A New Statistical Approach for Quantifying Change in Series of Retinal and Optic Nerve Head Topography Images

Andrew J. Patterson, David F. Garway-Heath, Nicolas G. Strouthidis, and David P. Crabb

PURPOSE. To describe and evaluate new statistical techniques for detecting topographic changes in series of retinal and optic nerve head images acquired by scanning laser tomography (Heidelberg Retinal Tomograph [HRT]; Heidelberg Engineering, Heidelberg, Germany).

METHODS. Proven quantitative techniques, collectively referred to as statistic image mapping (SIM), are widely used in neuroimaging. These techniques are applied to HRT images. A pixel-by-pixel analysis of topographic height over time yields a statistic image that is generated by using permutation testing, derives significance limits for change wholly from the patient’s own data, and removes the need for reference data sets. These novel techniques were compared to the Topographic Change Analysis (TCA super-pixel analysis) available in the current HRT software, by means of an extensive series of computer experiments. The SIM and TCA techniques were further tested and compared to linear regression of rim area (RA) against time, in real longitudinal HRT series of eyes of 20 normal subjects and 30 ocular hypertensive (OHT) patients that were known to have converted to glaucoma, on the basis of visual field criteria.

RESULTS. Computer simulation indicated that SIM has better diagnostic precision at detecting change. In the real longitudinal series, SIM flagged false-positive structural progression in two (10%) of normal subjects, whereas TCA identified three (15%), and linear regression of RA against time identified two (10%). SIM identified 22 (73%) of the OHT converters as having structural progression, whereas the TCA and linear regression of RA against time each identified 16 (53%) over the course of the follow-up.

CONCLUSIONS. SIM has better diagnostic precision in detecting change in series of HRT images than compared to current quantitative techniques. The clinical utility of these techniques will be established on further longitudinal data sets. (Invest Ophthalmol Vis Sci. 2005;46:1659–1667) DOI:10.1167/iovs.04-0953

Confocal scanning laser tomography yields reproducible, three-dimensional (3-D) images of the posterior segment of the eye. This imaging technology, described in detail elsewhere, and typified by the commercially available Heidelberg Retinal Tomograph (HRT; Heidelberg Engineering, Heidelberg, Germany) is widely used in the assessment of the glaucomatous optic nerve head (ONH). Quantitative assessment of these topographic images can separate glaucomatous and normal eyes with generally high levels of diagnostic precision, with some evidence that this can be done in cases where measurable standard perimetric defects at the time of testing. The real promise of this technology probably lies in objectively measuring progressive structural damage, or stability, in patients being observed over time, which is possible, because the local height measurements at each of the pixels of a topography image are sufficiently reproducible.

An alternative approach, devised by Chauhan et al., considers change over time at the level of groups of pixels within the image: the Topographic Change Analysis (TCA). Now included in the HRT software, this technique divides the image into a 64 × 64-superpixel array. (Each superpixel is 4 × 4 pixels, thus containing 16 pixels.) Change in topographic height in superpixels is quantified with a statistical method comparing a set of baseline images to the most recent follow-up images.

In neuroimaging, positron emission tomography (PET) or magnetic resonance imaging (MRI) scans yield a sequence of 3-D images of the subject’s brain from which the temporal and spatial characteristics of neuronal activity can be deduced. In the case of MRI, for example, this is done by measuring changes in cerebral blood oxygenation related to brain activity. The images are complex and high-dimensional, typically containing as many as 100,000 measured volume elements or voxels (3-D pixels). Consequently, the neuroimaging research community has been forced to develop an extensive suite of techniques to register, align, process, and analyze arrays of imaging data. We propose to exploit part of this catalog of proven methods, specifically the techniques collectively referred to as statistic image mapping (SIM) which are used for determining areas of activity and change in series of MRI- and PET-type images, by applying them to series of retinal and optic nerve head topography images. In particular, we use a non-parametric version of these techniques. These are intuitive to understand, and assessment of change in the image is based solely on the subject’s own data and within-subject image variability, rather than any a priori information or patient population characteristics.
The purpose of this study is to describe and apply SIM techniques to HRT images. We also evaluated the performance of this new statistical approach by comparing it to the TCA method currently made available on the HRT software. We did this by means of an extensive series of computer simulation experiments that used a novel technique for generating simulated series of stable and progressing “virtual patient” HRT images, in which noise, typical of the misalignment inherent in serial topographical images, was mimicked. In addition, we applied SIM techniques to longitudinal sets of real HRT data from patients and normal subjects, and made comparisons with the TCA method and trend analysis of HRT rim area measurements.

