Collaborative non-parametric two-sample testing

Alejandro de la Concha **Nicolas Vayatis** **Argyris Kalogeratos**
Centre Borelli, ENS Paris-Saclay, Université Paris-Saclay, Gif-sur-Yvette, France

Abstract

Multiple two-sample test problem in a graph-structured setting is a common scenario in fields such as Spatial Statistics and Neuroscience. Each node v in fixed graph deals with a two-sample testing problem between two node-specific probability density functions, p_v and q_v. The goal is to identify nodes where the null hypothesis $p_v = q_v$ should be rejected, under the assumption that connected nodes would yield similar test outcomes. We propose the non-parametric *collaborative two-sample testing* (CTST) framework that efficiently leverages the graph structure and minimizes the assumptions over p_v and q_v. CTST integrates elements from ϕ-divergence estimation, Kernel Methods, and Multitask Learning. We use synthetic experiments and a real sensor network detecting seismic activity to demonstrate that CTST outperforms state-of-the-art non-parametric statistical tests that apply at each node independently, hence disregard the geometry of the problem.

1 INTRODUCTION

Given two pdfs p and q, a *two-sample test* (TST) assesses if the null hypothesis, $\mathrm{H}_{\text{null}} : p = q$, is true, versus the alternative $\mathrm{H}_{\text{alt}} : p \neq q$. TST has been studied in the Machine Learning literature (Sugiyama et al., 2011a; Gretton et al., 2012; Harchaoui et al., 2013; Lopez-Paz and Oquab, 2017; Bargiotas et al., 2021). As in most statistical problems, passing from the typical univariate to a multivariate setting is non-trivial. Performing multiple two-sample tests encounters the *multiple comparison problem* (MCP), which refers to the fact that the probability of wrongly rejecting a set of null hypotheses (false positives, or Type-I error), increases artificially with the number of tests. MCP treatments include the Bonferroni correction that scales the π-values[1] by the number of tested hypotheses (Dunn, 1961), and the non-parametric resampling test that employs the maximum statistic and permutation tests (Westfall and Young, 1992).

Multiple Two-Sample Testing (MTST) emerges in fields such as Spatial Statistics and Neuroscience, where each test is applied to local data sampled from a different 'location', and the validity of null hypotheses often depends on the 'proximity' between locations. MTST is corroborated, for instance, by the Hebbian perspective stating "*Neurons that fire together wire together*" (Hebb, 1949), which that is common ground in Neuroscience, or Tobler's *first law of Geography* (Tobler, 1970) eloquently stating "*Everything is related to everything else, but near things are more related than distant things*", which is cornerstone in Spatial Statistics.

Multiple two-sample testing on graphs. Motivated by the above application fields, we study the particularly challenging problem of *graph-structured* MTST, where a TST is considered at each node $v \in V = \{1, ..., N\}$ of a given fixed graph G, comparing two node-specific pdfs p_v and q_v. Then, all hypotheses

$$\left\{ \mathrm{H}_{\text{null},v} : \;\; p_v = q_v \quad \text{vs.} \quad \mathrm{H}_{\text{alt},v} : \;\; p_v \neq q_v \right\}_{v \in V} \qquad (1)$$

are simultaneously tested to determine the set $R_{\text{MT}} = \{v \in V \,|\, \mathrm{H}_{\text{null},v} \text{ is found false}\}$ containing the nodes whose null hypotheses are to be rejected with a given level of confidence $1 - \pi^*$. The goal is for R_{MT} to be as close as possible to the set of hypotheses where $p_v \neq q_v$ truly holds, denoted by $\mathbf{I}_0^{\text{c}}$ (i.e. the set complement of $\mathbf{I}_0$). As in any *multiple hypothesis testing* (MT) approach, also in this case determining R_{MT} requires three components:

1. A test statistic S_v for $\mathrm{H}_{\text{null},v}$, to be estimated using the data of node v, to quantify the dissimilarity of p_v and q_v.
2. A π-value estimation framework to identify which of the $\{\mathrm{H}_{\text{null},v}\}_{v \in V}$ to be rejected.
3. A Type-I error correction strategy to control for the MCP.

