


## RESEARCH ARTICLE

# Evaluation of thresholding methods for activation likelihood estimation meta-analysis via large-scale simulations

Lennart Frahm<sup>1,2</sup>  | Edna C. Cieslik<sup>2,3</sup> | Felix Hoffstaedter<sup>2,3</sup> |  
Theodore D. Satterthwaite<sup>4,5</sup> | Peter T. Fox<sup>6</sup> | Robert Langner<sup>2,3</sup> |  
Simon B. Eickhoff<sup>2,3</sup>

<sup>1</sup>Department of Psychiatry, Psychotherapy and Psychosomatics, School of Medicine, RWTH Aachen University, Aachen, Germany

<sup>2</sup>Institute of Neuroscience and Medicine (INM7: Brain and Behavior), Research Centre Jülich, Jülich, Germany

<sup>3</sup>Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

<sup>4</sup>Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>5</sup>Penn Lifespan Informatics and Neuroimaging Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>6</sup>Research Imaging Institute, University of Texas Health Science Center, San Antonio, Texas, USA

## Correspondence

Lennart Frahm, Institute of Neuroscience and Medicine (INM7: Brain and Behavior), Research Centre Jülich, Wilhelm-Johnen-Straße, Jülich 52425, Germany.  
Email: [l.frahm@fz-juelich.de](mailto:l.frahm@fz-juelich.de)

## Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 269953372/GRK2150, EI 816/11-1; National Institute of Mental Health, Grant/Award Number: R01-MH074457; Jülich-Aachen Research Alliance (JARA); National Institute of Aging, Grant/Award Number: P30-AG066546

## Abstract

In recent neuroimaging studies, threshold-free cluster enhancement (TFCE) gained popularity as a sophisticated thresholding method for statistical inference. It was shown to feature higher sensitivity than the frequently used approach of controlling the cluster-level family-wise error (cFWE) and it does not require setting a cluster-forming threshold at voxel level. Here, we examined the applicability of TFCE to a widely used method for coordinate-based neuroimaging meta-analysis, Activation Likelihood Estimation (ALE), by means of large-scale simulations. We created over 200,000 artificial meta-analysis datasets by independently varying the total number of experiments included and the amount of spatial convergence across experiments. Next, we applied ALE to all datasets and compared the performance of TFCE to both voxel-level and cluster-level FWE correction approaches. All three multiple-comparison correction methods yielded valid results, with only about 5% of the significant clusters being based on spurious convergence, which corresponds to the nominal level the methods were controlling for. On average, TFCE's sensitivity was comparable to that of cFWE correction, but it was slightly worse for a subset of parameter combinations, even after TFCE parameter optimization. cFWE yielded the largest significant clusters, closely followed by TFCE, while voxel-level FWE correction yielded substantially smaller clusters, showcasing its high spatial specificity. Given that TFCE does not outperform the standard cFWE correction but is computationally much more expensive, we conclude that employing TFCE for ALE cannot be recommended to the general user.

## KEYWORDS

family-wise error, FWE, multiple comparison correction, neuroimaging meta-analysis, significance thresholding, threshold-free cluster enhancement cluster extent

Simon B. Eickhoff and Robert Langner contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

## 1 | INTRODUCTION

Task-based functional magnetic resonance imaging (fMRI) is a major approach in modern neuroimaging research. Measuring hemodynamic responses to experimental tasks allows researchers to map mental processes and functions to specific brain regions. While this approach enabled important insights into the functional organization of the brain, it comes with its own problems and pitfalls, which ultimately contribute to the current reproducibility crisis in science (Baker, 2016; Bossier et al., 2020; Turner et al., 2018). For instance, many task-fMRI studies are underpowered, featuring a large number of dependent variables and a comparatively low number of participants or observations. This limits the generalizability of findings, and any effects reported tend to overestimate the true population effect (Cremers et al., 2017). Further, there is great experimental flexibility in setting up studies that are supposed to investigate the same mental function coupled with analytical flexibility, which refers to the variability in data analysis workflows employed across the world to preprocess and analyze datasets. A recent study (Botvinik-Nezer et al., 2020) exemplified the impact of this latter aspect by showing that no two teams out of a pool of 70 independent research teams used identical processing pipelines for analyzing the same dataset with respect to the same 9 *ex-ante* hypotheses, leading to rather heterogeneous results.

A way to elegantly deal with the problems mentioned above are meta-analyses, offering a synthesis of findings across paradigms and processing pipelines and thereby allowing more conclusive inference, even if the average power level of the included studies was questionable (Botvinik-Nezer et al., 2020; Cremers et al., 2017; Eickhoff et al., 2016; Wager et al., 2007). For synthesizing the task-fMRI literature, coordinate-based meta-analyses (CBMA) are especially useful as they summarize findings via aggregating peak coordinates reported in standard stereotaxic space. Although this leads to a rather sparse representation of published results, including a lack of information on the strength of the evidence, it also allows CBMA to integrate the largest part of the literature because standardized peak coordinates are reported in almost all task-fMRI publications.

