Response to reviews

The authors are grateful for all the helpful comments of the two anonymous reviewers, which have resulted in an improved manuscript. Substantial changes have been made to the original manuscript, including the following: (1) the removal of convexity from the testing procedure; (2) the implementation of the Benjamini and Hochberg procedure to control the false discovery rate of the geometric test; (3) the addition of an in-depth discussion explaining the fundamental differences between the geometric and areawise tests; (4) the inclusion of sensitivity studies in Sect. 5; (5) the addition of a brief discussion of how the results of the methods would change if another analyzing wavelet was used; (6) the addition of two figures, which better illustrate the methods. Changes in nomenclature and mathematical notation have also been made. Discussions of figures have been improved to facilitate the interpretation of results. Finally, results from other climate studies are better integrated with the results obtained using the proposed methods, providing motivation for the application of the methods in future studies.

What follows is a point-by-point response to the two anonymous reviewers. Reviewer comments have been reproduced in bold text and our responses are in plain text. All references to line and page numbers pertain to the original manuscript.

Reviewer 1

The present study will be an important addition to the significance testing of wavelet spectra. It provides what may be a useful alternative to the existing "areawise" test and also provides a new topological approach. Before publication, however, several unresolved issues need to be addressed. First, the consequences of a major assumption made in the method need to be discussed. Second, the authors justify the need for this test by the "multiple testing problem" but then they do not show how their test improves upon the pointwise test that was the motivating issue. Third, they also need to improve the figures and their discussion of them. Fourth, the authors should justify their climate examples that use short segments from available long climate series or use longer ones. Additionally, many minor issues of clarity need to be addressed. Putting some of this work in the context of other climate studies would also be helpful and increase the importance of this study for climate science.

We are grateful for these thoughtful comments, which are addressed in detail in the responses to the general and specific comments listed below.

General comments

I. The proposed geometric test suffers from a binary decision of a pointwise threshold "significance" or not. The authors showed some sensitivity to that threshold. The authors should at least discuss an alternative test, very similar in spirit, which does not use binary assessments: the false discovery rate (Wilks 2006).

In the summary section, a discussion has been has added that offers a potential other method for minimizing the number of false positive results. The discussion includes the following paragraph: "One disadvantage of the geometric and areawise tests is that they require a binary decision in which pointwise and geometric significance levels must be chosen. The binary decision can be circumvented by applying a *p*-value adjustment procedure to the wavelet power coefficients directly. For example, one could apply the Benjamini and Hochberg (1995) procedure to the wavelet power coefficients or a modified version of the procedure developed by Benjamini and Yekutieli (2002), which is valid for any dependency structure among the local test statistics. The latter procedure would seem most appropriate given the autocorrelation structure of wavelet power coefficients; however, it is noted that the procedure has considerably less statistical power than the original procedure valid for independent local test statistics, though Wilks (2006) found the Benjamini and Hochberg (1995) procedure to remain powerful even when the assumption of independence is violated."

II. The authors need to justify the "coordinate system" being scale index. The width of the analyzing wavelet changes as a function of scale. I expect that the significance of a wide patch at a small scale to be different from a wide patch at a large scale. The calculation may be "simpler", but it could also be the wrong area to assess. This is my single biggest concern with this testing procedure. The authors need to show that using the actual coordination system would result in the same distribution of chi.

To remove any ambiguity, we have chosen to use the actual coordinate system in the testing procedure, though it was found that using the other coordinate system did not change the geometric significance of patches.

III. The areawise test based on the reproducing kernel of Maraun et al. 2007 is strictly limited to Gaussian white noise. The present authors are also making use of the reproducing kernel in their equation 10 and subsequent steps. a) How does what should be a changing kernel as a function of the noise alter the effectiveness or sensitivity of both the areawise and this geometric test? This is particularly relevant for the comparison testing in Section 4.2 and may be an additional strength of the geometric test if it is less sensitive to the form of the noise and the error in the kernel.

Maraun et al. (2007) found that the areawise test was insensitive to the form of noise. In particular, the areawise test does not depend on the choice of lag-1 autocorrelation correlation coefficient for red-noise processes. This independence to the form of noise is supported by the comparison of the areawise and geometric tests for different lag-1 autocorrelation coefficients, which has now been added in Sect. 4 and is also now shown in Fig. 4 (now Fig. 6).

b) The comparison testing of Section 4.2 needs to be performed on different AR1 noises from an AR1 parameter of 0 to nearly 1. The reproducing kernel of Maraun et al. 2007 becomes less and less relevant as the auto-correlation increases, but the area of random significant patches could continue to grow as the AR1 parameter is increased.

The comparison between the two tests for different AR1 parameters has been added to Sect. 4. It turns out that the area of random significance patches is weakly dependent on the AR1 parameter.