METHODS

The novel quantitative techniques described in these methods are applicable to several of the retina imaging modalities. In this work, we considered series of topography images. The principle of HRT image acquisition is described fully elsewhere.1,2 Briefly, the HRT uses a low-intensity diode laser and obtains 32 equally spaced confocal sections, centered on the optic disc and perpendicular to the optical axis of the eye. The 32 sections, each having an area of 256 × 256 pixels, are aligned to compensate for lateral eye movements during acquisition. A 3-D reconstruction of the image area is obtained by calculating the positions of maximum reflectivity along the z-axis, providing an image with discrete topographic height values at 65,536 (256 × 256) pixels. Typically, at each clinical visit, multiple scans are obtained in a subject, usually three.19 A known characteristic of patients with progressing glaucoma is increasing ONH excavation and nerve fiber layer thinning with time, often referred to as structural progression.20 –22 The ideal clinical tool for assessing a longitudinal set of these HRT images would highlight this structural progression as localized areas of the ONH that are changing beyond the natural within-test and between-test variation in the images.

Statistic Image Mapping

The methods take advantage of proven statistical techniques that have been developed to analyze series of MRI and PET images. These
analyses usually proceed by computing a statistic at the pixel level (or the voxel in the case of MRI and PET images), indicating evidence for the observed effect of interest, and resulting in an "image of statistics," or a statistic image map. The entire statistic image must be assessed for significant effects, using a method that accounts for the inherent multiplicity involved in testing all the pixels at once. This analysis can be accomplished in a classic parametric statistical framework, but we used an alternative that is based on permutation testing. The latter is conceptually simple, does not rely on any theoretical probability models that may or may not be appropriate for the HRT data, deals with the multiple comparison problem of testing a vast image space, and critically derives significance limits for change based only on the individual patient’s series of data. These specific techniques and the mathematics that underpin them, as applied to PET and MRI data, are extensively described elsewhere. What follows is a description of three component parts of this approach and how we applied them to a series of HRT data, with the descriptive order chosen to facilitate the explanation of the methods, rather than to replicate the computational sequence, details of which can be found in the Appendix.

Permutation Testing at Individual Pixels. Consider that three HRT images, at each patient visit, are acquired at regular intervals during a clinical follow-up. After registration of the image series, the topographic height at each individual pixel is considered in turn. Visually, this can be done by plotting the topographic height at each pixel as a time series (Fig. 1). Next, a suitable statistic is derived for summarizing the change, or stability, of the topographic height at that pixel over time: the line of best fit (slope) derived from ordinary least-squares regression. The standard error (SE) of this slope gives an indication of how well the data fit the linear trend, with relatively high values indicating a poor fit or a noisy series of observation. Our test statistic at each pixel is simply the absolute value of the slope divided by the SE. A relatively large test statistic would be evidence of clear linear change of topographic height at that pixel. This process is performed at all the pixels, and the patient’s series of data is reduced to a statistic image—no longer a physiological image, but a 256 × 256-pixel map of statistics summarizing change within the image. The next step is to determine whether the observed test statistic at each pixel is unusual, or more extreme, than would be expected by chance. This testing of the significance of the test statistic is not completed in the conventional manner, by considering the observed test statistic as a random variable from a probability model, but uses a permutation test. We randomly shuffle, or relabel, the order of the observed data

