*Table 2: Description of the time series: time series with units, monitoring station, period, instrument type, original granularity, ranges (1 and 99 percentiles (*1%ile and 99%ile*)), mean, median and standard deviation (STD), lag-1 autocorrelation of the observations (ak_{1}^{data}) and number of ss partial autocorrelations for the 10 y period (order), number of data in the 10y period (N) and reference.*

Time series	Station	Period	Instrument	Granularity	1%ile 99%ile	Mean Median STD	ak_1^{data} order	N	reference
Aerosol scattering coef. [Mm^{-1}]	BND	1995-2018	TSI Nephelometer	1 h	6.57 167.80	43.51 33.04 33.85	0.60 2	3485	Sherman et al., 2015
Aerosol absorption coef. [Mm^{-1}]	BND	1995- 2018	PSAP and CLAP	1 h	0.51 11.06	3.40 2.85 2.30	0.53 2	3431	Andrews et al., 2019
Aerosol number concentration [cm^{-3}]	BND	1995- 2018	CPC	1 h	283 11636	4139 3674 2517	0.58 2	2979	Laj et al., 2020
Aerosol optical depth	PAY	2006-2015	PFR	1 h	0.025 0.285	0.126 0.113 0.064	0.72 2	641	Nyeki et al., 2019
Tropopause level [m]	PAY	1958-2018	Radio-sonde	12 h	7540 14660	11178 11280 1425	0.70 2	3636	Brocard et al., 2013
Zero degree level [m]	PAY	1958-2018	Radio-sonde	12 h	-859 4437	2333 2457 1208	0.89 3	3640	Brocard et al., 2013
Water Vapor Mixing ratio [g/kg]	PAY	2009-2018	Ralmo Lidar	0.5 h	1.41 11.88	5.90 5.57 2.63	0.88 3	2868	Hicks-Jalali et al., 2019

PSAP=Particle Soot Absorption Photometer, CLAP=Continuous Light Absorption Photometer, CPC=Condensation Particle Counter, PFR=Precision Filter Radiometer.

Table 3: Percent of false positives ~~and false negatives~~ for all data sets relative to a reference test for the MK-tests and prewhitening methods for periods of at least 10y (decadal trends) or smaller than 8y. N is the number of considered trends. PW should be considered as the best reference so that the results are given in bold. MK, TFPW-Y and VCTFPW have a higher number of type 1 errors and should not be considered as reference so that these results are given in italic.

Period	~~Type of error~~	MK	TFPW-Y	TFPW-WS ~~PW~~	PW/PW-cor	VCTFPW
≥ 10y N=~~2185~~2219	~~False positive~~	~~33~~**32.5**	**37.**~~4~~1	**1.7**	**reference**	~~0~~**47.0** ~~48.5~~
		31.8	36.1	reference	0.7	46.4
		reference	*9.4*	*0.2*	*0.3*	*26.4*
	~~False negative~~	*5.*~~3~~*0*	*reference*	*0.2*	*0.2*	~~0.2~~*24.8* ~~26.1~~
		15.7	*18.4*	*4.0*	*3.5*	*reference*
< 8y N=~~1045~~1067	~~False positive~~	~~19.8~~**16.0**	**14.**~~3~~1	~~1.1~~**0.7**	**reference**	~~0.0~~**36.6** ~~44.9~~
		15.9	13.9	reference	0.5	36.7
	~~False negative 7.0~~	*reference*	*3.0*~~.3~~	*0.*~~0~~*1*	*0.*~~3~~*0*	~~36.6~~*28.1*
		5.0	*reference*	*0*	*0.0*	*29.7*
		8.4	*8.1*	*1.3*	*1.1*	*reference*