IV. The authors need to better motivate including convexity in the testing procedure.

After further investigation, the inclusion of convexity was found to only a play a minor role in the results of the testing procedure. On the other hand, convexity can explain the differences between the areawise and geometric tests. Thus, convexity has been removed from testing procedure but is included in the discussion of why the tests differ. It is noted that the removal of convexity makes the geometric test generally less conservative than the areawise test, a problem that was remedied by controlling the false discovery rate using the Benjamini and Hochberg (1995) procedure. See the response to comment 21 for details.

V. The present method still seems to suffer from the multiple testing problem. If I have 20 patches and find 2 that are geometrically significant, how is the probability that both were still the result of the noise process addressed?

Indeed, the present method still suffers from the multiple testing problem. The multiple testing problem has been resolved through the application of the Benjamini and Hochberg (1995) method. A new section (now Sect. 4.2) has been added, which describes the procedure and the procedure is used throughout the paper to control the false discovery at the 0.05 level. See response to comment 23 for details.

VI. Recent work (Hanna et al. 2014) claimed to detect a trend in the variance of the NAO. The present study's wavelet analysis of the NAO and new statistical significance testing procedure would be the ideal place to evaluate that claim. The authors should comment on any significant changes in variance detected.

A comment about this study was added on page 1343 line 25 to put the results of the method in the context of climate science.

VII. Can the authors provide some computer code or pseudo-code for how to implement their procedure?

We would be pleased to make Matlab code available and will provide such in the journal's supplementary information if this is permitted or by request to the first author.

Specific comments

1. pg 1332 line 19. The introductory juxtaposition of "random" and "meaningful" does not make sense. I think what is meant is stochastic or deterministic. Random structures are meaningful. The assessment is essential to understand the predictability of the system.

The words "random" and "meaningful" have been replaced by "stochastic" and deterministic" on page 1332 line 19.

2. line 1333 line 15. Some reference for the "climate science" procedure of comparing spectra to red noise should be given.

A reference (Hasselman, 1976) is now given on page 1333 line 15.

3. pg 1334 line 5. The use of the phraseology "Moreover, the areawise,..." is confusing. The term areawise has not been introduced or defined.
The term areawise in the context of significance testing has now been introduced on page 1333 line 3 in the introduction section.

4. pg 1334 line 13. "holes" has not been defined. The meaning here is unclear. Please clarify.

An informal definition of a hole has been inserted on page 1334 line 13 and the reader is now referred to Sect. 5 for a more formal definition of a hole.

5. pg 1335 line 20. What is meant by "Another interesting feature emerges: periods of reduced pointwise significance surrounded by regions of pointwise significance." I don't see anything like this indicated on the figure.

The phraseology has been changed on page 1335 line 20 to clarify the feature of interest. Also, included on the corresponding figure is a label to help guide the reader.

6. pg 1336 line 4, "reproducing kernel" should be defined before its importance is discussed.

The importance of the reproducing kernel is now discussed on page 1336 line 4.

7. pg 1336 lines 8 and 9. The structure of the vaguely referenced equation relating the reproducing kernel to the correlation between wavelet coefficients is important to the argument and explanation here. The equation should be reproduced and then cited.

The equation for the correlation structure of wavelet coefficients has now been reproduced and cited on page 1336 line 8.

8. pg 1336 line 9. Is the area "given by" the reproducing kernel, or is the typical patch area the area of the reproducing kernel?

The phraseology on page 1336 line 9 has been changed to "the typical patch area is the area of the reproducing kernel."

9. pg 1336 line 9 and following. What is meant precisely by area in this context should be defined. Particularly because the subsequent area of the geometric test is different.

A sentenced was added to page 1336 line 15 to clarify what the area is in the context of the areawise test. Another clarifying sentence was added on page 1336 line 20 to help distinguish between the area used in the geometric and areawise tests.

10. pg 1336 line 16. Is the test for "any" reproducing kernel or the reproducing kernel corresponding to the analyzing wavelet?

The test should be performed with the reproducing kernel corresponding the analyzing wavelet and the wording has been changed on page 1336 line 16 to reflect that.

11. page 1336 line 20. One is not assessing the significance of the wavelet coefficients. One is assessing the wavelet spectrum or the coefficients squared.

“Wavelet coefficients” have been changed to “wavelet power coefficients” on page 1336 line 20

12. page 1337 line 4-10. The discussion of the illustration needs improvement. Please provide the reader some specific examples so that they know what they are looking at. At what time and scale are some of these features seen? Why is Figure 1 plotted so differently from Figure 2? What are the red noise parameters being used?

Fig. 1 and Fig. 2. (now Figs. 3 and 4) are now plotted identically, with light gray shading representing those 5% pointwise significance patches that are geometrically significant and dark gray shading indicating those 1% pointwise significance patches that are geometrically significant. A more detailed discussion has been added to page 1337 line 4 to highlight some specific features.

13. equation 7. It may seem pedantic, but please include how this discrete equation 7 follows from Green’s theorem, which applies to integrals (perhaps in a small appendix or provide a reference).

A new appendix (Appendix C) has been added to the manuscript where the derivation of Green’s theorem for a polygon is given. Equation (7) is now also cited.