![Figure 2. Simulated change: active (changing) pixels with negative slopes are shaded, with the largest cluster in black. We show the observed statistic image and two of the 1000 permutations. The distribution of maximum cluster sizes was created by recording the largest cluster of active pixels in the statistical image for each unique permutation. In this case, one cluster in the observed statistic image (●), generated by simulating a progressing patient, is very unusual (P < 0.01). Therefore, the glaucoma in the virtual patient is classified as progressing.](https://iovs.arvojournals.org/pdfaccess.ashx?url=data/journals/iovs/933231/)
and recalculate the test statistic for all possible permutations of the order of images. If we let \( N \) denote the number of all possible labelings, \( t_i \) the statistic corresponding to labeling \( i \), then the set of \( t_i \) for all possible relabeling constitutes the permutation distribution. For example, there would be 369,600 (12!/(3! × 3! × 3! × 3!)) of these in a series of four clinical visits with three scans at each visit (see Appendix for more details of this calculation). We then assume that all the \( t_i \) are equally likely and determine the significance of the observed test statistic by counting the proportion of the permutation distribution as, or more, extreme than our observed value, giving us our \( P \)-value. If \( P \) is, for example, <5% we label the pixel as active or changing. (We therefore assume that images acquired at the same visit are no more correlated than images acquired between visits. Previous work on the influence of time separation on interimage topographic variability support the intuition behind this approach.32) This permutation test is performed pixel by pixel, and the statistic image becomes “thresholded” at the 5% level, with pixels flagged if they are significant (Fig. 1).

In practice, a sample of 1000 randomizations (drawn without replacement from all the possible labelings) are used to generate the permutations distribution.33,34 This cases the computation burden but still allows for a statistically exact test at standard levels of significance testing. (Larger samples would be needed to evaluate \( P < 0.1 \%).

**Permutation Testing for Thresholded Clusters.** Thus far, we have considered a separate analysis at each of the 65,356 pixels within the HRT image, with no attempt to take into account the multiplicity of testing. Statisticians refer to this as the **multiple comparisons problem** and the construction of a corrective analysis for high-dimensional MRI and PET data has occupied many researchers, with ideas ranging from the simple use of Bonferroni adjustments to other mathematical solutions (see Nicholls and Holmes35). In this work we again exploited an intuitive approach, once more using a permutation test, which has been successfully applied to sequences of MRI and PET images, and outperforms other approaches when there are few images involved (or experiments with low degrees of freedom).

Once we had thresholded the statistic image pixel by pixel (Fig. 2), we were left with an image that contained clusters of contiguous, significant, or active pixels. We then noted the size of the largest cluster in the observed image. To ascertain whether the spatial extent of the clusters in the observed image was unusually large by chance alone, we set about shuffling the images again, recomputed the statistic image, calculated the cluster sizes, and recorded the size of the largest cluster. (In fact, the shuffling for the pixel-by-pixel analysis and the cluster testing is all accomplished in one “sweep” in the computational algorithm.) This procedure was repeated, to generate a permutation distribution of the maximum cluster size (Fig. 2). Hence, we assessed the significance of the observed result by considering only the patient’s data, and no knowledge of the probabilistic behavior of the topographic heights at image pixels was required. This is particularly useful because of the **spatial correlation** that exists within the image (i.e., the topographic height of neighboring pixels is more similar in some parts of the image than in others) and, in part, this cluster testing accounts for this. The threshold value generated to determine progression was unique for each patient and varied depending on the patient’s signal-to-noise ratio. The criteria for progression included only depressed or more significant superpixels bound within the contour line for the optic disc.