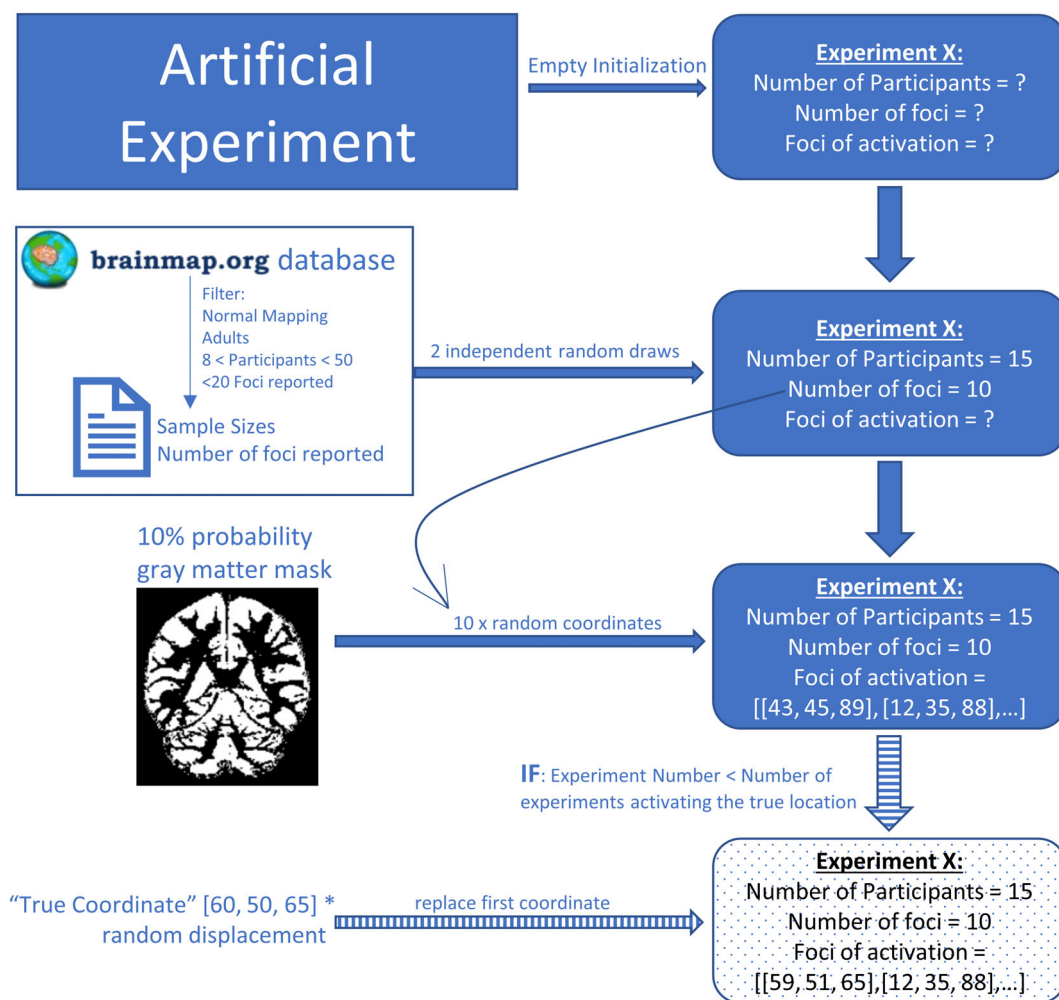
Activation-Likelihood Estimation (ALE) is one of the most common methods for CBMA, being used in over 100 published studies each year since 2011. It is part of the BrainMap software suite (<http://brainmap.org/ale>; Laird et al., 2009; Laird et al., 2005) as GingerALE and available for Python users as part of the NiMARE package (<https://nimare.readthedocs.io>). ALE enables the synthesis of findings across neuroimaging studies by modeling reported activation coordinates (foci) as probability distributions, thereby taking into account spatial uncertainty inherent to macroscale neuroimaging (Eickhoff et al., 2009). Originally introduced in 2002 (Turkeltaub et al., 2002), ALE has since undergone extensive evaluations and improvements in regard to statistical soundness (Acar et al., 2018; Eickhoff et al., 2009; Eickhoff et al., 2012; Turkeltaub et al., 2012). In 2016, Eickhoff et al. used a large-scale simulation set-up to evaluate four different approaches to multiple comparison correction in an ALE setting. In accordance with previous neuroimaging literature, they found both the reporting of uncorrected results and usage of false-discovery rate

correction to be inappropriate (Chumbley & Friston, 2009; Makin & de Xivry, 2019). Correcting for the family-wise error rate at voxel level (vFWE) was found to be valid but very conservative, while cluster-level family-wise error correction (cFWE) showed the best overall performance and was therefore recommended as the approach of choice.

In recent years threshold-free cluster enhancement (TFCE), a new technique to correct for alpha error inflation from multiple comparisons, has been rising in popularity (Han et al., 2019; Lett et al., 2017). TFCE enhances raw statistic images and aims to emphasize regions that show extended cluster-like activations, using spatial neighborhood information. It does so by calculating the product of cluster sizes at a large range of thresholds with the corresponding voxel-level activation heights and summing the results over all thresholds. In contrast to cFWE correction, this allows TFCE to work without a single, preset cluster-forming threshold. Even though both cluster extent and image height get raised to the power of two free parameters ( $E$  and  $H$ , respectively), theoretical analysis, and empirical results have shown that they can be seen as fixed values (Smith & Nichols, 2009). TFCE and cFWE also differ in that, by design, only rather large clusters survive cFWE correction, while TFCE can theoretically yield single significant voxels. This can be seen as an advantage for cFWE, allowing for easier interpretation, but also as a disadvantage, because spatially restricted but valid effects (e.g., in anatomically small structures) could be missed. Furthermore, TFCE performs much better when the assumption of signal stationarity has been violated (Salimi-Khorshidi et al., 2011; Spisak et al., 2019), which is a very common feature of neuroimaging where signal and noise levels are not consistent throughout the cortex. Most importantly, though, TFCE has been shown to be more sensitive than cFWE in numerous individual studies, especially when dealing with strong focal signals (Han et al., 2019; Li et al., 2017; Noble et al., 2020; Smith & Nichols, 2009; Spisak et al., 2019). These findings strongly suggest using TFCE as a means to correct for multiple comparisons in ALE settings as well. We, therefore, sought to comprehensively evaluate and compare TFCE performance to other standard corrections methods, namely vFWE and cFWE. We used a large-scale simulation approach, similar to the one employed by Eickhoff et al. (2016), running over ~200,000 distinct ALE analyses based on simulated datasets. We then applied TFCE, cFWE, and vFWE corrections to improve statistical inference and assessed the three methods via a multitude of evaluation outcomes, namely sensitivity, susceptibility to spurious convergence, and the resulting cluster size.

## 2 | METHODS

Our simulation study consisted of four sequential steps: (1) creating artificial datasets based on empirically derived parameters and random sampling, (2) performing ALE analysis on each dataset, (3) applying three different thresholding techniques to the results to correct for multiple comparisons, and (4) evaluating the performance of the thresholding techniques based on various outcome measures. As a note on terminology, we here will use the terms



**FIGURE 1** Simulation of an experiment. Two independent draws from the filtered Brainmap database were used to determine the sample size and the number of foci reported by the experiment. Next, we sampled the corresponding number of coordinates from a lenient gray-matter mask. Last, the first coordinate got replaced by the true coordinate multiplied with a displacement factor. This last step only happened if the experiment was an experiment activating the target location

“experiment” and “paper” when referring to a particular experimental contrast versus an entire published scientific work (possibly reporting results of multiple experiments), respectively, avoiding the ambiguous word “study,” which could be used to describe both concepts.

## 2.1 | Artificial datasets

The first step to run large-scale ALE simulations was to create artificial meta-analysis datasets, aiming to represent naturalistic datasets as well as possible. As in any real meta-analysis, the dataset was filled with different experiments, but here each experiment's characteristics and “results” were randomly generated following certain boundary conditions (Figure 1). Both the experiment's sample size and the number of foci reported for a given experiment were randomly sampled from the BrainMap database. This database contains meta-data and result coordinates of thousands of fMRI experiments and therefore

serves as a good approximation of the distribution of parameter values encountered when putting together a real-life meta-analysis dataset. We filtered the BrainMap database to only include normal mapping experiments conducted in adult samples. Additionally, we excluded all experiments with less than 8 or more than 50 participants and those for which more than 20 foci were reported. This left us with just over 6000 eligible experiments (number of participants per experiment:  $M = 15.56$ ,  $SD = 6.31$ ; number of reported foci per experiment:  $M = 10$ ,  $SD = 5.15$ ). Note that we only sampled the number of reported foci per experiment from the database and not the actual coordinates reported in the database. These were instead uniformly sampled from a relatively lenient gray-matter mask based on the ICBM tissue probability maps (>10% probability for gray matter; Evans et al., 1994).