14. equation 7. The variable n is not defined.

n, which has been changed to *m,* is now defined on page 1338 line 7

15. equations 8 and 9. Provide a reference for this definition of a centroid. Doesn’t it have a fundamental problem when polygons intersect, such as we see in Fig. 2 at a scale of 5 years around 1990?

A reference (Worboys and Duckham, 2004) has been added on page 1338 line 12. Although graphically the polygons appear to be intersecting, problems in computing the centroids have never arose because the polygons don’t actually intersect when examined more closely.

16. pg 1340 line 3. Why is it “...noted that all holes...” When would holes be relevant for this procedure? Please clarify or remove.

The exclusion of holes in the calculation of holes is now justified on page 1340 line 3. The convex hull is not defined for sets with holes because line segments can always leave sets containing holes.

17. pg 1340 line 14. Do patches of equal area “need to be distinguished”? I would think that they should have the same significance that would depend on how often they occur.

The procedure has been changed to not include convexity in the calculation of the null distribution, though the removal of convexity was found to make the test, on average, less conservative than the areawise test. Instead, convexity is now used to explain differences between the areawise and geometric tests in Sect. 4.2.

18. pg 1340 line 17. I don't understand why two patches with the same normalized significant area, regardless of shape, don't have the same significance. The authors need to better motivate in what context this difference in geometry matters. One could simply be testing the area without regard to this shape issue. If there was no reliance on the reproducing kernel (which becomes less and less relevant for strongly auto- correlated noise), the test should be on the distributions of A and nothing else.

After a careful investigation, it was determined that convexity only plays a minor role in the testing procedure but still remains an essential part of the manuscript. See responses to comments 17 and 21.

19. pg 1340 line 20. Is the null distribution of chi independent of the form of the null hypothesis noise? If not, then the dependence should be explicitly stated here.

The choice of null hypothesis does not seem to impact the null distribution substantially. The lack of dependence on the lag-1 autocorrelation coefficients is now explicitly stated on page 1340 line 20.

20. pg 1343 line 2. The "large number" should be stated. Their length in time should also be stated. Does the length matter?

"Large number" has been change to "1000" on page 1343 line 2. The length of the time series were 1000. The length of the time series does matter, at least for the pointwise significance levels used in this study.

21. pg 1343 line 10. How should one interpret the differences between the two tests? If a patch is areawise significant but not geometrically significant in particular, seems to possibly point to a substantial issue in the testing procedures, a problem with including convexity in the geometric procedure, or a problem with the reproducing kernel approach in both tests. These discrepancies need to be addressed and discussed in depth, as it goes to the heart of the point of this paper.

The differences between the two tests arise from the fact that implicit in the calculation of areawise significance level are the convexity and other geometric parameters of a typical patch generated from red noise. An in-depth discussion has been added on page 1343 line 10 describing the difference between the two tests and how such differences should be interpreted.
Two additional paragraphs have been added:
"According to the areawise test, patches with smaller values of $\mathcal{C}$ are less likely to be areawise significant so that it is expected that patches deemed significant by the areawise test will be primarily convex. To test this hypothesis, 10,000 patches arising from red-noise processes with different lag-1 autocorrelation coefficients were generated and the convexity of those patches

deemed areawise significant at the $\alpha_{aw} = 0.05$ level was calculated. The results in Fig. 6c show the mean convexity as a function of the lag-1 autocorrelation coefficients, together with the 95% confidence bound. The mean convexity of the patches was found to be approximately 0.8, regardless of the lag-1 autocorrelation coefficient. An identical experiment was also performed for geometrically significant patches but with the convexity of patches that are geometrically significant at the $\alpha_{geo} = 0.05$ being computed. In contrast to areawise significant patches, patches that were found to be geometrically significant, on average, had lower convexity, the reason for which is that the calculation of α_{geo} makes no assumption about convexity. The *p*-value for the geometric test is thus $p_{geo} = f(A; H_0)$ for some function *f*, contrasting with p_{aw} that depends on convexity. The results of the experiments are consistent with Figs. 5a and 5b, where both the ideal patches have the same geometric significance but the ideal patch in Fig. 5b has a larger p_{aw} so that $p_{aw} > p_{geo}$."

"Convexity cannot fully explain the differences between p_{aw} and p_{geo} for a given patch. More generally, $p_{aw} = g(\mathcal{C}, A, S_1, \ldots, S_R; H_0)$, where S_1 to S_R are shape parameters of the patch, such as aspect ratio and symmetry. For example, for a convex patch whose length in the time direction is long with respect to the reproducing kernel (at some critical level) but thin in the scale direction with respect to the reproducing kernel would be deemed insignificant by the areawise test, though it may have an area much larger than the critical area of the areawise test. Asymmetry with respect to the scale axis, as another example, may also result in a patch being deemed insignificant by the areawise test if, for example, the width of the patch in the scale direction decreases with time. If the normalized areas of such patches are larger than the critical level of the geometric test, the patches will be geometrically significant, though may not be areawise significant if the reproducing kernel is unable to fit inside the narrow portion of the patch. The above arguments suggest that $f(A; H_0) \neq g(\mathcal{C}, A, S_1, \ldots, S_R; H_0)$ and thus the significance of patches as determined by the geometric and areawise tests need not be equal."