Cellules supprimées

Cellules supprimées

Cellules supprimées

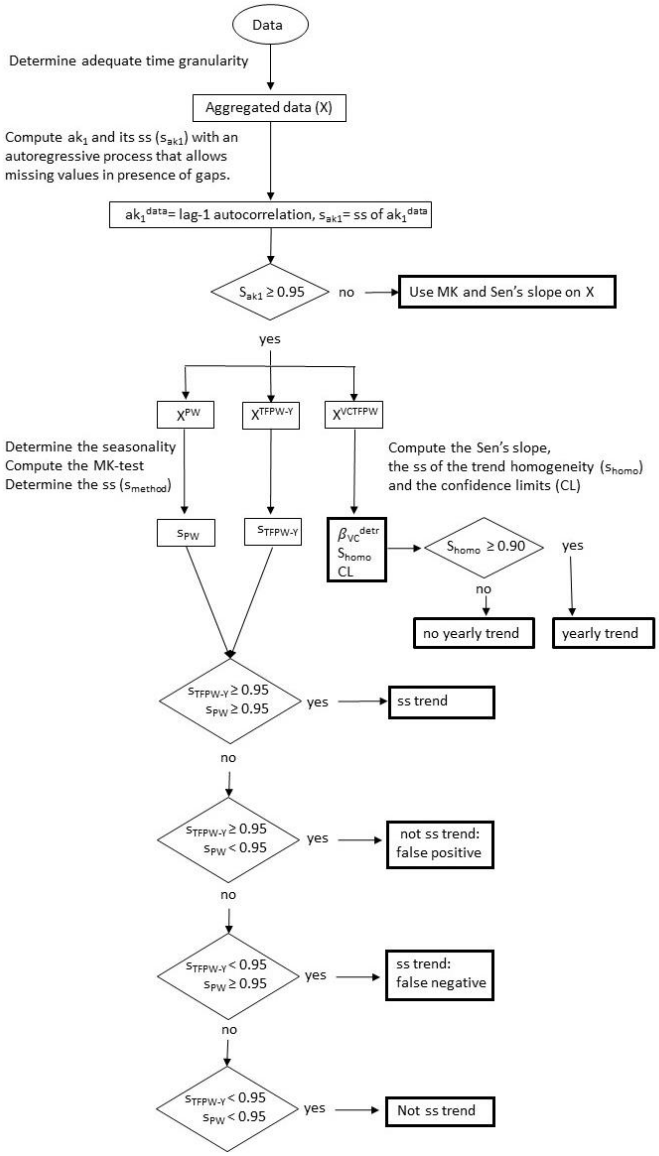
Cellules supprimées

Cellules supprimées

Table 4: Percentage of yearly trends with homogeneous temporal segments as a function of the type of segment (month or season), of the prewhitening method and of the length of the periods based on all seven time series considered in this study.

Period	Method	Months	Meteorological seasons
≥ 10y N=115	VCTFPW	26.1 %	80.0 %
	TFPW_Y	25.2 %	86.1 %
< 8y N=55	VCTFPW	5.5 %	74.5 %
	TFPW_Y	5.5 %	80 %

Figures


Data
Determine adequate time granularity
Aggregated data (X)
Compute ak1 and its ss (sak1) with an autoregressive process that allows missing values in presence of gaps.
ak1data= lag-1 autocorrelation, sak1= ss of ak1data
Sak1 ≥ 0.95
no
Use MK and Sen's slope on X
yes
XPW
XTFPW-Y
XVCTFPW
Determine the seasonality
Compute the MK-test
Determine the ss (smethod)
Compute the Sen's slope,
the ss of the trend homogeneity (shomo)
and the confidence limits (CL)
sPW
sTFPW-Y
βVCdetr
Shomo
CL
Shomo ≥ 0.90
yes
no
no yearly trend
yearly trend
sTFPW-Y ≥ 0.95
sPW ≥ 0.95
yes
ss trend
no
sTFPW-Y ≥ 0.95
sPW < 0.95
yes
not ss trend:
false positive
no
sTFPW-Y < 0.95
sPW ≥ 0.95
yes
ss trend:
false negative
no
sTFPW-Y < 0.95
sPW < 0.95
yes
Not ss trend

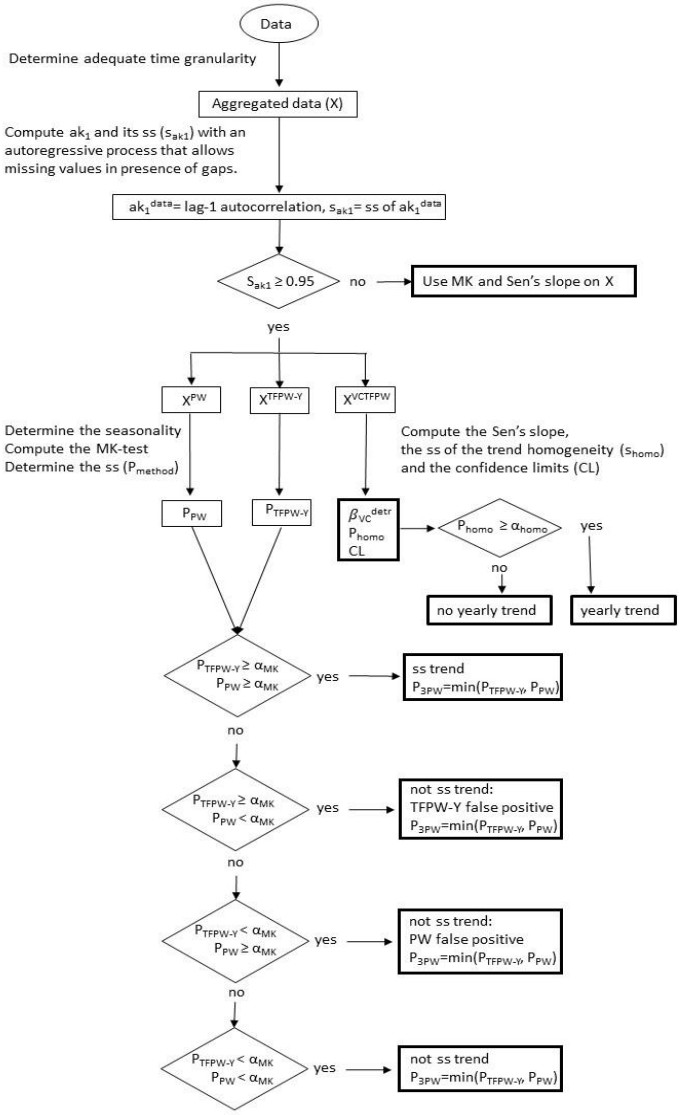


Figure 1: Scheme of the new 3PW algorithm. α_{MK} is the desired confidence limit for the MK test and α_{homo} the desired confidence limit for the homogeneity test between temporal segments. The values applied for this study are α_{MK}=0.95 and α_{homo}=0.90.