[1] p-values appear as π-values to distinguish from pdf p

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

In the context of graph-structured MTST, to the best of our knowledge, there exist mostly plug-in methods, in the sense that: they perform (1) and (2) independently for each node; then, for (3) they apply post-hoc Type-I error correction using aggregation mechanisms over the estimated π-values, or alternatively they avoid the MCP by defining a single test statistic from the multiple test statistics $\{S_v\}_{v\in V}$, and then they estimate a π-value based on that quantity. The main drawback of these approaches is that the individual test statistics that fail to quantify accurately the difference between each pair (p_v vs. q_v) may lead to inaccurate conclusions.

Notable graph-structured MT techniques include the *Permutation Cluster Test* (PCT) (Maris and Oostenveld, 2007), *Threshold-free Cluster Enhancement* (TFCE) (Smith and Nichols, 2009), and the *Structure-Adaptive Benjamini Hochberg Algorithm* (SAHBA) (Li and Barber, 2018), which all assume that the rejected null hypotheses should be associated with a group of connected nodes. The π-values of PCT and TFCE are estimated via a permutation test over a maximum test statistic. In contrast, SAHBA uses a reweighting mechanism of the node-level π-values, under the assumption that connected nodes should show similar π-values.

Contribution. We present the *Collaborative Two-Sample Test* (CTST): a graph-structured TST built upon non-parametric methods and the notion of graph smoothness. Fig. 1 gives a visual overview of the approach. Distinct from existing works, CTST's core novelty is that it not only *estimates jointly and in an associative manner* all node-level test statistics, but it also intertwines estimation with the identification of the hypotheses to be rejected.

2 BACKGROUND ELEMENTS

This section gives general notations and a brief background for each of the diverse technical elements that we combine in this work.

General notations. A_{ij} denotes the entry at the i-th row and j-th column of a matrix A, and $A_{i,:}$ is its i-th row. $\text{vec}(a_1,..,a_n)$ denotes the concatenation of the input vectors $a_1,...,a_n$ in a single vector. $\mathbf{1}_M$ is a vector with M ones (resp. $\mathbf{0}_M$), I_M is a $M\times M$ identity matrix, and $\mathbb{1}\{\cdot\}$ is the indicator function. The Euclidean norm and the dot product are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, and when endowed a functional space $\mathcal{F}$ by $\|\cdot\|_{\mathcal{F}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{F}}$. For an observation x belonging to a d-dimensional input space, we write $x\in\mathcal{X}\subset\mathbb{R}^d$.

A fixed undirected weighted graph $G=(V,E,W)$ is defined by the set of N nodes V, and the set of edges E. We suppose positive-weighted and undirected edges and nodes without self-loops, i.e. for the weight matrix $W\in\mathbb{R}^{N\times N}$ it holds $W_{uu}=0$, $\forall u\in V$, and $W_{uv}=W_{vu}\geq 0$. In the rest, composite objects (vectors, matrices, sets, etc.) that refer to all the nodes of a graph appear in bold font. Finally, *graph smoothness* is central in this work; the smoothness of a graph function $\vartheta: V\to\mathbb{R}$ over G is defined as $\sum_{(u,v)\in E} W_{uv}(\vartheta(u)-\vartheta(v))^2$. This notion generalizes for N estimates over the nodes of G, which is justified by the expected smooth variation of a phenomenon over a graph, and in turn motivates the use of graph regularization techniques.