22. pg 1343 line 17. Why is significance level of 0.9 being used (I think you mean 0.1).Why not 0.05, as it more common and reduces the risk of the Type I error more?

The significance level has been changed to 0.05 for the areawise test to reduce Type 1 errors.

23. pg 1343 lines 25-27. The multiple testing problem is still not resolved, at least in this discussion and application of the test. In Fig. 1, I see more than 20 patches and 3 are geometrically significant. Couldn't I have gotten that result from chance at the 10% level of the test?

The authors appreciate the critical evaluation of the proposed method. The motivation for constructing the geometric test was to offer an alternative method for dramatically reducing the number of spurious results. However, a simple way of further reducing spurious results detected using the geometric test is to apply the Benjamini and Hochberg (1995) method to the *p*-values of individual patches in a given wavelet power spectrum to control the false discovery rate, the expected proportion of rejected null hypotheses that are actually true. A discussion of the method has been added to page 1342 and the method has been used throughout the text.

Section 4.2 includes the following discussion of the Benjamini and Hochberg (1995) procedure:

“If the geometric test was performed on *K* significance patches at the α_{geo} level, then, on average, one can expect $\alpha_{geo}K$ false positive results, which would make the geometric test permissive for large *K*. It is therefore necessary to reduce the number of false positive results. There are various ways to reduce the number of false positives, including the Walker test, Bonferroni correction, and other counting procedures (Wilks, 2006). Recently, methods for controlling the false discovery rate (FDR) have been developed, where the FDR is the expected proportion of rejected local null hypotheses that are actually true (Benjamini and Hochberg, 1995). In particular, Benjamini and Hochberg (1995) developed a method for controlling the FDR based on the number of local hypotheses being tested and the degree to which the local hypotheses were rejected, contrasting with other procedures that ignore the confidence with which the local tests reject the local hypotheses (Wilks, 2006). Moreover, the method has proven to have high statistical power, especially when only a small fraction of the *K* local tests correspond to false null hypotheses (Wilks, 2006). The procedure will therefore be used to control the false discovery rate of the geometric test, which will facilitate the interpretation of results.

Suppose that *K* local hypotheses were tested, where, in present case, the local hypotheses refer to the testing of each patch individually under the assumption that the results of the individual tests are independent. A global geometric test can be performed at the α_{global} level as follows: Let $p_{(l)}$ denote the *l*th smallest of *K* local *p*-values; then, under the assumption that the *K* local tests are independent, the FDR can be controlled at the *q*-level by rejecting those local tests for which $p_{(l)}$ is no greater than

$$p_{FDR=} \max_{r=1,\ldots,K}\left[p_{(r)}: p_{(r)} \leq q(r/K)\right] \tag{15}$$

$$\max_{r=1,\ldots,K}\left[p_{(r)}: p_{(r)} \leq \alpha_{global}(r/K)\right] \tag{16}$$

so that the FDR level is equivalent to the global test level. According to the procedure, any local test resulting in a *p*-value less than or equal to the largest *p*-value for which Eq. (16) is satisfied is deem significant. If no such local *p*-values exist, then none are deemed significant and, therefore, the global test hypothesis cannot be rejected. The global geometric test will thus only deem those significant patches with *p*-values satisfying Eq. (16) as significant. Throughout the paper $q = \alpha_{global}$ will be set to 0.05.”

24. pg 1343 line 27. I don’t see any obvious seasonality in the wavelet power spectrum shown. The time-averages of the wavelet power for each season would help to make the “variability” point. It is currently not supported by the figure.

Line 27 on page 1343 was removed because the modified geometric test did not find any significant patches.

25. pg 1344 line 3 and following. I don’t see a period of 32 months or of 12 months plotted on the figure 2. It only goes to a period of 7 months.

The axis label is incorrect and should be “years.” However, to be consistent with other plots the axis labels and limits have been set to months in Figure 2.

26. pg 1344 line 26. The definition and method of calculating a “hole” needs to be given.

A formal a definition of a “hole” has been inserted on page 1344 line 26. The definition uses notions from topology to define what it means for a patch to contain a hole.

27. pg 1345 line 1. What is the sensitivity of the shape and amplitude of Fig. 5 to the choice of autocorrelation. Would 0.9 and 0.1 be different or the same? The authors are making generalizations based on just one parameter setting.

The amplitude of Fig. 5 was found to be independent of the autocorrelation coefficient. Such a lack of dependence has been explicitly stated on page 1345 line 1.

28. pg 1345 line 11. Why is an 80% significance level used here? 90% was used earlier. Both have a larger risk of a Type I error than the traditional 95%. (note that the nomenclature should actually be 20%, 10%, and 5% when the “significance” is being considered rather than the error bar).