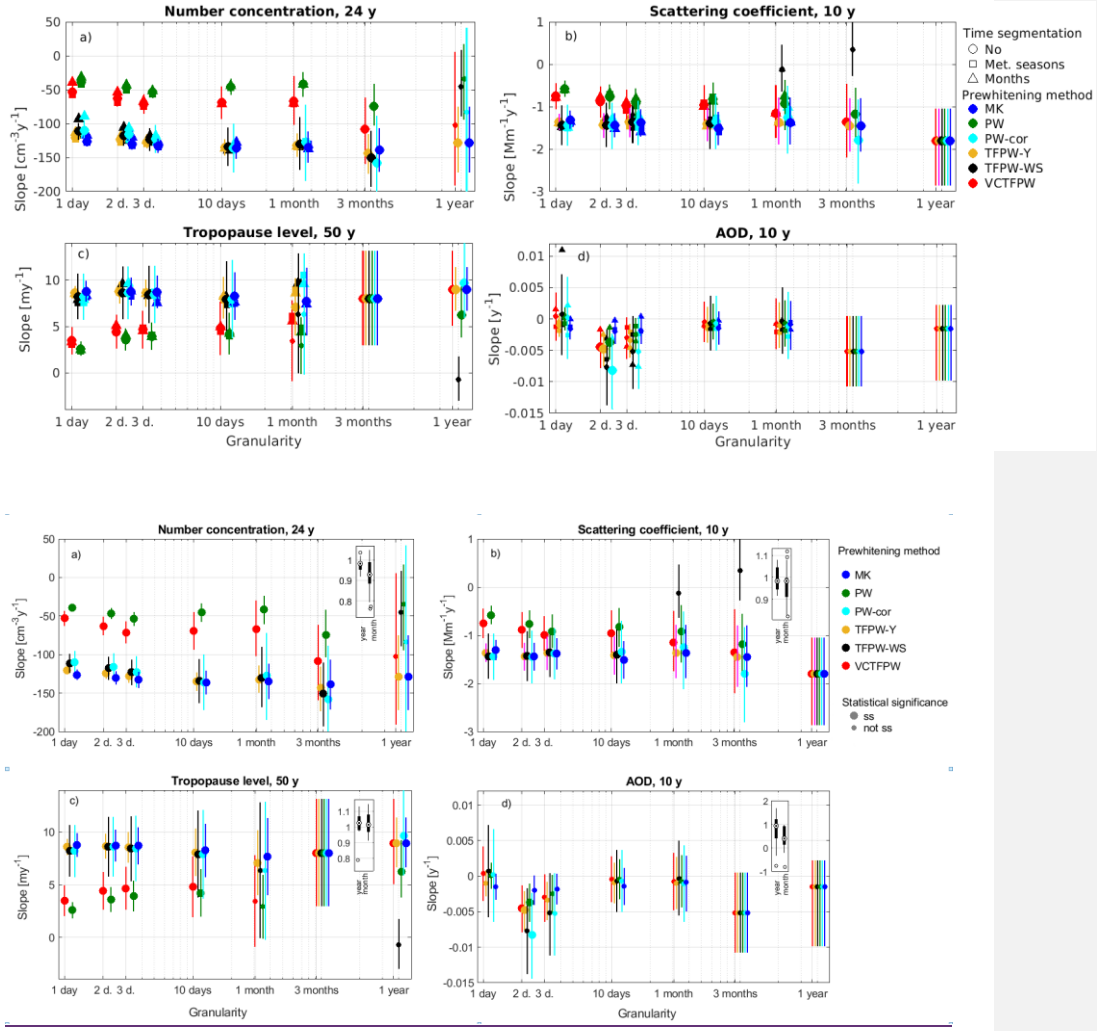


Figure 2: Slope and confidence limits as a function of the time granularity for MK and the five prewhitening methods (indicated by colors) ~~*and for various temporal segmentation choices (indicated by symbols)*~~ *for a) the aerosol number concentration for the 24 y period, b) the aerosol* ~~*absorption*~~*scattering coefficient for the 10 y period, c) the tropopause level altitude for the 50 y period, and d) the AOD for the 10 y period. Larger symbols indicate ss trends and confidence limits are plotted only without time segmentation for clarity purposes.* *Inserted boxplots indicate the median, the quartiles and the whiskers of the ratio between the slopes computed with no temporal segmentation (year) and with the temporal segmentation of 12 months (month) over the slopes computed with the temporal segmentation of four meteorological seasons.*