**Preprocessing: The Pseudo Test Statistic.** A prerequisite for any pixel-by-pixel analysis of a series of images is that any given pixel represents precisely the same anatomic region across the series. Even with the HRT software alignment procedures, such representation is a considerable leap of faith. Spatial smoothing improves signal to noise across the series, and the TCA superpixel method (described later) is a simple, but workable, example of this—essentially, with averaged topographic height effects being considered within a 4 × 4-pixel region. Again, a proven solution to this problem is available that involves the generation of a **pseudo test statistic**. Rather than divide the 256 × 256 matrix of individual slope values by the 256 × 256 matrix of individual SEs to yield the test statistic, the slope values are divided by a spatially filtered SE. The latter is the matrix of SEs smoothed with a weighted Gaussian kernel. Thus, a pseudo-test-statistic image is formed by dividing the slope matrix by the smoothed SE matrix. Hence, all the analyses, including the permutation cluster testing, proceeded with these pseudo test statistics. In essence, the noise from the variance image (the matrix of SEs) is smoothed, but not the signal. Statistic image maps constructed with smoothed variance estimates have been shown to improve the power of the approach substantially and can only be used in the nonparametric or permutation setting outlined herein.26–29 We therefore included this in our approach.

**Evaluation of the New Approach.** We compared the performance of our new approach against the TCA method in a computer virtual-patient simulation. The TCA method was replicated in consultation with the authors of the technique (David Hamilton, Department of Mathematics and Statistics, Dalhousie University, Canada, private communication, 2004). In short, the 256 × 256-pixel array from each topographical image is divided into a 64 × 64-superpixel array. An ANOVA is conducted to measure the extent of a constant shift in the topographic height over all 16 pixels within each superpixel from the set of images (three replicates at baseline) to another (three replicates at the follow-up visit), but the method also considers an interaction term allowing for different changes at different pixels within each superpixel. The significance of change at each superpixel is evaluated using an F distribution, in which the degrees of freedom are adjusted by a correction. It is worth highlighting that this adjustment is used for analysis within a superpixel and does not correct for the spatial correlation or multiple testing across the whole image. The criteria for change implemented in a recent publication were applied exactly.33 Any virtual patient who showed a cluster of 20 or more significant superpixels bound within the contour line for the optic disc, where the topographic change compared with baseline occurred in three consecutive sets of follow up images, was considered to have confirmed progression of glaucoma.

**Computer Simulation.** Simulations of topographic images have been used previously to test the ability of new techniques to distinguish glaucomatous from normal optic discs.35,36 In this study, simulation experiments were designed to quantify the specificity and sensitivity of our technique and the TCA superpixel method. Subjects with stable or unstable images were simulated. Those with unstable images had gradual and episodic change applied to a region of the neuroretinal rim. Each subject comprised a longitudinal series of 30 images: 10 sets of 3 images (baseline set, and nine follow-up visits).

We simulated each stable series by using 30 identical copies of an HRT topographic image (replicating 10 visits with three scans per visit) and then applied noise to each image. Progressing patients’ series with gradual change (linear) were simulated by creating 30 identical images and applying a cumulative decay of 5 \( \mu \)m per visit to a cluster of 480 pixels to the neuroretinal rim. Progressing patients series with episodic change (sudden) were simulated by applying a height decay of 50 \( \mu \)m to the cluster at a randomly selected visit between visits 2 and 10, inclusive.

Between-image variability had two elements: “misalignment noise” and background noise. Misalignment noise was simulated by applying a series of transformations to each image in translations (\( x', y' \), and \( z' \) and rotations about each axis \( (\alpha_{x'}, \alpha_{y'}, \text{and} \alpha_{z'}) \). Transformations are applied at a subpixel level, using bicubic interpolation algorithms.37 The magnitude of each of the six transformations was made unique in each simulation by using a random number sampled from a normal distribution, wherein the mean of the size of the transformation is set at zero and a variance is fixed for each transformation. To mimic background noise, Gaussian noise was added to each pixel with variance \( \nu \) and mean zero. (A proven nonbiased random-number generator was used to sample from a normal distribution.38) To replicate the
RESULTS