ϕ-divergences and likelihood-ratio. ϕ-divergences are non-negative functions measuring the dissimilarity between two probability measures. For two probability measures P and Q defined as:

$$\mathcal{D}_\phi(P\|Q)=\begin{cases}\displaystyle\int\phi\left(\frac{dQ}{dP}\right)(x)dP(x), & \text{if } Q\ll P\\ +\infty & \text{if } Q\not\ll P\end{cases} \tag{2}$$

where $Q\ll P$ means Q is absolutely continuous w.r.t. P, and $\phi:\mathbb{R}\to\mathbb{R}$ is a proper closed convex function from $(-\infty,\infty)$ to $[0,\infty]$ with $\phi(1)=0$ and such that its domain $\text{Dom}(\phi)=\{x\in\mathbb{R}|\phi(x)<\infty\}$ is an interval with endpoints $a_\phi<1<b_\phi$ (Csiszár, 1967; Broniatowski and Keziou, 2006; Birrell et al., 2022a). A ϕ-divergence measures the dissimilarity between P and Q as it satisfies $\mathcal{D}_\phi(P\|Q)\geq 0$. Moreover, if ϕ is strictly convex in a neighborhood around 1 then $\mathcal{D}_\phi(P\|Q)=0$ if and only if $P=Q$. Notably, for $z\in\mathbb{R}$, when $\phi(z)=-\log(z)$ we recover the well-known Kullback-Leibler divergence (KL-divergence) (Kullback, 1959), and when $\phi(z)=\frac{1}{2}(z-1)^2$ we obtain the Pearson's χ^2-divergence (Pearson, 1900).

The function $r(x)=\frac{dQ}{dP}(x)$ is called the Radon-Nikodym derivative. When both probability measures P and Q, admit a pdf w.r.t. a reference measure ρ, denotes as p and q, respectively, the Radon-Nikodym derivative becomes the quotient $r(x)=\frac{q(x)}{p(x)}$, which is known in statistics as *likelihood-ratio* or *density-ratio*.

3 THE COLLABORATIVE NON-PARAMETRIC TWO-SAMPLE TEST (CTST)

3.1 Problem statement

Let a fixed undirected and positive-weighted graph be $G=(V,E,W)$, and suppose each node $v\in V$ has $n+n'$ (same for all nodes) iid observations from two unknown pdfs, p_v and q_v, respectively. The two data observations subsets taking values in the input space $\mathcal{X}\subset\mathbb{R}^d$ are:

$$\begin{cases}\mathbf{X}=\{\mathbf{X}_v\}_{v\in V}=\left\{\{x_{v,i}:x_{v,i}\overset{\text{iid}}{\sim}p_v\}_{i=1}^{n}\right\}_{v\in V};\\ \mathbf{X}'=\{\mathbf{X}'_v\}_{v\in V}=\left\{\{x'_{v,i}:x'_{v,i}\overset{\text{iid}}{\sim}q_v\}_{i=1}^{n'}\right\}_{v\in V}.\end{cases} \tag{3}$$

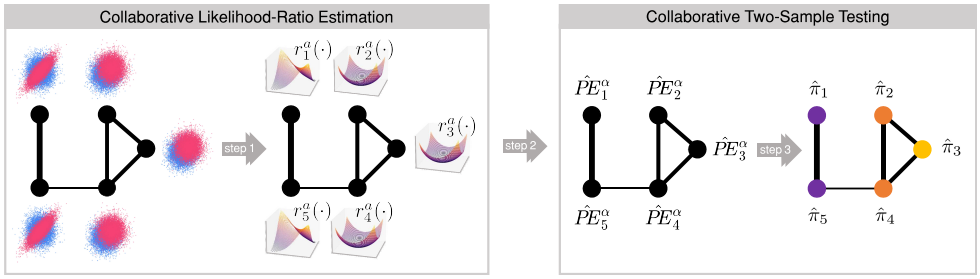


Figure 1: Collaborative multiple two-sample testing (CTST) based on the collaborative likelihood-ratio estimation (LRE) over a graph. Left: Given observations from two pdfs, p_v (blue) and q_v (pink) at each node v of a graph, the associated relative likelihood-ratios $\{r_v^\alpha\}_v$ are estimated in a collaborative manner. Any given $x \in \mathcal{X}$ gets mapped to the graph signal $(r_1^\alpha(x),...,r_N^\alpha(x))^\mathsf{T}$. **Right:** The likelihood-ratio approximations are used to estimate the node-level χ^2-divergence between p_v and q_v, which are used as test statistics. π-values $\hat{\pi}_v$ are computed for each node that allow the identification of nodes where the null hypothesis $p_v = q_v$ is rejected, while controlling the Type-I error rate.