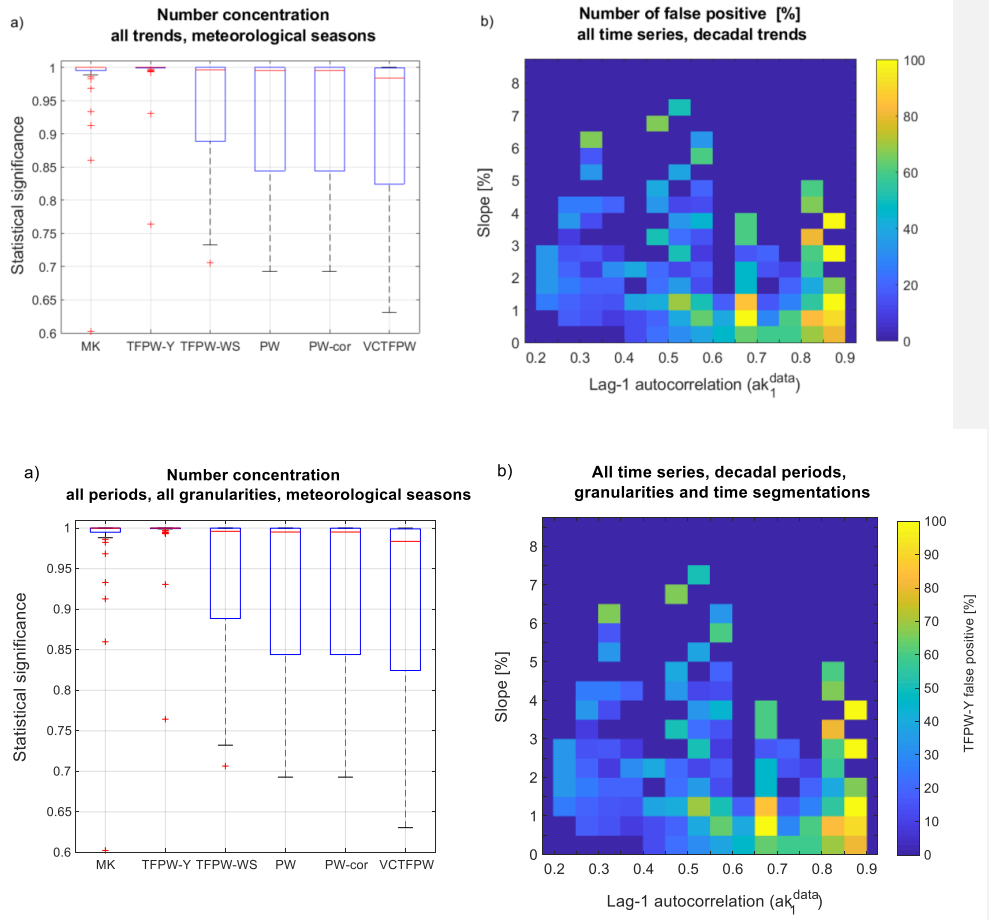


Figure 3 a) Statistical significance of ~~slopes~~trends as a function of the prewhitening methods for the aerosol number concentration for the yearly trends computed from four meteorological seasons time segmentation, for all periods (5y to 24y) and all time granularities (1 day to 1 month). This represents 40 trends~~).~~. The median is represented by the red line, the boxes are the 25% and 75% percentiles, the whiskers the 0.7 and 99.3 percentiles and the red plus signs the outliers. Some outliers are not on the figure for purposes of clarity.

b) Number of TFPW-Y false positives as a function of ak_1^{data} and slope categories for all the computed trends of all time series for all decadal periods. Categories with less than 3 points are not plotted.