The random generation of experiments was then used to fill artificial datasets. We systematically evaluated the effects of sample size and degree of convergence across datasets on ALE results by varying two parameters:

### 1. Number of experiments:

We ran ALE meta-analyses on datasets containing between 15 and 45 experiments, instead of 5–30 as in Eickhoff et al. (2016), because Eickhoff et al. (2016) showed that when using the current standard thresholding approach, meta-analyses of less than 17 experiments are easily underpowered and overly influenced by single experiments. Additionally, we ran analyses on three substantially greater dataset sizes ( $n = 75, 100, \text{ or } 150$ ) to evaluate the behavior of multiple-comparison correction techniques in larger-scale ALE analyses.

### 2. Number of “true activations”

The key question addressed by neuroimaging meta-analyses is how much agreement there is between studies regarding the activation of a given brain location. In simulations, of course, this true activation location is defined in advance and so is the number of simulated experiments that contain a focus at this ground-truth location. In our case we chose a voxel in the left motor cortex ( $-30/-26/58$  MNI space), purely based on later ease of visualization. To create a more realistic representation of true activation, the reported foci were not all located precisely at the same coordinate, but rather in the vicinity of the “true” location, displaced by virtue of a spread distribution devised and tested by Eickhoff et al. (2016). This spread distribution was based on the analysis of 15 hand-coded datasets and 105 datasets automatically extracted from the BrainMap database. Eickhoff et al. (2016) performed ALE analyses on all datasets, identified peaks in the resulting Z-maps and checked the distance of contributing foci to the corresponding peaks. Using this distribution of distances to displace the true activations, we ensured that the simulated data mirrored “real” data as well as possible. To independently assess the effect of the size of spread around the “true” location, we ran supplementary analyses in which the displacement was multiplied by a factor between 0.5 and 1.5 (number of studies = 30; experiments activating the target location = 2–10). In contrast to Eickhoff et al., we investigated a range from 0 to 10 experiments activating the target location (instead of 1–10). By including zero true convergence, we aimed to get a better estimate of the performance of the multiple-comparison correction methods when confronted with spurious activation only. Including scenarios of convergence across more than 10 studies was shown to be unnecessary by Eickhoff et al., as both cFWE and vFWE corrections achieved near-perfect detection rates already when 10 studies activated the simulated target location.

Varying these two parameters gave us a total of 341 combinations for which we created 500 datasets each to account for the random sampling in experiment generation. This resulted in 170,500 datasets being generated for the main analysis. Taken together with all supplementary analyses, which included another 84,400 datasets, we analyzed a total of 254,400 datasets.

## 2.2 | ALE and multiple-comparison correction

We ran ALE analyses on all datasets and tested for statistical significance using voxel- or cluster-level FWE (cluster-forming

threshold  $p < .001$ ) correction at  $p < .05$  as well as TFCE ( $E = 2, H = 0.5$ ) corrected at  $p < .05$ . All three methods are based on permutation testing to determine a threshold value as follows:

1. An empty version of the current dataset (i.e., featuring a given number of experiments, each with a given number of subjects and foci) was filled with random foci sampled from the ICBM gray-matter mask.
2. ALE was run on it.
3. We saved the highest ALE value (vFWE), the size of the biggest cluster with a cluster-forming threshold of  $p < .001$  (cFWE), and the highest TFCE value.
4. Steps 1–3 were repeated 10,000 times.

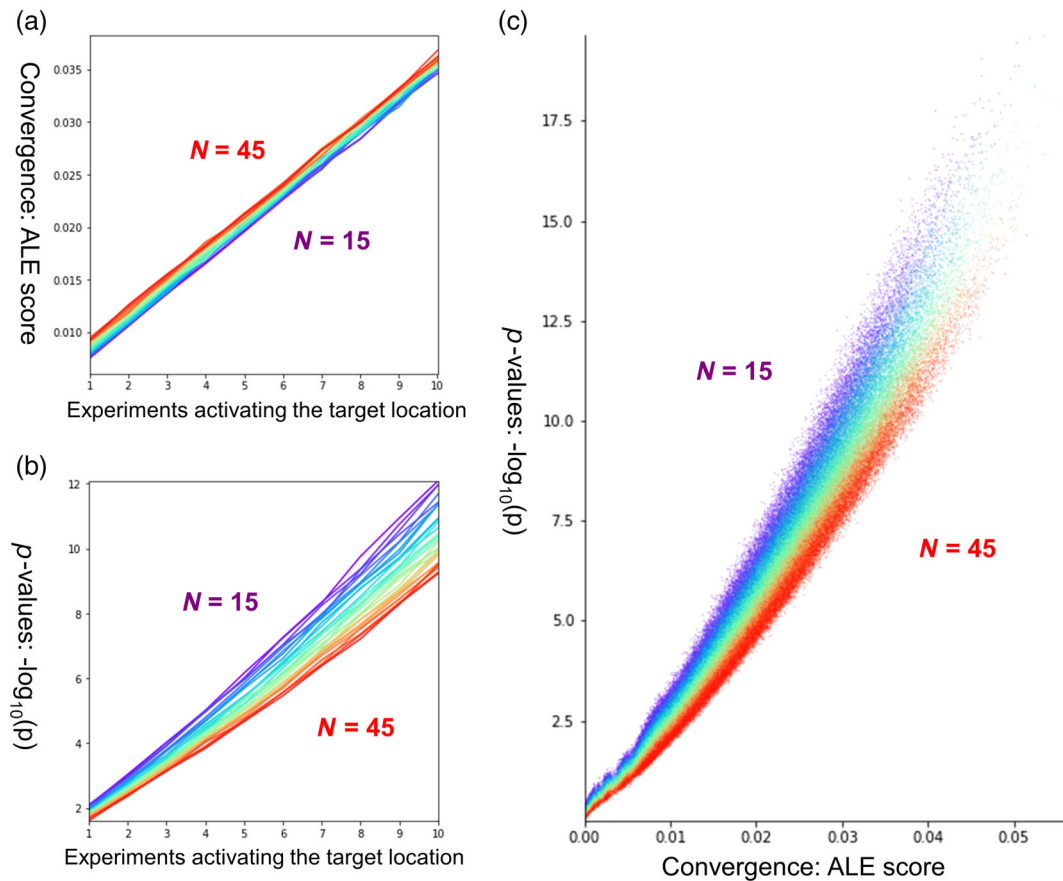
This gave us a distribution of maximum values under the assumption of spatial independence, from which we selected the value representing the 95th percentile as a significance threshold for the results of the original analysis. This way a family-wise error rate of 5% was established.

For datasets of  $n = 30$  studies, we performed additional exploratory analyses looking at TFCE parameter combinations between  $E = [1.8, 2.0, 2.2]$  and  $H = [0.3, 0.5, 0.7]$  to assess the impact of these settings on statistical inference.

All analyses were run with a new implementation of the ALE algorithm in Python (<https://github.com/LenFrahm/pyALE>), the results of which were confirmed to be identical to those obtained with the original version implemented in Matlab as used in Eickhoff et al. (2016). TFCE was independently implemented to extend the original ALE implementation, according to the description provided by Smith and Nichols (2009) and taking inspiration from a python-based implementation publicly available on GitHub (<https://github.com/Mouse-Imaging-Centre/minc-stuffs/blob/master/python/TFCE>).