In this case, a pointwise significance level of 20% is used to find “holes” and not to assess the significance of the wavelet coefficients squared. The nomenclature on page 1345 line 11 has been changed to 20%, 10%, and 5%. The nomenclature in all the figures and figure captions have been changed as well.

29. pg 1345 and following. What null hypothesis is being compared in this simple test of white noise and a sinusoid. How different were the amplitudes? Is this actually a general result or specific to the parameters chosen for the series? A similar lack of specificity and detail applies to the rest of the discussion through page 1349. The results only have theoretical implications if they are generalizable. From the present discussion, this cannot be assessed.

The authors appreciate the reviewer’s careful reading of pages 1345-1349 and constructive criticism. The comments have resulted in significant improvements to this part of the manuscript.

The null hypotheses used were white noise and red-noise spectra. Their uses are now explicitly stated on page 1345 lines 16 and 17. For the experiment of a single sine wave, sine waves of varying amplitudes were generated to determine if there is an amplitude dependency. It was determined that there was no amplitude dependence, which is now explicitly stated on page 1345 line 16.
A discussion of several other experiments has been added to Section 5.1 page 1348 in order to test how the results would change under a different set of parameters. The additional experiments include using different noise backgrounds, amplitude of the cosines, and signal-to-noise ratios. It turns out, however, that there is only a small difference between the theoretical critical delta *r* for the original experiment and that obtained under very low-noise situations with the cosines having large amplitudes. The low-noise, high-amplitude situation is considered the best-case scenario so that it represents a theoretical maximum.

The following paragraph has been added to address the reviewer's concerns:
"It turns out that even if the above experiment (not shown) was repeated using white-noise background spectra Δr_{crit} would still be equal to 0.45, though more holes were found to appear at signal-to-noise ratios less than 2. It was expected, however, that Δr_{crit} also depends on the amplitudes of the cosines in Eq. 24. Thus, a third experiment was conducted in which the amplitudes of the cosines were allowed to vary from 1 to 50 and f_1 and f_2 were allowed to vary from 0 to 0.5. The experiment was repeated for signal-to-noise ratios from 1 to 20. The results from the experiments (not shown) indicate that for red-noise background spectra and for a signal-to-noise ratio of 20 that Δr_{crit} = 0.53, contrasting with the case for white noise background spectra where Δr_{crit} was found to be 0.51."
Overall, the discussions from pages 1345 through 1349 have been refined by including more details of the experiments performed.

30. pg 1349 line 9. Couldn't the same effect be found in the linear AR2 model for some choices of its two parameters? I don't think that nonlinearity needs to be invoked to see this behavior of "holes".

An AR2 process could certainly produce holes but not to the extent that a nonlinear time series could. In fact, the amount of holes generated from AR2 processes was found to be similar to that of AR1 processes.

31. pg 1349 line 19. What is meant by "phase coherent oscillations"?

A definition of phase coherence was added to page 1349 line 19.

32. Is there a sensitivity to the dj used in the wavelet analysis? If so, this should be stated.

dj controls the spacing between discrete scales, where a smaller dj will give better scale resolution. If the dj is too large there will not be adequate sampling in scale so that some features will be missed. The maximum value of dj depends on the analyzing wavelet used, though dj is not intrinsic to the wavelet function so sensitivity of results to dj seems unlikely.

If dj is somehow intrinsic to the wavelet function, this should be referenced or shown. I do not know of any support for this idea. It is a tunable parameter as far as I know.

To the authors knowledge, dj is not intrinsic to the wavelet function itself but to how the scales are discretized.

33. Fig. 1. I recommend some other symbol or method to indicate the geometrically significant patches. Stippling or hatching them would help them better stand out. The x makes me think that these patches have been eliminated, rather than highlighted.

Gray shading has been used to highlight those significance patches that are geometrically significant and thick red contours are used to indicate the areawise significance regions in Figs. 1 and 2 (now Figs. 3 and 4).

34. Fig. 1. The "I_sim" indicated on the figure should be defined in the caption.

Because the false discovery rate is used in the ideal and climatic examples, I_sim no longer appears on Figs. 1 and 2 now (Figs. 3 and 4) and related figures.

35. Fig. 1, Somewhere the actual wavelet spectral values should be shown to get a sense of how the regions passing the pointwise test compare to those not passing.

For clarity, the full wavelet power spectra have been plotted separately from the areawise and geometric test results and are now Figs. 1 and 2. The results for the areawise and geometric tests are now Figs. 3 and 4.

36. Fig. 1. I don't see in the text where "normalized" has been defined.

Normalized wavelet power has now been defined on page 1335 line 16.

37. The red noise equation being used should be shown and how the parameters are fit should be stated. Some discussion of why one is testing against discrete red noise compared to continuous red noise should be given. The spectra are not the same.