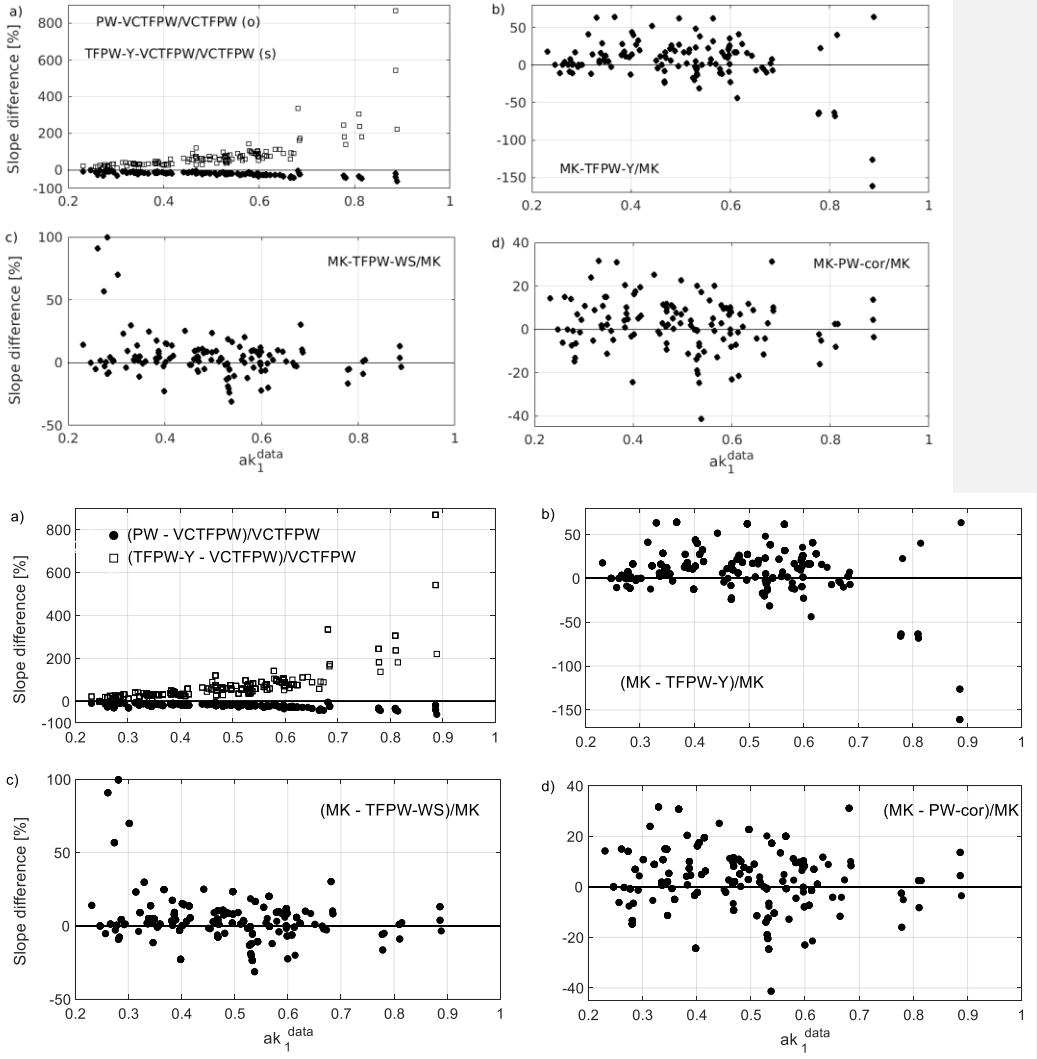


Figure 4: Slope differences as a function of ak_1^{data} from the original data for all datasets, granularities and periods and for meteorological season time segmentation: a) PW minus VCTFPW slope (filled dots) and TFPW-Y minus VCTFPW slope (open squares) normalized by the VCTFPW slope, b) MK slope minus TFPW-Y slopes, c) MK minus TFPW-WS slopes and d) MK minus PW-cor slopes. The slope difference in b) c) and d) are normalized by MK slope. Not ss trends (PW taken as reference) are not plotted since the slopes cannot be distinguished from zero trend. Note the different y-axis ranges on these plots.

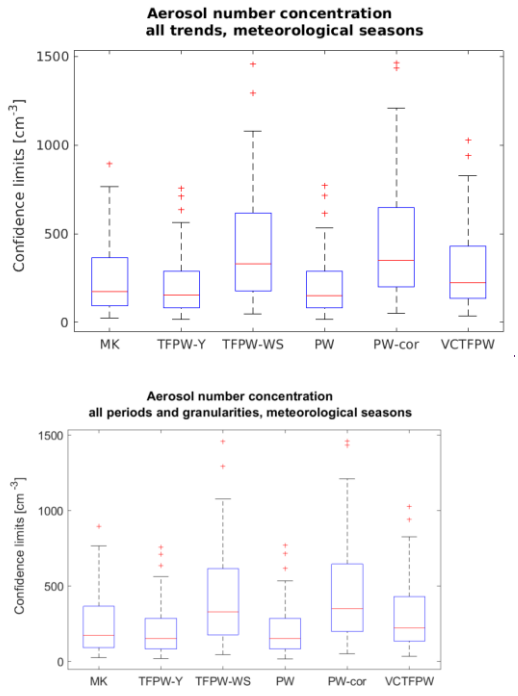


Figure 5: Distribution of the confidence limit intervals of the slope for the trend in aerosol number concentration for all periods (5y-24y) and time granularities (1 day-1 month) as a function of the method for the meteorological ~~seasons~~season temporal segmentation. Box-whisker plotting as described for figure 3a.