## 2.3 | Outcome measures

We first compared the thresholding methods based on their sensitivity, which describes the method's power to find convergence when more than one study activates the target location. Sensitivity was quantified by the average rate of significant findings in a 4 mm radius around the “true location” over the 500 iterations of each parameter combination and should be high. The opposite outcome measure to sensitivity is the susceptibility to spurious convergence, which is measured by the average rate of significant clusters outside of a 4 mm radius around the “true location” over the 500 iterations of each parameter combination. Such clusters reflect the chance of incidental convergence under the known spatial independence of results, which should be as low as possible. The next outcome measure was cluster size, which is quantified by the average size of significant clusters that includes the location of true activation. Lastly, we looked at computational efficiency, measured by the time it takes for a single



**FIGURE 2** Behavior of ALE scores and the corresponding  $p$ -values under the different levels of the two simulation parameters (number of experiments and number of experiments activating the target location) and their 341 combinations. The total number of experiments included in the ALE analysis is color coded in a spectral sequence from 15 experiments (purple) to 45 experiments (red). (a) Average ALE-score (over 500 iterations) at the ground-truth location. ALE scores increased linearly as a function of the number of experiments activating the target location but also with the total number of experiments due to the increased chance of (positive) interference by noise foci. (b) Average  $p$ -value over 500 iterations at the ground-truth location.  $p$ -values decreased with a higher number of experiments activating the target location.  $p$ -values increased with the total number of experiments because of a right shift of the null-distribution. (c) ALE scores versus  $p$ -values at the ground-truth location for all 170,500 simulations. The more experiments are included in an ALE analysis, the more convergence (higher ALE score) was needed to obtain the same  $p$ -values.

null permutation to be calculated for each thresholding methods, averaged over 50 permutations for each of the 30 dataset sizes we looked at.

### 3 | RESULTS

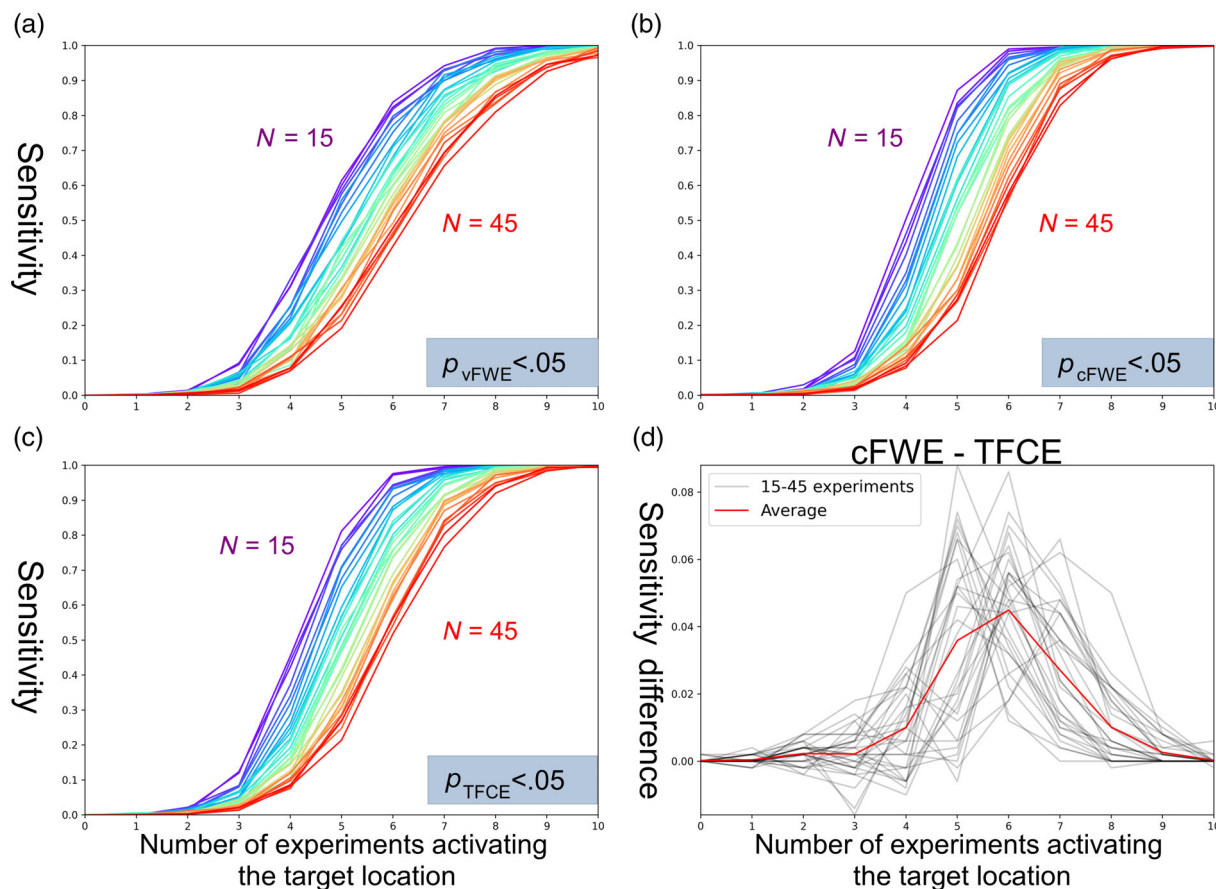
#### 3.1 | ALE scores and $p$ -values

We examined the maximum ALE score and  $p$ -value in a 4-mm radius around the chosen true location as a function of the total number of experiments and the number of experiments activating the target location (Figure 2). ALE scores increase with more experiments activating the target location, while  $p$ -values decrease. Even though we analyzed a parameter range that slightly differed from Eickhoff et al.'s (2016), we observed very similar patterns. Especially in the larger meta-analysis datasets (30–45 studies) included here, we

observed that the trends for ALE score and  $p$ -values described by Eickhoff et al. held up well.

#### 3.2 | Sensitivity

To find out if there is any added value in using TFCE, as compared to cFWE or vFWE correction, we examined the sensitivity of each method. As can be seen in Figure 3, cFWE correction performs best overall, closely followed by TFCE, whereas vFWE shows notably lower sensitivity than the other two. Upon closer examination of the difference between cFWE and TFCE corrections, it becomes clear that cFWE on average performs better between four and eight experiments activating the target location. The reason for this is that sensitivity is close to zero for  $n < 4$  experiments activating the target location and close to one for  $n > 8$  experiments activating the target location. The only range where methods can outperform each other is



**FIGURE 3** (a–c) Sensitivity of ALE when applying different multiple-comparison correction methods for statistical inference. The number of experiments activating the target location is represented on the x-axis, while each total number of experiments has its own curve in the graph following a spectral color sequence (15–purple; 45–red). The curves show the average sensitivity over the 500 iterations of each parameter combination. For all three methods, sensitivity increased in an approximately sigmoid fashion as a function of the number of experiments activating the target location. Additionally, having more experiments in the dataset required having more experiments activating the target location to achieve the same sensitivity. (d) Zooming in on the difference in sensitivity between cFWE correction and TFCE: The differences between individual dataset sizes are displayed in gray and the average over all dataset sizes in red. cFWE correction performed better on average, especially between 4 and 8 experiments activating the target location. There were a few dataset sizes in which TFCE has a slight sensitivity advantage at 3–4 experiments activating the target location.

in the rising part of the sigmoid. When looking at sensitivity for each dataset in this range individually cFWE and TFCE performed equally in 94% of datasets, cFWE outperformed in 4.3% of datasets, and TFCE in the remaining 1.7%.