Computer Simulation

In the 300 stable virtual patients, under the conditions of these computer experiments, the TCA method flagged 16%, 17%, and 17% at MPHSD of 15, 25, and 35 μm, respectively, at some point in the follow-up series (false positives). These values were closer to 10% in the first half of the follow-up, but worsened as more visits were considered. SIM had much better specificity, with 6%, 5%, and 5% flagged at the different levels of noise (Fig. 3a). In the simulations of progressing patients, the TCA method identified progression at some point in follow-up in 95%, 31%, and 28% with linear change, and 82%, 47%, and 42% with episodic change, for the MPHSD of 15, 25, and 35 μm, respectively. SIM identified 100%, 68%, and 62% with linear change, and 86%, 57%, and 55% with episodic change (Figs. 3b–d). For these experiments, the TCA had slightly better or similar sensitivity than did SIM at detecting gradual (linear) change, up to about visit 6 or 7, with SIM outperforming TCA as more data became available. A similar pattern emerged when episodic loss was specified, but with equivalent sensitivity when the noise was low (MPHSD 15 μm).

Real Longitudinal HRT Series

The results are summarized in Table 1. Examples of the similarity and differences between the SIM, TCA, and RA results are illustrated in Figure 4. Cases 1 and 2 are both OHT converters. In case 1, both SIM and TCA confirmed progression at visit 4, and the linear regression of RA against time was statistically significant (P < 0.001). In case 2, SIM identified progression at visit 6, whereas the TCA did not detect progression at all, the linear regression of RA against time was reached slight statistical significance at visit 7 (P = 0.042).

Although our technique is computationally intensive, by developing the algorithms in a low-level programming language and designing the code to reduce function calls and variable passing, the computer burden is not prohibitive. Analysis of a patient having 10 visits (30 images with 3 scans per visit) took less than 3 minutes on our computer with a 3-GHz processor. Shorter series took less time to analyze, but even a very long series of patient records were manageable on a standard computer during a patient visit. Further improvements to the computer code are likely to reduce this time further.

DISCUSSION

Reproducible scanning laser tomography images of the ONH potentially present an objective method for measuring disease progression in glaucoma. Serial analysis (using trend analysis or statistical tests comparing baseline and follow up images) of topographic indices, such as cup-to-disk ratio, RA, and the like, derived after a disc margin contour line has been defined, have been typically used to quantify change. These methods may be subject to similar inadequacies associated with using the global indices to summarize progression in visual fields: chiefly loss of spatial information and poor sensitivity to identify the localized damage. In this work, we have presented and evaluated statistical procedures for the analysis of pixel level longitudinal data in structural HRT images by using existing techniques primarily developed for neuroimaging data.

The computer simulation and analysis of real longitudinal HRT data provided evidence that SIM is more powerful at detecting localized change than the TCA method and analysis of HRT rim area measurements against time. This result was achieved without the expense of more false positives. The
developers of the TCA method originally demonstrated a high level of sensitivity and specificity in detecting change in computer simulation experiments, but series of confirmation tests and a requirement for a certain cluster size were needed to produce similarly adequate levels of diagnostic precision in real longitudinal image data sets. This necessity is not surprising, as the original simulations centered on a single superpixel rather than results across the whole image. A statistical adjustment (the Satterthwaite correction) was used to correct for similarity (spatial correlation) of the topographic height within a superpixel, but no real account was made for the multiplicity of testing across the whole image. The empirical solution to the problem of multiplicity of testing included the requirement for clusters of pixels to be above a certain size, based on observed series of normal subjects. However, this empirical solution is based on observations of variability within a population and not on observed variability within a subject’s own data series. One possible reason our technique outperforms this analysis in our computer simulation, and in real longitudinal data, is that it inherently corrects for the multiple comparison problem. Handling this aspect of imaging data is one of the key features of the SIM approach.