The proposed CTST aims at solving the graph-structured MTST problem of Expr. 1 (see Fig. 1), and in general terms, comprises three steps:

- *Step 1 – Collaborative likelihood-ratio estimation*: Joint estimation of the node-level relative likelihood-ratios, $\mathbf{r}^\alpha = (r_1^\alpha,...,r_N^\alpha)$, using the available data (Eq. 3). The vector-valued function $\mathbf{r}^\alpha$ is then used to approximate for each node v the χ^2-divergence between the associated p_v and q_v.
- *Step 2 – Node-level test statistics*: The ϕ-divergences' properties (see Sec. 2) make them good candidates for node-level test statistics. To deal with their non-symmetricity, at each node the pair of node-level test statistics $\{S_v\}_{v\in V}$, $\{S'_v\}_{v\in V}$ are used, which corresponds to both direction in which the χ^2-divergence is approximated.
- *Step 3 – π-value estimation*: A permutation test is used for estimating two sets of node-level π-values, $\{\pi_v\}_{v\in V}$ and $\{\pi'_v\}_{v\in V}$. These sets are used to identify the set of null hypotheses to be rejected (R_{CTST}). The permutation test guarantees weak FWER control.

LRE and χ^2-divergence maximization. Hypothesis testing defines a decision rule based on a test statistic that estimated from the available data, hence any test depends on its underlying estimation approach. In this paper, we capitalize over the connection between the likelihood-ratio and the χ^2-divergence estimation problem to avoid making distributional hypothesis on the node level pdfs.

To start, consider P, Q two probability measures with pdfs p, q respectively. $Q \not\ll P$ implies the likelihood-ratio does not exist, which is a situation may occur in practice when the support of Q is not included in the support of P. To overcome such issues, consider the convex combination between the measure P and Q can be considered: $P^\alpha = (1-\alpha)P + \alpha Q$, where $0 < \alpha \leq 1$. It is easy to verify that $Q \ll P^\alpha$. Under this formulation, the likelihood-ratio $r^\alpha(x) = \frac{q(x)}{p^\alpha(x)}$ is a bounded function ($\sup_{x\in\mathcal{X}} |r^\alpha(x)| \leq \frac{1}{\alpha}$). We call r^α the relative likelihood-ratio. Then, we can use r^α to estimate the χ^2-divergence between p^α and q:

$$PE^\alpha(P,Q) := PE(P^\alpha\|Q) = \int \frac{(r^\alpha(y)-1)^2}{2} dP^\alpha(y). \quad (4)$$

Notice that $PE(P^\alpha\|Q)$ is a valid dissimilarity measure between P and Q despite the α-regularization. For any $0 < \alpha < 1$, $PE(P^\alpha\|Q) \geq 0$ and, as $\phi(\zeta) = \frac{(\zeta-1)^2}{2}$ is strictly convex around 1, $PE(P^\alpha\|Q) = 0$ iff $P^\alpha = Q$, which implies $P = Q$.

Theorem 3.1 states that estimating $PE(P^\alpha\|Q)$ is equivalent to estimating r^α directly in terms of a Reproducing Kernel Hilbert Space (RKHS). The technical details of this result is left to Appendix A.

Theorem 3.1. *Let us denote by $\mathbb{H}$ a RKHS containing as elements functions $f : \mathcal{X} \to \mathbb{R}$. $\mathbb{H}$ is equipped with the inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}} : \mathbb{H}\times\mathbb{H} \to \mathbb{R}$, which will be reproduced by a positive semi-definite kernel function $K(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Let us assume the kernel function $\mathrm{K}(\cdot,\cdot)$ is measurable and there exists a constant κ, $0 < \kappa < \infty$ such that $\sup_{x\in\mathcal{X}} \sqrt{\mathrm{K}(x,x)} \leq \kappa$, then the variational formulation of the χ^2-divergence between P^α and Q takes the form:*

$$PE^\alpha(P\|Q) \geq \sup_{f\in\mathbb{H}} \int f(x')\, dQ(x') - \int \frac{f^2(y)}{2} dP^\alpha(y) - \frac{1}{2} \quad (5)$$

where f is an approximation to the likelihood-ratio r^α. If $r^\alpha \in \mathbb{H}$, the equality in Expr. 5 is attained.