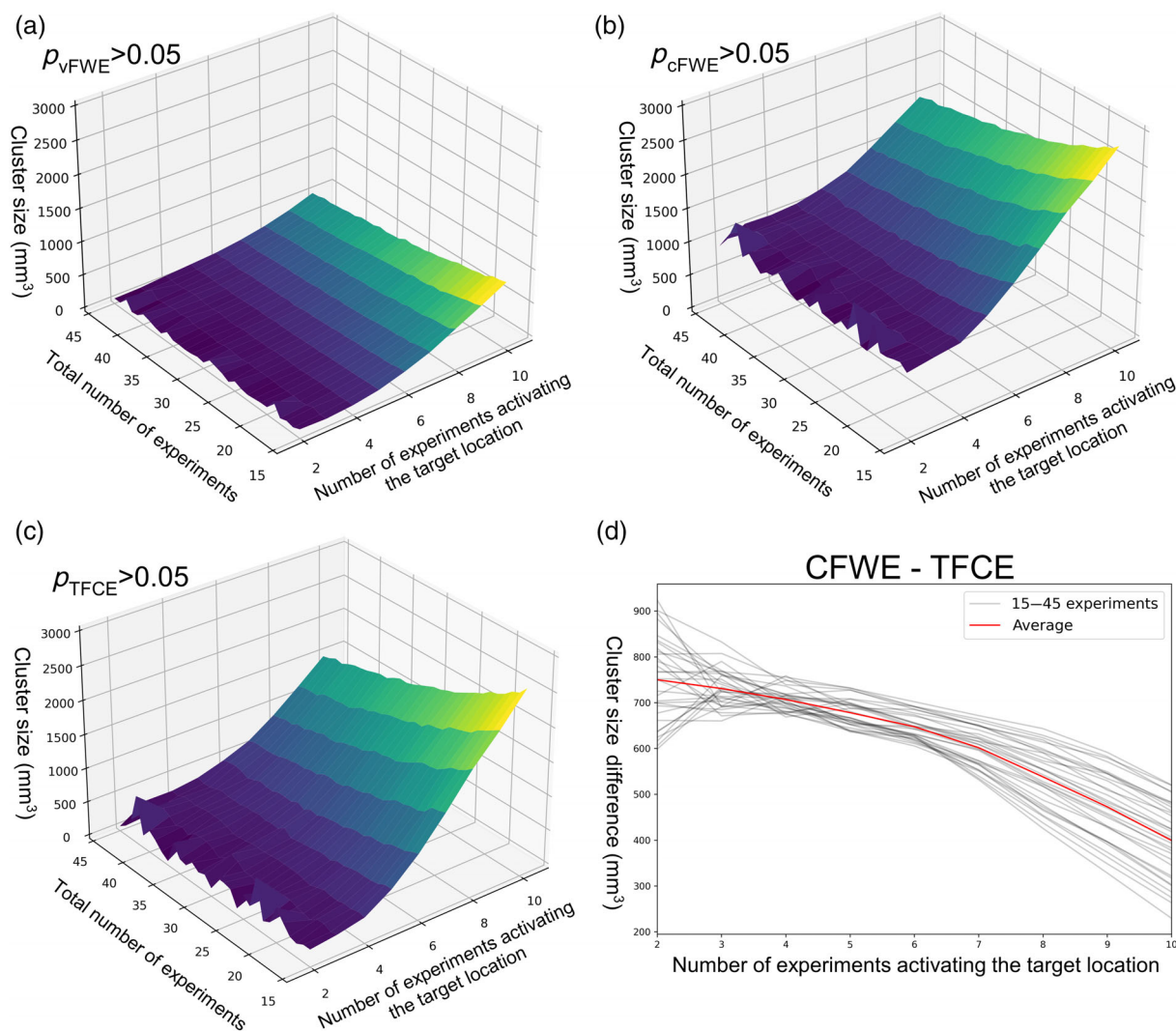
### 3.3 | Cluster size

The size of the cluster of voxels considered significant is strongly influenced by both the total number of experiments and the number of experiments activating the target location, but in different directions. We observed smaller significant clusters with a higher total number of experiments due to stricter significance thresholds resulting from the permutation test. In contrast, having more experiments activating the target location led to bigger clusters, because of the higher probability of larger displacement around the true activation location.

Using cFWE correction yielded the largest significant clusters, while vFWE correction yielded the smallest, which is in accordance with the findings of Eickhoff et al. (2016). TFCE performed more similar to cFWE than to vFWE but still showed noticeably smaller clusters than did cFWE, especially at lower true activation numbers (Figure 4). Taking together the results on sensitivity and cluster size, it appears that TFCE is not more likely to detect smaller clusters but just yields smaller clusters for the given level of convergence, relative to cFWE correction.

### 3.4 | Susceptibility to spurious convergence

The next measure we compare the three multiple comparison correction methods on was their susceptibility to spurious convergence. As can be seen in Figure 5, all three multiple-comparison correction methods showed an average level of spurious convergence of around 0.05. This is exactly what was expected given the fact that we



**FIGURE 4** (a–c) Cluster size of statistically significant areas of convergence that include at least one voxel in a 4-mm radius around the true location, under the different levels of the two simulation parameters (number of experiments and number of experiments activating the target location) and their 341 combinations. The number of experiments activating the target location was strongly positively correlated with cluster size, while the total number of experiments showed a negative correlation. cFWE correction featured the largest clusters closely followed by TFCE. The clusters declared significant by vFWE correction were exceedingly small in comparison. (d) Zooming in on the difference in cluster sizes between cFWE and TFCE corrections, it can be observed that the difference became more pronounced with fewer experiments activating the target location. This is because cFWE correction will always only result in relatively large clusters, while TFCE can potentially yield single significant voxels. This difference was more pronounced at lower convergence levels.

controlled the family-wise error rate to be 5% with all three methods. As described above, this was done by virtue of calculating a permutation null-distribution of ALE scores and selecting a threshold for which only 5% of random simulations feature more extreme values, which corresponds to the likelihood observed here.