The red-noise equation used is now shown on page 1335 line 12 and the equation for a theoretical red-noise background is also shown. Two methods are now cited that are used for estimating AR1 parameters in Sect. 3.1. A discussion of those methods seems beyond the scope of the paper so that the reader is referred to books describing them in-depth. The use of a discrete red-noise spectrum was discussed in Torrence and Compo (1998) and has since been routinely applied to wavelet power spectra of climatic time series. Instead of adding a discussion of the use of a discrete Fourier spectrum in wavelet analysis, which would add to the overall length of the paper, the reader is referred to Torrence and Compo (1998) for a more in-depth discussion.

38. Why are Fig. 1 and Fig 2 plotted so differently? Also, please double check that the time series in Fig 2a is monthly resolution. It does not appear to be monthly. It looks like it has been smoothed.

Figs. 1 and 2 are now plotted identically. See response to comment 12. The data were checked and found to be monthly resolution.

39. Why are such short time series considered? The NAO extends back to the early 1800s. Nino3.4 goes back to 1850 in several datasets. I would think that the longest possible record would help in defining the distribution of areas. It would also push out the cone of influence.

Longer time series for the NAO and Nino 3.4 index are used throughout the paper. In particular, the time period has been extended to 1870-2013 to better illustrate the applicability of the proposed methods.

40. pg 1350 line 10. No one has shown that "spurious results" are "ubiquitious" in wavelet spectra and neither has this paper. In contrast, Maraun et al. 2007 showed (Appendix C)

that the sensitivity of pointwise and areawise tests depends on the signal to noise of the series. As exemplified in the discussion of Fig. 1, this geometric test still has the multiple testing problem.

We agree. Line 10 on page 1350 has been deleted.

Technical corrections

1. pg 1338 “would have it did not contain” needs an “if”. 2. Fig. 2. I think something is wrong with the y-axis as given. Nino3.4 should not have so much power at periods of 5 months and the cone-of-influence for monthly data should be at much longer scales than 7 months

The text on page 1338 has been corrected. The labels on the y-axis should have been months, not years. The axis label and axis limit of Fig. 2 have been corrected.

Reviewer 2

The manuscript describes new methodology for advancing statistical significance testing of wavelet power spectra. The methodology builds from previous work in significance testing and addresses several problems that previous work did not address. The manuscript is generally very well written and concise, in particular given that it blends sophisticated time-frequency decomposition, statistical, and topological concepts. The work represents advancement in the quantitative interpretation of wavelet analysis, which is a topic that has received criticism. Therefore, I recommend it for publication in Nonlinear Processes in Geophysics. However, I have several general and specific comments that should be addressed before publication.

We are grateful for these thoughtful comments, which are addressed in detail in the responses to the general and specific comments listed below.

General comments

(1) The manuscript describes significance testing based on geometric and topological properties of regions within the wavelet power spectrum. These properties are closely tied to the parameters of the wavelet. The manuscript only considers the Morlet wavelet with ω0=6. The authors should discuss the sensitivity of their results to other commonly used wavelets or wavelet parameters that provide more (less) precision in the time domain and less (more) in the frequency domain compared to Morlet.

The reviewer is correct that the results are sensitive to the analyzing wavelet, but we believe that a complete exposition of this point is beyond the scope of this paper because the Morlet wavelet is suitable for most geophysical applications; discussion of other wavelets in this context is mainly of mathematical interest. The following brief discussion of the sensitivity of the results to the chosen analyzing wavelet has been included in the summary section (Sect. 7):

“It is noted that the geometric test was only applied to patches arising from the convolution of the Morlet wavelet with a time series. The results presented in this paper are not valid for wavelet power spectra obtained using other analyzing wavelets, the reason for which is that each wavelet function has different time- and scale-localization properties that inevitably impact the geometry of patches. For example, patches found in the wavelet power spectrum obtained using a Paul wavelet are elongated in the scale direction relative to those obtained using a Morlet wavelet with $\omega_0 = 6$, resulting in nearby patches at different scales merging together. The merging of patches at different scales will alter their geometry with respect to the relatively thin (in scale) patches obtained using the Morlet wavelet.”

(2) What is the purpose of using the NAO time series as an example to assist with describing and testing the new methods? After it is introduced, it is largely dismissed as being a poor choice for this task and the focus shifts to the Nino 3.4 index.

The idea behind using the NAO index was to show, using the new methods, that the NAO is a stochastic process. Climate implications have been added to the paper to better motivate the use of the NAO in the application of the methods. Feldstein (2002), for example, found the NAO to be consistent with a first-order Markov process with a typical lifetime of 7 to 10 days. Hanna et al. (2014), as an another example, found that the variability of the NAO has increased but the results from geometric test suggests that such changers have been stochastic in nature.

(3) The Cone of Influence (COI) is referenced in the figure captions, but not described in the text. For pointwise significance, identifications are independent, so pointwise significance outside the COI can be ignored in the same way that wavelet power can be ignored outside the COI. It seems like this might not be true for geometric and topological methods. Therefore, are topological and geometric tests sensitive to edge effects (i.e., can edge effects influence significance even in regions where power is not influenced by edge effects)? If so, please provide more information about the importance of the COI and how it might influence results using the proposed methods.