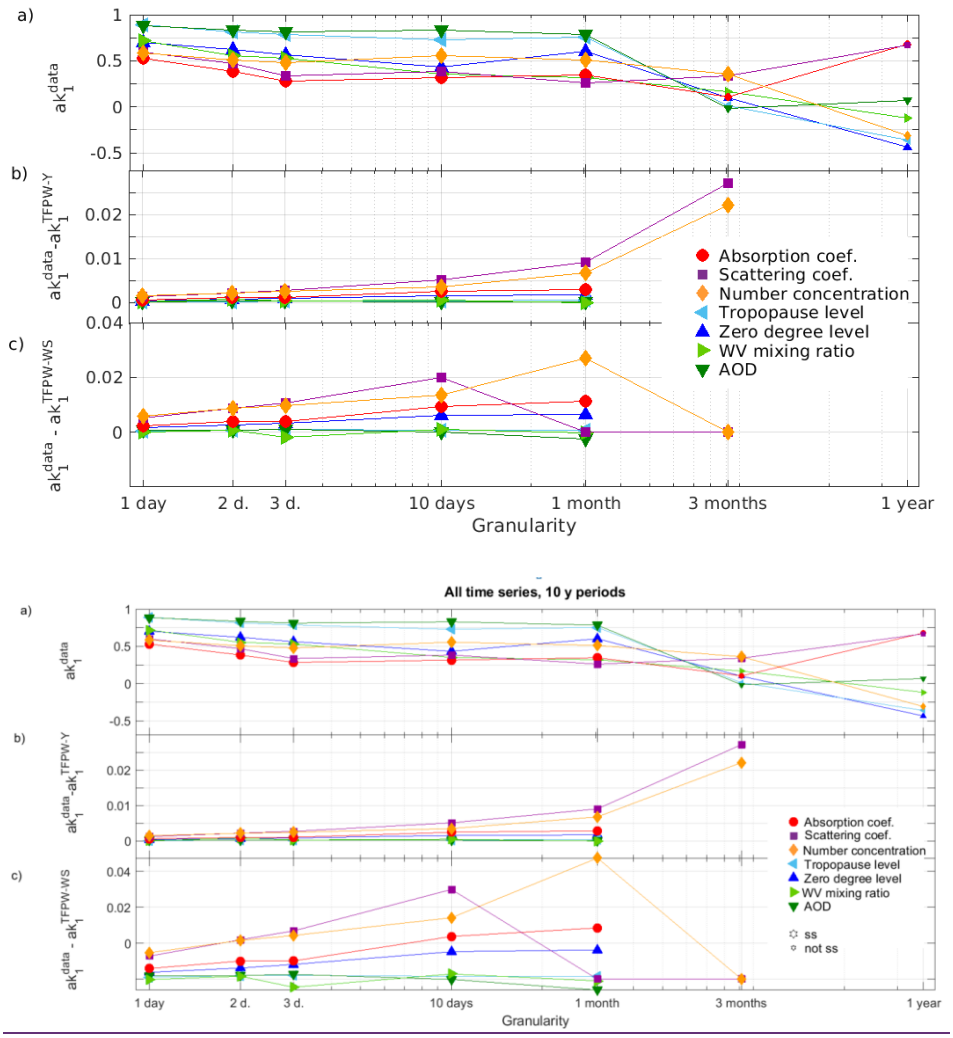


Figure 6: a) Lag-1 autocorrelation (ak_1^{data}) of the original data as a function of the time granularity for the 10 y time series of all ~~time series,~~ parameters, bigger symbols correspond to ss ak_1^{data}, b) ak_1 difference between the original data and the TFPW-Y data, and c) ak_1 difference between the original data and the TFPW-WS data. For b) and c) only ss cases are plotted because prewhitening methods are not applied when ak_1 is not ss.

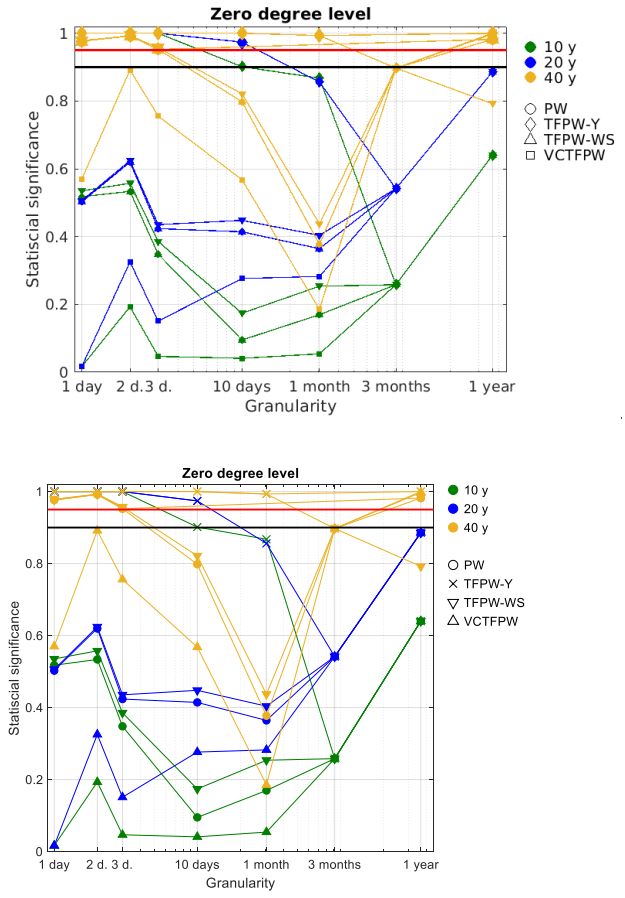


Figure 7: Statistical significance of the trends as a function of the time granularity and prewhitening methods for the zero degree level time series for 10y, 20y and 40y periods without temporal segmentation to address seasonality. The horizontal red and black lines correspond to the threshold of 95% and 90% confidence ~~level~~levels, respectively~~, and ss trends are also emphasized by bigger symbols~~.