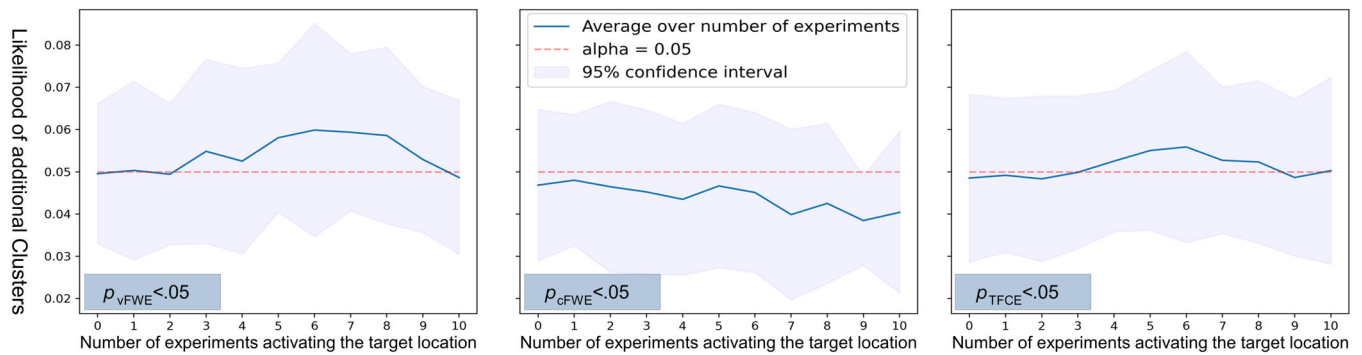
### 3.5 | Supplementary analyses

To further substantiate our simulation results we ran additional analyses examining larger datasets, different parameter settings for TFCE, as well as varying the size of spread around the “true” location. When applying ALE to datasets of 75, 100, or 150 experiments, respectively,

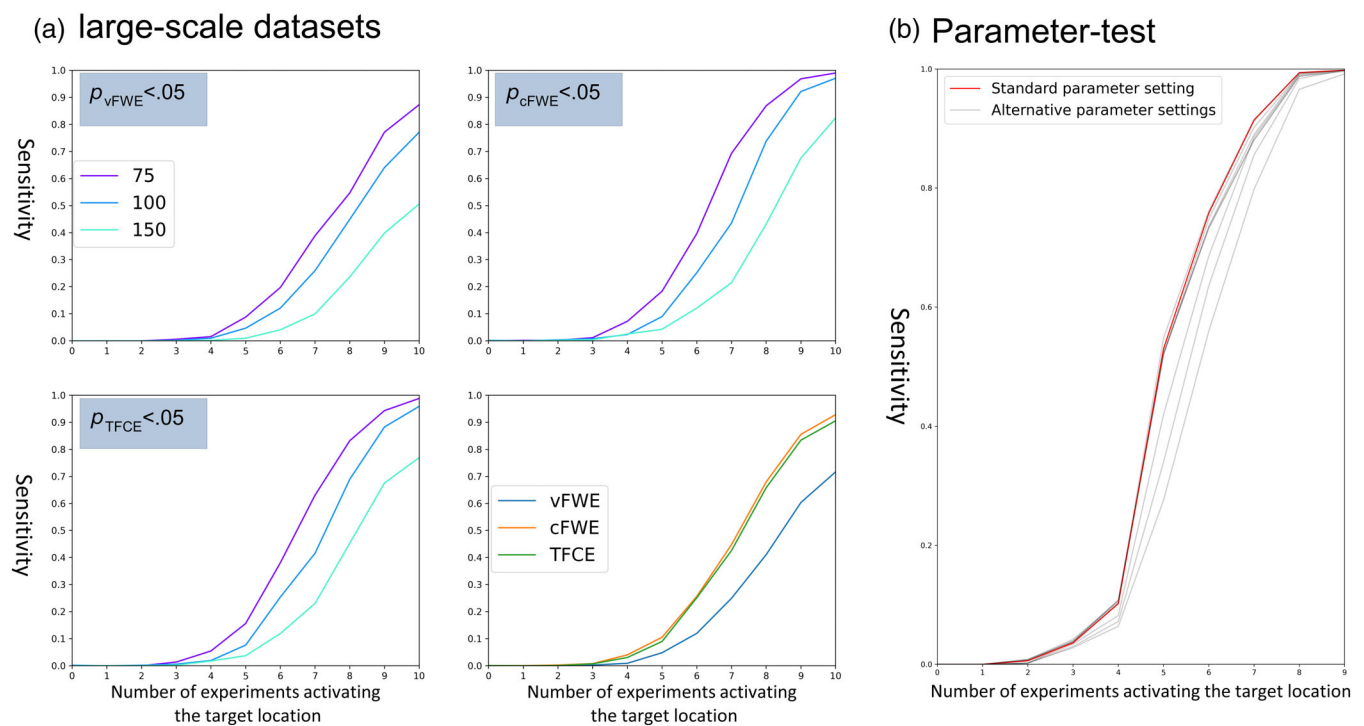
we observed sensitivity patterns that corresponded to the ones found in the main analysis: cFWE correction performed slightly better than TFCE, and vFWE correction performed least well (Figure 6a).

The TFCE parameter testing confirmed previous literature by showing that the standard parameter setting ( $H = 2$ ,  $E = 0.5$ ) performed best in almost every scenario, and the previously recommended parameter values should therefore be considered optimal for ALE analyses, too (Figure 6b).

As a last supplementary analysis, we multiplied the displacement around the “true location” by a factor between 0.5 and 1.5 with 0.1 increments. Sensitivity of TFCE improved slightly at lower displacements, while cFWE correction performed comparatively better at higher displacements. The general sensitivity patterns remained the same as in the main analysis.



**FIGURE 5** The likelihood of additional significant clusters as a function of the number of experiments activating the target location, averaged across the total number of experiments (blue line). As can be seen, all three multiple-comparison correction methods largely succeeded at controlling for an alpha error of .05.



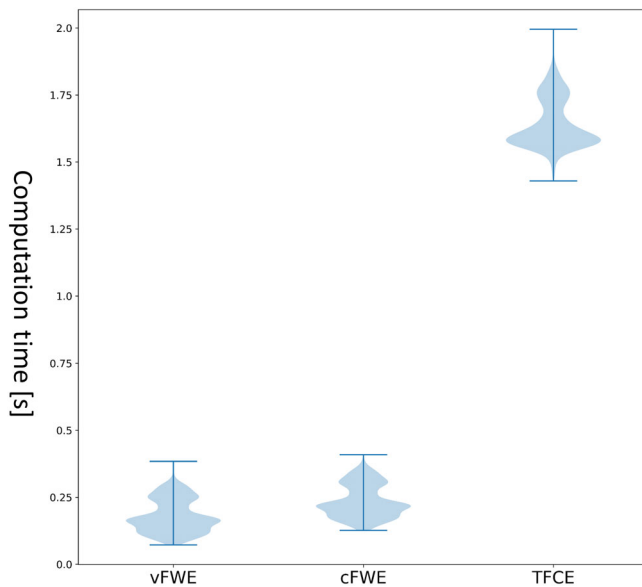
**FIGURE 6** (a) Sensitivity of ALE in a large-scale meta-analysis setting when applying different multiple comparison correction methods for statistical inference. The general trend observed in the main simulations holds for large-scale datasets as well. Sensitivity increased as a function of experiments activating the target location in a sigmoid fashion and the lower the rate of experiments activating the target location is in comparison to the total number of experiments, the lower sensitivity became. Lower right: Zooming in on the difference between cFWE and TFCE corrections: Even though TFCE performed slightly better at 7 and 8 experiments activating the target location for datasets including 150 studies, cFWE showed higher sensitivity on average. (b) Sensitivity of ALE corrected with TFCE using different parameter levels for the E and H exponent looking at a dataset of  $n = 30$  experiments. The standard setting (indicated in red), described in the literature as a fixed setting, is  $H = 2$  and  $E = 0.5$ . We used combinations of  $H = [1.8, 2.0, 2.2]$  with  $E = [0.3, 0.5, 0.7]$  (indicated in gray) to see if other values would improve performance. Overall, the standard parameter setting performed best or at least on par with other parameter settings.