The definition of the COI has been added on Page 1335 line 16. The cone of influence is the region of the wavelet spectrum in which edge effects become important. The areawise, geometric, and topological methods are all sensitive to edge effects. The effect of the COI on a patch located outside the COI is to shrink the patch, as the wavelet power and thus the significance associated with the patch are reduced. A paragraph discussing the impact of the COI on the results of the geometric test has been added after line 10 Page 1342. The paragraph reads as follows:

“Another situation that may arise in practice is the application of the geometric test to patches located both inside and outside the COI. In the case of the pointwise significance test, the edge effects only influence those wavelet power coefficients that lie inside the COI; however, for the geometric test, the significance of the entire patch will be impacted even if the patch only partially lies inside the COI. The reason is that the COI will act to decrease the size of significance patches through the reduction of wavelet power in the COI and subsequently the total area of the patch. One should thus be cautious when interpreting the results of the geometric test for patches near the COI.”

(4) Significance is determined by the 90% confidence level for the areawise and geometric tests, but 95% is used for the pointwise test. What is the reason for this inconsistency?

Throughout the paper we now use the 95% confidence level for the pointwise and areawise tests. The false discovery rate of the geometric test is controlled at the 5% level

(5) Throughout the text and captions, Figures 1, 2, 6, and 7 are described as "wavelet power spectra", but wavelet power is not shown in the figures. The authors should find another way to describe the contents of the figures or include contours of wavelet power.

"Wavelet Power Spectra" are now referred to as the significance of wavelet power throughout figure captions and text.

(6) There is no discussion about how "holes" are identified and I don't feel that there is enough information for future work to adopt the method. Minimally, holes should be defined quantitatively somehow, but it might also be helpful to describe how an algorithm could be developed to identify them. I am further confused because I cannot see all the holes that are identified in Figure 6 and 7, in part perhaps because the wavelet power is not shown, and in part because the significance patch does not completely encircle them.

A formal definition of a hole has been added on page 1344 line 19. Moreover, a discussion of how a hole is calculated is also provided after the definition is given. The following text has been added to include the definition of a hole:
"A more formal definition of a hole will require some notions from topology. Let $I = [0,1]$ be the close unit interval. Then a path from a point a to a point b in a significance patch P is a continuous function $f : I \rightarrow P$ with $f(0) = a$ and $f(1) = b$, where in the case that $f(0) = f(1) = c$ the path is said to be closed (Hatcher, 2002). Note that a point is a special kind of closed path called the constant path. A patch will be said to contain a hole if there exists a path in the significance patch such that it cannot be continuously deformed into a point, where the feature obstructing the path from such a deformation is a hole. The definition is consistent with notions of simply-connectedness in topology (Hatcher, 2002). Figure 4 shows an example of a closed path (blue curve) in a patch that cannot be contracted to a point because it surrounds a hole located in the patch."

The calculation of a hole is discussed in the following paragraph:
"For a patch with a hole, there will be two boundaries, an external boundary and an internal boundary representing the boundary between the hole and the patch. Thus, if a patch contains an internal boundary or contour it will contain a hole, whereas a patch without a hole will contain no internal contours. In practical applications, the existence of a hole can be determined by orienting external contours in the clockwise direction and internal contours in the counter-clockwise direction, a procedure automatically implemented using the standard Matlab contour function. The number of counter-clockwise oriented contours is thus the number of holes in the wavelet power spectrum at a given pointwise significance level."
Figures 6 and 7 have been changed so that the reader can identify the location of the holes at different pointwise significance levels. Table 1 has also been changed to better illustrate the power of the topological method in identifying significant wavelet power coefficients. The discussion in Sect. 5 has been changed to reflect the changes in Table 1 and Figs. 6 and 7 (now Figs. 8 and 9).

Specific Comments

(1) S1333L20: To better orient the reader, can you please provide a sentence or two that describes the main problems with pointwise testing?

Added on page 1333 line 21 is a brief example of what would happen if the pointwise significance test was applied to a wavelet power spectrum with a large number of wavelet power coefficients. The example will better orient the reader.

(2) S1335L16-18: This sentence should reference Figure 1.

Figure 1 has been referenced on page 1335 line 16.

(3) S1335L18-20: This sentence should reference Figure 2. The sentence states that periods from 16 to 64 months are significant, but Figure 2b only goes from 1 to 7 months. I suspect that the axis is actually in years or that the values are j not sj. However this is resolved it would be good to maintain consistency between Figs. 1 and 2.

Figure 2 has been referenced on page 1335 line 18. The scale axis has been corrected (should be months and labeled in months).

(4) Section 3.1: Please define the term "patch". Section 3.1 is good place to define the term similarly to how it is defined in the captions to Figs. 1 and 2, but the introduction might be a good place too.