3.2 Step 1: Collaborative LRE

The CTST framework estimates the difference between p_v and q_v by exploiting the graph smoothness hypothesis. Problem 3.1 and the geometry of the RKHS allows a more formal statement, relying on the fact that two

functions $f_u, f_v \in \mathbb{H}$ close in the RKHS, give similar evaluations at the same point $x \in \mathcal{X}$:

$$|f_u(x) - f_v(x)| = |\langle \mathrm{K}(x,\cdot), f_u - f_v\rangle_{\mathbb{H}}| \leq \kappa \|f_u - f_v\|_{\mathbb{H}}, \quad (6)$$

thus, enforcing graph smoothness where small $\|f_u - f_v\|_{\mathbb{H}}$ for adjacent u and v nodes is expected to lead to similar test statistics.

Optimization problem. To estimate the χ^2-divergence between p_v and q_v for all nodes, first we estimate jointly the node-level relative likelihood-ratios, $\mathbf{r}^\alpha = (r_1^\alpha, ..., r_N^\alpha)$, using the available data (Eq. 3) by solving the following optimization problem:

$$\min_{\mathbf{f}\in\mathbb{H}^N} (1-\alpha) \sum_{x\in\mathbf{X}_v} \frac{f_v^2(x)}{n} + \alpha \sum_{x'\in\mathbf{X}'_v} \frac{f_v^2(x')}{n'} - \sum_{x'\in\mathbf{X}'_v} \frac{f_v(x')}{n'} + \frac{\lambda}{4} \sum_{u,v\in V} W_{uv} \|f_u - f_v\|_{\mathbb{H}}^2 + \frac{\lambda\gamma}{2} \sum_{v\in V} \|f_v\|_{\mathbb{H}}^2. \quad (7)$$

The first line is the negative variational representation of the χ^2-divergence at each node; the next term penalizes the non-smoothness of the estimates over the graph; last is a penalty term reducing the risk of overfitting (Sheldon, 2008); λ, γ are regularization constants.

Numerical implementation. To solve Problem 7, we employ GRULSIF that was proposed in de la Concha et al. (2024). Provided a dictionary $D_{\hat{L}}$ with $\hat{L}$ basis functions, such that the finite dimensional space $\mathbf{F} = \mathrm{span}(\{\varphi(x) : x \in D_{\hat{L}}\})$ approximates $\mathbb{H}$, it was further proposed to use Nyström approximation to replace the feature map $\varphi(x)$ by its orthogonal projection into the space $\mathbf{F}$. By determining a set of so-called *anchor points* in $\mathbb{H}$, $\varphi(x_1), ..., \varphi(x_{\hat{L}})$, and via the associated kernel matrix, $\mathcal{K}_{\hat{L}} \in \mathbb{R}^{\hat{L}\times\hat{L}}$, $[\mathcal{K}_{\hat{L}}]_{ij} = \mathrm{K}(x_i, x_j)$, the new feature map $\psi(\cdot) = \mathcal{K}_{\hat{L}}^{-\frac{1}{2}} \left(\mathrm{K}(\cdot, x_1), .., \mathrm{K}(\cdot, x_{\hat{L}})\right)^\mathsf{T}$ derives.