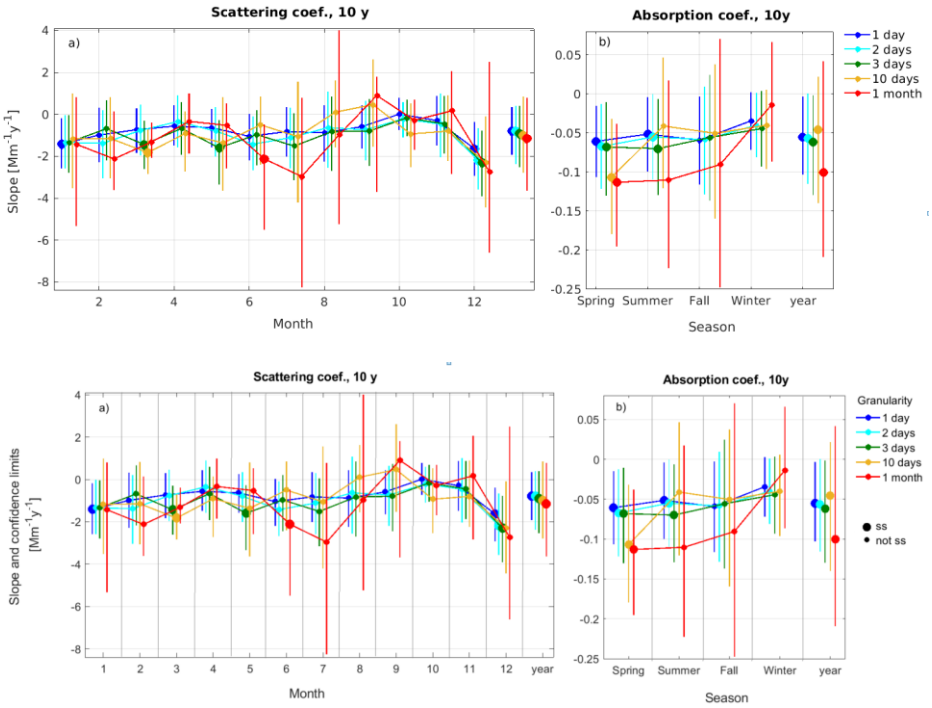


Figure 8: VCTFPW slope (dots) and CL (vertical lines) as a function of the time granularity for the division of the time series into a) 12 months for the 10 y aerosol scattering coefficient and b) into four meteorological seasons for the 10 y aerosol absorption coefficient. Larger symbols indicate statistically significant slopes computed from ~~the new algorithm~~3PW.

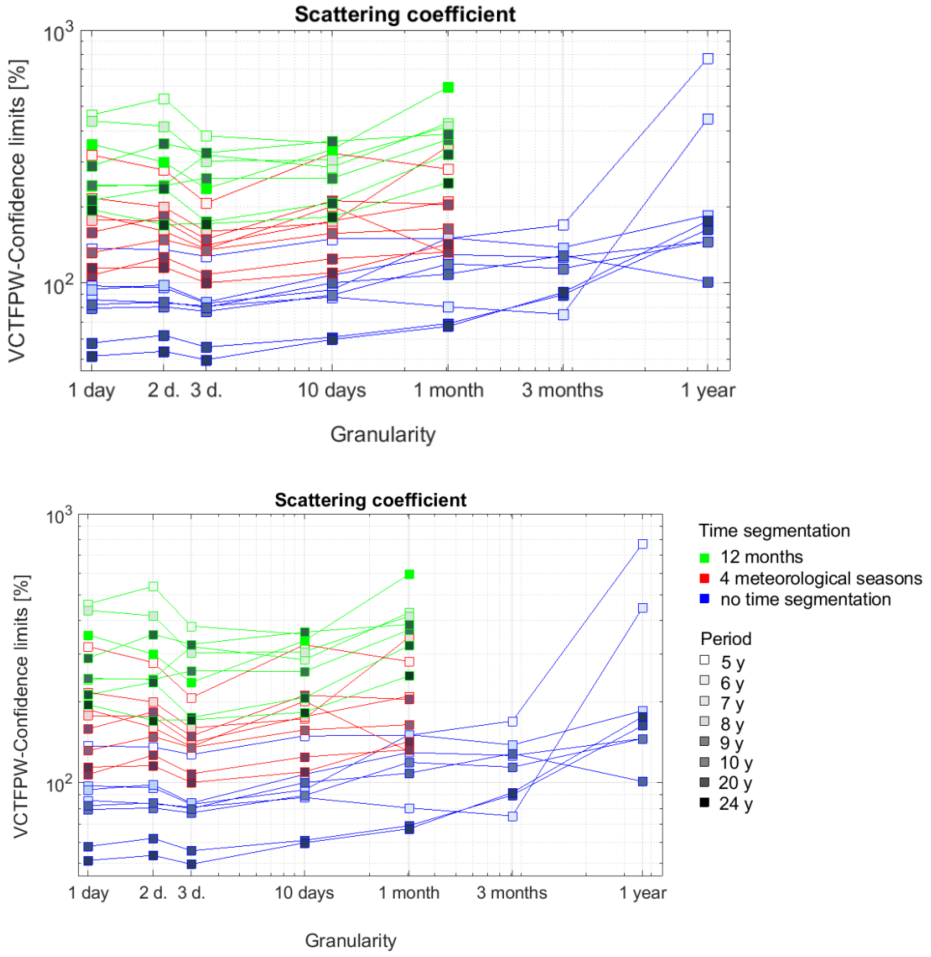


Figure 9: Confidence limits of VCTFPW as a function of the time granularity for various ~~lengths~~periods of the aerosol scattering coefficient time series. Blue represents for no consideration of seasonalities; red represents ~~division~~time segmentation into ~~4~~four meteorological seasons and green ~~represents division~~ into 12 months. The color shading corresponds to the length of the period from 5 y (lightest) to 24 y (darkest).