At the center of our technique is the use of permutation testing, tailoring the analysis to the data itself without incorrectly assuming that topographic heights, across the whole image, follow the behavior of a random variable from a known probability distribution, or without reliance on some reference patient population database. Statistically speaking, permutation methods are known to be both flexible and exact. With the increased computational power now available, there seems to be no important argument against their preferred use in situations in which there may be arbitrary properties of the observed data that cannot be accounted for by a probability model. The most plausible explanation, however, for the better diagnostic precision of SIM in comparison to the TCA technique in these computer experiments is simply the use of the whole series of the data. The TCA method uses only the baseline images and three follow-up images. This may be reasonable when the follow-up is short, but when the available series of data lengths beyond four visits, it results in considerable data redundancy, as illustrated in Figure 3 when the difference between the two methods appears approximately halfway through the potential follow-up of 10 visits. It is also interesting to note that there is no discernible difference between the power of the methods when episodic or sudden loss is specified (Fig. 3). This aspect of the results is reassuring, because our choice of the pixel-by-pixel test statistic is essentially a rate (trend) parameter, which might not be considered sensitive to detecting a sudden change. However, we have reported for threshold measurements in the visual field that linear regression adequately identifies sudden change, unless a series of data becomes very long. At the same time, there is a real advantage of using a rate parameter, as it may provide clinically interpretable information once the technique has identified a significant region of change. Of course, there is no firm evidence about structural loss in glaucoma being either gradual or sudden, but it seems the new technique described herein will be sufficiently sensitive to both types of deterioration.

### Table 1. Results of the Longitudinal HRT Series

<table>
<thead>
<tr>
<th></th>
<th>SIM</th>
<th>TCA</th>
<th>RA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Control subjects</td>
<td>2 (10)</td>
<td>3 (15)</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Converters</td>
<td>22 (73)</td>
<td>16 (53)</td>
<td>16 (53)</td>
</tr>
</tbody>
</table>

Data represent the number of eyes determined to be progressing with SIM, TCA, and linear regression of the RA against time, applied to a real longitudinal HRT series. Results are shown for 20 normal subjects and 30 patients with OHT whose eyes converted to a diagnosis of glaucoma, according to visual field criteria (converters).
This work serves an additional purpose in reporting the statistical image mapping techniques as an example of the catalog of proven methodology developed by the neuroimaging community that should be exploited by clinicians and scientists who are developing analysis techniques for retinal images. Moreover, the specific statistical techniques described in this work should not be restricted to quantifying glaucomatous structural progression, but could be used to identify features in topography images acquired to monitor other disease processes. A good example would be the recently proposed macula thickness maps for diabetic edema. Furthermore, these techniques should not be confined to one type of retinal imaging modality and could be applied, for example, to series of images acquired from optical coherence tomography and scanning laser polarimetry.

The computer simulation of series of HRT images reported in this work is the first of its kind, with previous computer simulations restricted to separating normal and glaucomatous. It is also novel because it imitates noise by replicating the repeatability of topographic height measurements by applying Gaussian and misalignment noise that typical remain in serial topographical images, even after they have been registered by the HRT software. Of course, an evidence base for the clinical validity of SIM, as applied to series of retinal images, can only be achieved in a study of a longitudinal cohort of patients and control subjects large enough to provide sufficient statistical power, but this was beyond the scope of the present study and is the subject of future work. In addition, we see the main value of these techniques as being a way of providing the clinician with a much needed, reliable method of visualizing, quantifying, and assessing rates of glaucomatous change in small localized areas in series of retinal images, rather than binary progression or stable classifications that rely on topographic summary parameters. The permutation test provides great advantages over the parametric approach, because it always works, given a short series of data, and makes no assumptions about the underlying distribution of the slopes or the topographic heights. In conclusion, we have demonstrated...
the application of a new set of techniques to detect change in series of retinal and ONH topography images. The evidence provided by the results from computer simulation experiments and the real longitudinal HRT data and their proven use in the field for which they were originally developed suggest they can be a useful clinical tool.