Writing Problem 7 in terms of the empirical expectations and by involving the Nyström approximation, the solution $\hat{\mathbf{f}} = (\hat{f}_1, ..., \hat{f}_N) \in \mathbb{H}^N$ takes the form: $\hat{f}_v(\cdot) = \psi(\cdot)^\mathsf{T}\hat{\theta}_v$, where $\hat{\theta}_v \in \mathbb{R}^L$. By vectorizing all node parameters in $\mathbf{\Theta} = \mathrm{vec}(\theta_1^\mathsf{T}, ..., \theta_N^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{N\hat{L}}$, Problem 7 is rewritten as a quadratic problem over $\mathbf{\Theta}$:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{N\hat{L}}} \frac{1}{N} \sum_{v\in V} \theta_v^\mathsf{T} \left(\frac{(1-\alpha)}{2} H_{\psi,v} + \frac{\alpha}{2} H'_{\psi,v} \right) \theta_v - h'^\mathsf{T}_{\psi,v}\theta_v + \frac{\lambda}{4} \sum_{u,v\in V} W_{uv} \|\theta_v - \theta_u\|^2 + \frac{\lambda\gamma}{2} \sum_{v\in V} \|\theta_v\|^2 \quad (8)$$

where

$$H_{\psi,v} = \frac{1}{n} \sum_{x\in\mathbf{X}_v} \psi(x)\psi(x)^\mathsf{T}, \quad H'_{\psi,v} = \frac{1}{n'} \sum_{x\in\mathbf{X}'_v} \psi(x)\psi(x)^\mathsf{T}$$
$$h'_{\psi,v} = \frac{1}{n'} \sum_{x\in\mathbf{X}'_v} \psi(x). \quad (9)$$

In the same work, it was proposed to solve Problem 8 with the Cyclic Block Coordinate Descent (CBCD) (Beck and Tetruashvili, 2013; Li et al., 2018). If $n = n'$, then the final computational cost is $\mathcal{O}(N\hat{L}^3 + nN\hat{L}^2 + N\hat{L}^2\log^2(N\hat{L}))$, where $\hat{L} \ll Nn$, which makes it scalable to large graphs.

Other implementation elements of GRULSIF is the selection of anchor points and hyperparameters, namely the parameters of the kernel function $\mathrm{K}(\cdot,\cdot)$ and λ, γ. To identify the anchor points, we employ the greedy method from Richard et al. (2009): it starts with the first element in the sequence, $D_1 = \{x_1\}$, and then a new element x_t is added if its coherence (μ) to the current dictionary D_{t-1} is below a given threshold μ_0:

$$\mu := \max_{x\in D_t} |\langle \varphi(x_t), \varphi(x)\rangle_{\mathbb{H}}| = \max_{x\in D_t} |K(x_t, x)| \leq \mu_0. \quad (10)$$

Given a unit-norm kernel (i.e. $K(x,x) = 1$, $\forall x \in \mathcal{X}$) with hyperparameters σ and a threshold coherence $\mu_0 \in (0,1)$, we run the method for all the observations.

For the regularization parameter α, since it requires special attention, we provide several insightful experiments for the CTST task in Appendix C.

3.3 Step 2: Node-level test statistics

There are multiple avenues to go from node-level LRE to the selection of hypotheses to be rejected. For example, in parametric hypothesis testing it is common to use the sum of the log-likelihood over the available data as a test statistic. Alternatively, instead of node-level test statistics, a global statistic that summarizes all the data could be used. In this paper, we propose to work with node-level test statistics that are approximations of the χ^2-divergence that was used as auxiliary cost function in the LRE problem (Expr. 5).

Notice that, the integration in Eq. 2 is w.r.t. P, hence in practice the output is more sensitive to points where P has higher mass, which also means that ϕ-divergences may be non-symmetric functions, i.e. $\mathcal{D}_\phi(p\|q) \not\equiv \mathcal{D}_\phi(q\|p)$. To address this non-symmetricity, we identify the set of hypotheses to be rejected (R_{CTST}) by considering both the comparisons $PE^\alpha(p\|q)$ and $PE^\alpha(q\|p)$ to derive two sets of test statistics:

$$\{S_v\}_{v\in V} = \{\hat{PE}_v^\alpha(\mathbf{X}_v\|\mathbf{X}'_v) \sim PE^\alpha(p_v\|q_v)\}_{v\in V};$$
$$\{S'_v\}_{v\in V} = \{\hat{PE}_v^\alpha(\mathbf{X}'_v\|\mathbf{X}_v) \sim PE^\alpha(q_v\|p_v)\}_{v\in V}. \quad (11)$$