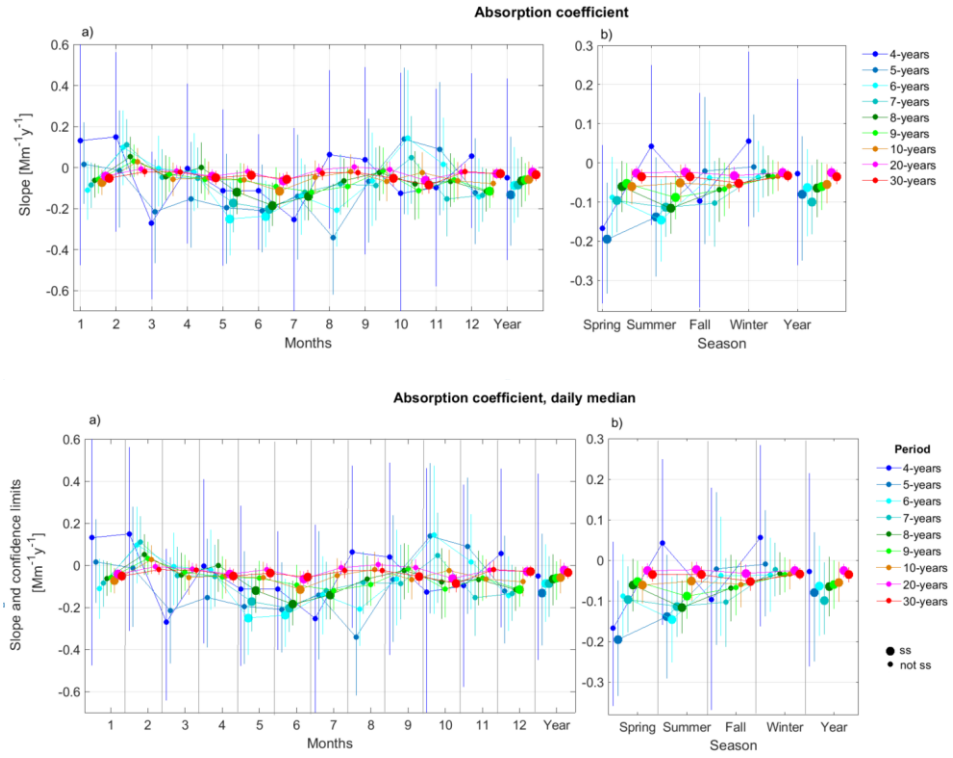


Figure 10: VCTFPW slopes (dots) and CL (vertical lines) as a function of various periods ending in 2018 for the daily aerosol absorption coefficient for the division of the time series into a) 12 months and b) four meteorological seasons. Colors represent time period lengths and bigger symbols represent ss trends.

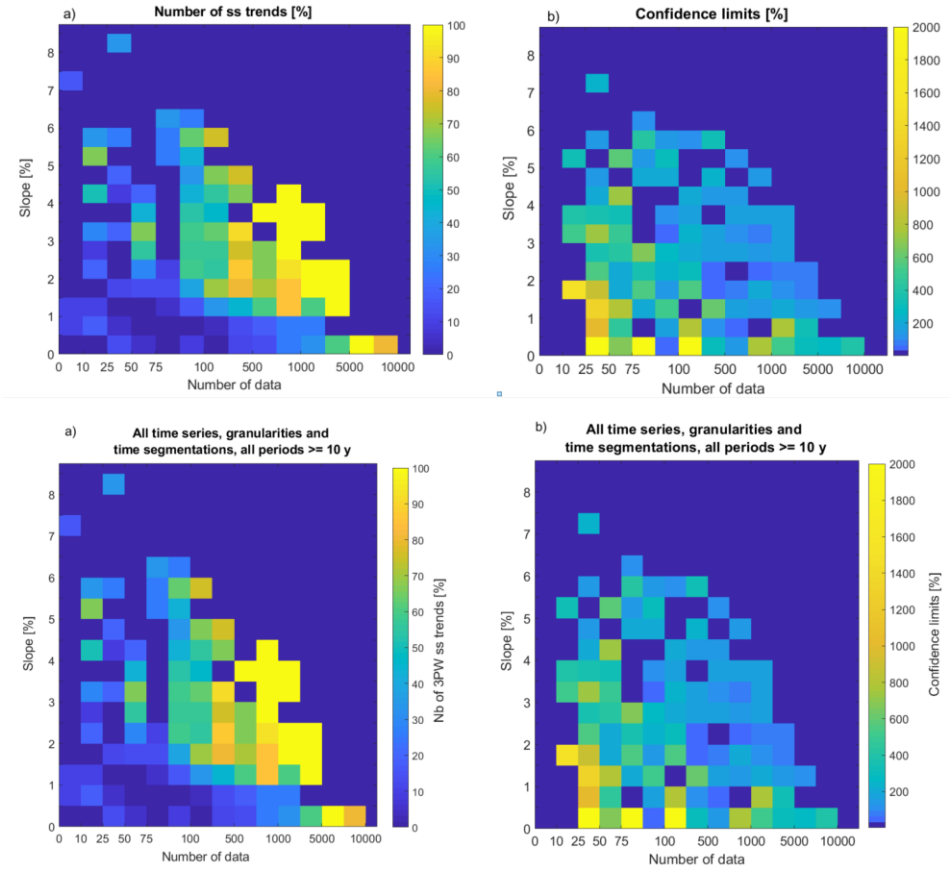


Figure 11: a) The percentage of 3PW ss trends ~~from the new algorithm~~ (sect. 2.3) and b) mean confidence limits normalized by the slope as a function of slope normalized ~~slope~~by the median and N categories for all ~~computed trends with~~time series, granularities and time segmentations and all period of at least a decade. The slopes are binned regularly (bin size = 0.5%) but N categories are irregular. Cells with less than 3 results were discarded in panel a).