Acknowledgments

The authors thank Fred Fitzke (Institute of Ophthalmology, London) for advice on this work, David Hamilton (Mathematics and Statistics, Dalhousie University, Halifax, Canada) for his contribution in replicating the TCA superpixel technique, and Paul Artes (Ophthalmology and Visual Sciences, Dalhousie University, Halifax, Canada) for technical help with the TCA superpixel analysis.

References

APPENDIX

Computational Aspects of the Statistic Image Mapping Approach

Permutation Testing at Individual Pixels. The only limiting factor of permutation tests is the number of combinations necessary for testing a probability limit. In practice, a sample of 1000 randomizations (drawn without replacement from all the possible labelings) are used to generate the permutations distribution. These sample data ease the computation burden but still allow for a statistically exact result at standard levels of significance testing. (Larger samples would be needed to evaluate $P < 0.01$.)

The number of possible unique permutations is expressed as

$$\frac{(s \times n)!}{(s!)^n}$$

where $s$ is the number of scans per visit and $n$ is the number of visits. For example, with four visits and three scans per visit, there are

$$\frac{(3 \times 4)!}{(3!)^4} = 369600$$

unique permutations.

The following steps represent the computational paradigm to compute a permutation distribution and test statistical significance:

1. At each pixel$(i,j)$ calculate by least-squares linear regression the slope $b(i,j)$, SE $se(i,j)$, and absolute test statistic $t(i,j)$ of time (dependent variable) against topographic height (independent variable).
2. Shuffle the order of the dependent variable (time) to generate a unique permutation and recalculate $b$, $se$, and $t$.

Quantifying Change in Topography Images

3. Repeat step 2 1000 times, calculating a unique permutation each time. As each permutation must be unique, the algorithm must perform sampling without replacement.
4. We reject the null hypothesis at a significance level of $P < 0.05$. Thus, for the mechanics of the permutation distribution, we reject the null hypothesis if the observed test statistic is greater than or equal to the 95th percentile of the permutation distribution. Therefore, sort the array of test statistic $t$ produced at each pixel$(i,j)$ in ascending order, and test whether the absolute observed test statistic is equal to or more than the 950th ($0.95 \times 1000$) value of $t$. Note that we retain the sign of the observed test statistic to indicate the direction of change—that is, a negative sign indicates a depression in topographic height values over time, whereas a positive sign indicates an elevation in topographic height over time.

Preprocessing: The Pseudo Test Statistic. The pseudo test statistic $t_{stat}(i,j)$ is calculated by dividing slope $b(i,j)$ with a smoothed SE $se(i,j)$. The smoothed SE is calculated by convolving the SE set$(i,j)$ with a Gaussian kernel. We used a square Gaussian kernel of symmetrical full width at half maximum of 11 and size $17 \times 17$ to smooth the SE $se(i,j)$. The pseudo test statistic is calculated for the observed case and for each unique permutation.

Permutation Testing for Thresholded Clusters. The following paradigm is a programming methodology for thresholding clusters:

1. Compute the observed pseudo test statistic.
2. Compute the pseudo test statistic for each unique permutation.
3. Compute an observed statistical image $s(i,j)$ by setting $s(i,j)$ to equal active_depressed or active_elevated, if the observed absolute pseudo test statistic is within or higher than the 95th percentile of the permutation distribution of the absolute pseudo test statistic at pixel$(i,j)$. Record the size of the maximum depressed and elevated clusters within the observed statistical image, bound within the contour line. An active pixel within a statistical image $s(i,j)$ is defined as part of a continuous cluster if one of the eight pixels within its neighborhood is also active (i.e., 8-connectivity).
4. Compute a statistic image at each of the 1000 unique permutations. Record the size of the maximum depressed and elevated clusters for each unique permutation, bound within the contour line.
5. Sort the array of maximum clusters into ascending order.
6. A depressed or elevated cluster (or clusters) within the observed statistic image is defined as statistically significant if it (or they) are larger than the 99th percentile of the maximum depressed and elevated cluster distributions. Progression is defined if a depressed cluster is larger than the 99th percentile of the maximum depressed distribution.