### 3.6 | Computational efficiency

An important aspect in evaluating the usefulness of statistical methods is assessing how computationally expensive (i.e., time intensive) they are. TFCE turned out to be the by far computationally most expensive option (Figure 7). Calculating a single TFCE null permutation took about 1.62 s (SD = 0.118), averaged over

50 repetitions of each dataset size (15–45). Comparing this to vFWE (M = 0.178 s, SD = 0.069) and cFWE (M = 0.233 s, SD = 0.070), TFCE took approximately 7–9 times longer for each permutation. As even a standard ALE analysis will require about 10,000 permutations for a robust estimation of the null distribution, this could become burdensome for everyday research practice.





**FIGURE 7** Computation time required for a single null permutation of each multiple-comparison correction method. Times measured for 50 datasets per dataset size (15–45), totaling 1550 timepoints. vFWE and cFWE corrections run almost equally fast, while TFCE takes up to nine times as much time as the other two methods.

## 4 | DISCUSSION

This study aimed to compare the performance of TFCE, as a means to correct for alpha error inflation in multiple statistical comparisons, to the current standard methods employed in the context of ALE meta-analyses. To achieve this, we created 170,500 unique datasets comprising between 15 and 45 experiments each, of which 0 to 10 experiments featured an activation focus close to a target location chosen beforehand as the location of ground-truth. Running ALE on these datasets, we found that on the most important metric, sensitivity, TFCE performed on average at best equally well as cFWE and most of the time slightly worse. This performance difference was also found in larger datasets (75, 100, or 150 experiments) and was not reduced by varying TFCE parameters. Furthermore, TFCE was found to be highly computationally expensive, taking up about 70% of the computation time used for all simulations.

### 4.1 | TFCE versus cFWE

Evaluating TFCE in an ALE setting, we found TFCE to perform very well, featuring high sensitivity at an expected level of susceptibility to spurious convergence. When comparing TFCE to the other established methods, TFCE was more sensitive than vFWE correction and very similar to cFWE correction in terms of sensitivity. However, when looking at the differences between TFCE and cFWE correction in more detail, it can be seen that on average cFWE correction showed a higher power to detect the true effect, especially in the

range of four to eight experiments activating the target location. It has to be noted though that in around 1.5% of datasets TFCE was able to detect the true effect, while cFWE did not. In certain situations, it might therefore be beneficial for the researcher to additionally apply TFCE to check for clusters of activation. The exact conditions for these situations will need to be explored by future research.

Applied to simulated ALE results, TFCE's performance has fallen a little bit short of the expectations raised by the current neuroimaging literature. In a methodological comparison by Noble et al. (2020), examining the sensitivity of cFWE correction and TFCE in resampled data from the Human Connectome Project (Van Essen et al., 2013), it was found that TFCE achieved approximately double the sensitivity of cFWE. This superiority of TFCE was also reported in other recent methodological papers (Chen et al., 2018; Han et al., 2019). The reason for the subpar performance of TFCE in the context of ALE analyses, in comparison to the literature, is not completely clear but we can offer two hypotheses. One possible explanation is the fact that due to the Gaussian modeling included in the ALE algorithm, there will be no highly focal signals in the resulting statistical images, the proper consideration of which is supposed to be a particular strength that TFCE has over cFWE (Smith & Nichols, 2009). The other option is that TFCE might not deal well with the sparsity of input, namely the peak coordinates, which form the basis of any CBMA.

### 4.2 | Comparison to previous ALE simulation

In 2016, Eickhoff and colleagues were the first to employ a large-scale simulation approach to quantify performance of multiple-comparison correction methods in an ALE setting. The current study showed very similar ALE behavior patterns and reproduced the relative difference in performance between cFWE and vFWE corrections. This corroborates the robustness of the simulations and ALE in general as there were several differences between this study's design and the one used in 2016, as detailed below:

1. Here, we used a different range for the number of experiments included and showed that Eickhoff et al.'s (2016) conclusions also hold for substantially larger meta-analysis samples. Nevertheless, future simulations might include even larger sample sizes, as with more neuroimaging studies being published each year, the size of typical meta-analysis samples will potentially increase as well.
2. The empirical parameters used in this study were sampled from a newer, more comprehensive version of the BrainMap database, and we applied stricter filter criteria for study selection.
3. All simulations were run on a new python-based ALE pipeline which is a translation of the in-house MATLAB scripts. The replication of findings additionally serves as a validation of the new python pipeline.

Taking into account these differences, the consistency of our results with those reported in Eickhoff et al. (2016) is highly

reassuring, attesting to the robustness of ALE in general and to the validity of our simulation setting in particular.

### 4.3 | Transferability to other CBMA techniques

Besides ALE, there are multiple other algorithms used to perform CBMA, most importantly multi-level kernel density analysis (MKDA; Kober & Wager, 2010; Nee et al., 2007; Wager et al., 2009) and signed-difference map analysis (SDM; Palaniyappan et al., 2012; Radua et al., 2012; Radua et al., 2010). While all of these methods share core features, the exact implementation of the algorithms differs quite drastically. Therefore, any conclusions drawn regarding the behavior of statistical tools within ALE might not hold for the other methods. This means that TFCE might still confer tangible advantages when used as a thresholding method in SDM, MKDA, or other CBMA algorithms, pending proper evaluation.

## 5 | CONCLUSION

Using a large-scale simulation approach, we evaluated whether TFCE would be a valuable addition to the current ALE algorithm for CBMA, serving as a sophisticated method to correct for multiple comparisons in mass-univariate analyses of neuroimaging data. Based on previous evaluations, we expected TFCE to confer a higher power to detect true effects while also improving on some of the methodological shortcomings of the current standard cFWE correction. However, our simulation results did not support this expectation, as TFCE's sensitivity was at best on par with that of the standard cFWE correction. Given that TFCE (vs. cFWE correction) incurs substantially higher computational costs, we conclude that, in most cases, ALE analyses would not benefit from employing TFCE, which therefore cannot be recommended as a standard approach in this context for the sake of efficiency.

### AUTHOR CONTRIBUTIONS

Simon B. Eickhoff, Theodore D. Satterthwaite, and Peter T. Fox conceived the design of the presented study and the analysis techniques used. Lennart Frahm and Felix Hoffstaedter wrote the code pipeline and collected the data. Lennart Frahm, Edna C. Cieslik, Robert Langner, and Simon B. Eickhoff analyzed and interpreted the collected data. Lennart F. Edna C. Cieslik, and Robert Langner came up with the first drafts of the paper. All authors discussed the results and contributed to the final manuscript.