Specifically for GRULSIF, after estimating the parameter vector $\hat{\mathbf{\Theta}}$, Theorem 3.1 allows to approximate $PE(p_v^\alpha\|q_v)$ by:

$$\hat{PE}_v^\alpha(\mathbf{X}_v\|\mathbf{X}'_v) := h'^\mathsf{T}_{\psi,v}\hat{\theta}_v - \frac{1-\alpha}{2}\hat{\theta}_v^\mathsf{T} H_{\psi,v}\hat{\theta}_v - \frac{\alpha}{2}\hat{\theta}_v^\mathsf{T} H'_{\psi,v}\hat{\theta}_v - \frac{1}{2} \quad (12)$$

It has been shown that $\hat{PE}^{\alpha}_{v}(\mathbf{X}_v \| \mathbf{X}'_v)$ is an asymptotic unbiased estimator of $PE^{\alpha}(p_v \| q_v)$, and that the graph smoothness hypothesis and the collaborative LRE becomes more relevant as the estimation problem becomes more challenging, e.g. when fewer observations per node are available (de la Concha et al., 2024).

3.4 Step 3: π-value estimation

Our MT strategy applies a threshold η^* to each estimated node-level π-value $\{\hat{\pi}_v\}_{v\in V}$, hence considers the set of rejected hypotheses $R_{\text{MT}} = \{v \in V \,|\, \hat{\pi}_v < \eta^*\}$. We denote by $TP = \#\{v \,|\, v \in \mathbf{I}_0^c \cap R_{\text{MT}}\}$ the number of true positives, and by $FP = \#\{v \,|\, v \in R_{\text{MT}} \cap \mathbf{I}_0\}$ the number of false positives. We address the MCP by weak control of the *k-Family-Wise Error Rate* (FWER for $k=1$), which controls the probability to occur at least one false rejection of the individual node-level hypotheses:

$$\mathbb{P}(FP \geq 1 \,|\, \mathbf{I}_0) = \mathbb{P}(\{\exists v \in \mathbf{I}_0 : \hat{\pi}_v < \eta^*\}) \leq \pi^*, \qquad (13)$$

where π^* is a user-defined rate (e.g. 0.01, 0.05). Henceforth, we consider the following null hypothesis:

$$\text{H}_{\text{null}} : p_v = q_v, \forall v \in V. \qquad (14)$$

Unlike *strong FWER control* that refers to any subset $\mathbf{I}_0 \subset V$, *weak FWER control* is less demanding as it deals only with the case where $\mathbf{I}_0 = V$. Weak control in MTST is particularly relevant when studying the behavior of complex systems under two different experimental conditions. When there is no statistically significant difference between the conditions, then all nodes are expected to satisfy the null hypothesis (Expr. 14). Neuroscience offers a good example of this setting: several sensors are used to monitor brain activity, and MTST aims at detecting clusters of firing neurons to a given stimulus. Classical methods, such as PCT (Maris and Oostenveld, 2007) and TFCE (Smith and Nichols, 2009), account for the inherent graph structure of the brain function and perform weak FWER control. By this, they mitigate the risk of reporting a false difference between two experimental conditions, while still maintaining the sensitivity necessary for detecting true neural activity patterns.

The graph regularization introduced by the collaborative ϕ-divergence estimation of Step 1 leads to robust estimators against outliers in the node-level test statistics. This is particularly relevant under H_{null} where we want to avoid false positives, thus it is natural to exploit this feature by designing a π-value estimation procedure with weak FWER control. Bear in mind that the flexibility of non-parametric LRE allows a certain level of heterogeneity of the pdfs in $\{p_v\}_{v\in V}$ and $\{q_v\}_{v\in V}$, as long as the relative likelihood-ratios of the pairs $((p_v, q_v)$ and (p_u, q_u), $(u,v) \in E)$ in adjacent nodes can be approximated by functions that are close in the shared RKHS. This feature complicates the distribution of the test statistic under H_{null} and, consequently, the derivation of an explicit formula for FWER control.