The definition of a patch was added to page 1333 line 23.

(5) S1336L5: Can this sentence be rearranged to define a and b at the beginning? Also, is it necessary to use b and a? Does b = t=t appendix A and does a = s = a appendix A ?

The notation has been changed from *b* to *t* and from *a* to *s* throughout Sect. 2 and in Appendix A.

(6) S1336L15-17: This sentence is not quite clear. If a kernel fits within the patch is the entire continuous patch interpreted as significant or only the points that fall within the kernel?

Two sentences have been added on page 1336 line 16 to clarify that only points within the kernel should be deemed significant.

(7) Section 4.1: It would be helpful to lead this section (or alternatively close the previous section) with a sentence that reminds the reader of the objectives of the developing a new geometric test to improve upon the areawise test.

A short paragraph has been inserted in the beginning of Sect. 4.1 to explicitly state the objectives of the test and to motivate the reader before the geometric test is developed. The paragraph reads as follows:

“A disadvantage of the areawise test is the complexity of the α_{aw} calculation, which involves a root-finding algorithm. It is therefore desirable to construct an alternative test whose significance level is easy to calculate, readily allowing the following: (1) the application of the test to patches at various pointwise significance levels; (2) the adjustments of the significance level of the test; (3) the application of the test to wavelet power spectra obtained using other analyzing wavelets; and (4) the implementation of *p*–value adjustment procedures to control the family-wise error rates and false discovery rates.”

(8) S1337L15 & Eq. 6: Please define j. Shouldn’t it be tn and sj instead of ti and si since s and t are independent (n need not equal j) and also have different maximum values (i.e., J ≠N)? Thus, pn,j not pi?

sj has been defined on page 1337 line 15. The *j* now refers to the *j*th scale value in the set of scales determined by the equation inserted after Eq. (6). Notation has been changed to ensure consistency among the different indices such as *i* and *j*.

(9) S1338 Equations: Again, I’m confused about i. I think it is used appropriately for t, but the index of s is an entirely different coordinate than that of t. additionally, n is defined as the index of time (1 to N) in Eq. (2). It seems to be redefined here.

Yes, the indices are not used appropriately for *s*. To remedy the problem, notation has been changed throughout Sect. 4 so that the indices are consistent. Instead of redefining *n*, *m* has been used to denote the number of vertices of the polygon for Eqs. (7), (8), and (9) (now Eqs. (11), (12), and (13)).

(10) S1340L24: The p-value here is not the same as p in S1337L15, yes?

The *p* on page 1337 line 15 is not same as the *p*-value on page 1340 line 24. A change of notation has been made to reflect that on page 1337 line 15.

(11) S1346L18: Is “top panel” actually the bottom panel (Fig. 6c)?

“Top Panel” has been changed to “Fig. 6c” on page 146 line 18.

(12) Captions for Figures 1 and 2: It might be helpful to call out relevant sections throughout the captions, as was done for Fig. 2a.

The reader is now referred to specific sections in the text in the captions for Figs. 1 and 2 (now Figs. 3 and 4).

(13) Caption for Figure 3: Please clarify in the caption that the reproducing kernel is associated with the areawise test and not the geometric test, but is shown for reference.

A clarification sentence has been added in the caption of Fig. 3 (now Fig. 5), indicating that the reproducing kernel is for the areawise test.

References added

Baxandall, P. and Liebeck, H.: Vector Calculus, Dover Publications, INC., Mineloa, New York, 550, 2008.Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Power Approach to Multiple Testing, J. Roy. Stat. Soc., 57, 289-300, 1995.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Power Approach to Multiple Testing, J. Roy. Stat. Soc., 57, 289-300, 1995.

Benjamini, Y. and Yekutieli, D.: The Control of the False Discovery Rate in Multiple Testing under Dependency, Ann. Stat., 29, 1165–1188, 2001.

Feldstein, S. B.: The Time Scale, Power Spectra, and Climate Noise Properties of Teleconnection Patterns, J. Climatol., 13, 4430-4440, 2000.

Hasselmann, K.: Stochastic Climate Models Part I. Theory, Tellus. , 28, 473-485, 1976.

Hatcher, A.: Algebraic Topology, Cambridge University Press, New York, 544, 2001.

Hanna, E., Cropper, T. E., Jones, P. D., Scaife, A. A., and Allan, R.: Recent seasonal asymmetric changes in the NAO (a marked summer decline and increased winter variability) and associated changes in the AO and Greenland Blocking Index, Int. J. Climatol., 2014.

Kay, S. M.: Modern Spectral Estimation: Theory and Application, Prentice Hall, Englewood Cliffs, NJ, 560, 1988.

Wilks, D. S.: On “Field Significance” and the False Discovery Rate, J. Appl. Meteor. Climatol., 45, 1181-1189, 2006.

Worsby, F. M., Duckham, M.: GIS: A Computing Perspective, CRC Press, Boca Raton, FL, 448, 2004.