### ACKNOWLEDGMENTS

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, EI 816/11-1 and International Research Training Group 2150, 269953372/GRK2150), the National Institute of Mental Health (R01-MH074457), the National Institute of Aging (P30-AG066546), and the Jülich-Aachen Research Alliance (JARA) granting computation time on the supercomputer JURECA (Jülich Supercomputing Centre, 2018) at

Forschungszentrum Jülich. Open access funding enabled and organized by Projekt DEAL.

### CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

### DATA AVAILABILITY STATEMENT

The data and code that support the findings of this study are openly available in TFCE\_ALE at [https://github.com/LenFrahm/TFCE\\_ALE](https://github.com/LenFrahm/TFCE_ALE).

### ORCID

Lennart Frahm  <https://orcid.org/0000-0001-8907-883X>

### REFERENCES

- Acar, F., Seurinck, R., Eickhoff, S. B., & Moerkerke, B. (2018). Assessing robustness against potential publication bias in Activation Likelihood Estimation (ALE) meta-analyses for fMRI. *PLoS One*, 13(11), e0208177. <https://doi.org/10.1371/journal.pone.0208177>
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604), 452–454.
- Bossier, H., Roels, S. P., Seurinck, R., Banaschewski, T., Barker, G. J., Bokde, A. L., Quinlan, E. B., Desrivieres, S., Flor, H., & Grigis, A. (2020). The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage*, 212, 116601.
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810), 84–88.
- Chen, X., Lu, B., & Yan, C. G. (2018). Reproducibility of R-fMRI metrics on the impact of different strategies for multiple comparison correction and sample sizes. *Human Brain Mapping*, 39(1), 300–318.
- Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(1), 62–70. <https://doi.org/10.1016/j.neuroimage.2008.05.021>
- Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLoS One*, 12(11), e0184923. <https://doi.org/10.1371/journal.pone.0184923>
- Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, 59(3), 2349–2361. <https://doi.org/10.1016/j.neuroimage.2011.09.017>
- Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, 30(9), 2907–2926. <https://doi.org/10.1002/hbm.20718>
- Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., Bzdok, D., & Eickhoff, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *NeuroImage*, 137, 70–85. <https://doi.org/10.1016/j.neuroimage.2016.04.072>
- Evans, A. C., Kamber, M., Collins, D., & MacDonald, D. (1994). An MRI-based probabilistic atlas of neuroanatomy. In *Magnetic resonance scanning and epilepsy* (pp. 263–274). Springer.
- Han, H., Glenn, A. L., & Dawson, K. J. (2019). Evaluating alternative correction methods for multiple comparison in functional neuroimaging research. *Brain Sciences*, 9(8), 198.
- Jülich Supercomputing Centre. (2018). JURECA: Modular supercomputer at Jülich supercomputing Centre. *Journal of Large-scale Research Facilities*, 4, A132. <https://doi.org/10.17815/jlsrf-4-121-1>

- Kober, H., & Wager, T. D. (2010). Meta-analysis of neuroimaging data. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2), 293–300.
- Laird, A. R., Eickhoff, S. B., Kurth, F., Fox, P. M., Uecker, A. M., Turner, J. A., Robinson, J. L., Lancaster, J. L., & Fox, P. T. (2009). ALE meta-analysis workflows via the brainmap database: Progress towards a probabilistic functional brain atlas. *Frontiers in Neuroinformatics*, 3, 23 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2715269/pdf/fninf-03-023.pdf>
- Laird, A. R., Lancaster, J. J., & Fox, P. T. (2005). BrainMap: The social evolution of a human brain mapping database. *Neuroinformatics*, 3(1), 65–77.
- Lett, T. A., Waller, L., Tost, H., Veer, I. M., Nazeri, A., Erk, S., Brandl, E. J., Charlet, K., Beck, A., Vollstadt-Klein, S., Jorde, A., Kiefer, F., Heinz, A., Meyer-Lindenberg, A., Chakravarty, M. M., & Walter, H. (2017). Cortical surface-based threshold-free cluster enhancement and cortexwise mediation. *Human Brain Mapping*, 38(6), 2795–2807. <https://doi.org/10.1002/hbm.23563>
- Li, H., Nickerson, L. D., Nichols, T. E., & Gao, J. H. (2017). Comparison of a non-stationary Voxelation-corrected cluster-size test with TFCE for group-level MRI inference. *Human Brain Mapping*, 38(3), 1269–1280.
- Makin, T. R., & de Xivry, J.-J. O. (2019). Science forum: Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, 8, e48175.
- Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience*, 7(1), 1–17.
- Noble, S., Scheinost, D., & Constable, R. T. (2020). Cluster failure or power failure? Evaluating sensitivity in cluster-level inference. *Neuroimage*, 209, 116468.
- Palaniyappan, L., Balain, V., Radua, J., & Liddle, P. F. (2012). Structural correlates of auditory hallucinations in schizophrenia: A meta-analysis. *Schizophrenia Research*, 137(1–3), 169–173.
- Radua, J., Mataix-Cols, D., Phillips, M. L., El-Hage, W., Kronhaus, D. M., Cardoner, N., & Surguladze, S. (2012). A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps. *European Psychiatry*, 27(8), 605–611.
- Radua, J., van den Heuvel, O. A., Surguladze, S., & Mataix-Cols, D. (2010). Meta-analytical comparison of voxel-based morphometry studies in obsessive-compulsive disorder vs other anxiety disorders. *Archives of General Psychiatry*, 67(7), 701–711.
- Salimi-Khorshidi, G., Smith, S. M., & Nichols, T. E. (2011). Adjusting the effect of nonstationarity in cluster-based and TFCE inference. *NeuroImage*, 54(3), 2006–2019.
- Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1), 83–98. <https://doi.org/10.1016/j.neuroimage.2008.03.061>
- Spisak, T., Spisak, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., & Kincses, T. (2019). Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. *NeuroImage*, 185, 12–26. <https://doi.org/10.1016/j.neuroimage.2018.09.078>
- Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: Method and validation. *NeuroImage*, 16(3 Pt 1), 765–780. <https://doi.org/10.1006/nimg.2002.1131>
- Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., & Fox, P. (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, 33(1), 1–13. <https://doi.org/10.1002/hbm.21186>
- Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, 1(1), 1–10.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & Wu-Minn HCP Consortium. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, 80, 62–79.
- Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: Current and future directions. *Social Cognitive and Affective Neuroscience*, 2(2), 150–158. <https://doi.org/10.1093/scan/nsm015>
- Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., & Van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, 45(1), S210–S221.

**How to cite this article:** Frahm, L., Cieslik, E. C., Hoffstaedter, F., Satterthwaite, T. D., Fox, P. T., Langner, R., & Eickhoff, S. B. (2022). Evaluation of thresholding methods for activation likelihood estimation meta-analysis via large-scale simulations. *Human Brain Mapping*, 43(13), 3987–3997. <https://doi.org/10.1002/hbm.25898>