Algorithm 1 – Collaborative two-sample test (CTST)

Input: $\mathbf{X}, \mathbf{X}'$: two samples with observations over a given graph;
$\alpha \in [0,1)$: parameter of the relative likelihood-ratio;
n_{perm}: number of permutations for π-value computation;
π^*: FWER rate for the test required by the user.
Output: $\{\hat{\pi}_v\}_{v\in V}$, $\{\hat{\pi}'_v\}_{v\in V}$: a pair of π-values per node;
R_{CTST}: nodes with rejected null hypothesis $\text{H}_{\text{null},v} : p_v = q_v$.

■ **Produce the elements to define** $\hat{PE}^{\alpha}_1(\mathbf{X}, \mathbf{X}')$, $\hat{PE}^{\alpha}_2(\mathbf{X}', \mathbf{X})$
1: Select the hyperparameters $\sigma_1^*, \lambda_1^*, \gamma_1^*$, $\sigma_2^*, \lambda_2^*, \gamma_2^*$ (see Sec. C.1)
2: Find the anchor points given a kernel $\text{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (see Sec. 3.2)
■ **Compute node-level test statistics using observed data**
3: Estimate $\hat{\mathbf{\Theta}}_1(\mathbf{X}, \mathbf{X}') = \text{GRULSIF}(\mathbf{X}, \mathbf{X}', \alpha, \sigma_1^*, D_1, \gamma_1^*, \lambda_1^*)$,
$\hat{\mathbf{\Theta}}_2(\mathbf{X}', \mathbf{X}) = \text{GRULSIF}(\mathbf{X}', \mathbf{X}, \alpha, \sigma_2^*, D_2, \gamma_2^*, \lambda_2^*)$
4: Compute $S_v = \{\hat{PE}^{\alpha}_v(X_v \| X'_v)\}_{v\in V}$, $S'_v = \{\hat{PE}^{\alpha}_v(X'_v \| X_v)\}_{v\in V}$
using $\hat{\mathbf{\Theta}}_1(\mathbf{X}, \mathbf{X}')$ and $\hat{\mathbf{\Theta}}_2(\mathbf{X}', \mathbf{X})$ (see Expr. 12)
■ **Permutation test**
5: **for** $i \in \{1, \ldots, n_{\text{perm}}\}$ **do**
6: Generate a random permutation τ of the set such that
$\mathbf{X}^{(\tau)} = \{X_{:,\tau(1)}, \ldots, X_{:,\tau(n+n')}\}$
7: Assign the first n elements of $\mathbf{X}^{(\tau)}$ to the set $\dot{\mathbf{X}}$
and the rest n' to the set $\dot{\mathbf{X}}'$
8: Compute $\hat{\mathbf{\Theta}}_1(\dot{\mathbf{X}}, \dot{\mathbf{X}}')$ and $\hat{\mathbf{\Theta}}_2(\dot{\mathbf{X}}', \dot{\mathbf{X}})$
9: Compute $\{\hat{PE}^{\alpha}_v(\dot{\mathbf{X}}_v \| \dot{\mathbf{X}}'_v)\}_{v\in V}$ and $\{\hat{PE}^{\alpha}_v(\dot{\mathbf{X}}'_v \| \dot{\mathbf{X}}_v)\}_{v\in V}$
using $\hat{\mathbf{\Theta}}_1(\dot{\mathbf{X}}, \dot{\mathbf{X}}')$ and $\hat{\mathbf{\Theta}}_2(\dot{\mathbf{X}}', \dot{\mathbf{X}})$
10: Compute the test statistics $s_1^i = \max_v \{\hat{PE}^{\alpha}_v(\dot{\mathbf{X}}_v \| \dot{\mathbf{X}}'_v)\}_{v\in V}$,
$s_2^i = \max_v \{\hat{PE}^{\alpha}_v(\dot{\mathbf{X}}'_v \| \dot{\mathbf{X}}_v)\}_{v\in V}$
11: **end for**
12: **for** $v \in \{1, \ldots, N\}$ **do**
13: $\hat{\pi}_v = \frac{1}{n_{\text{perm}}} \sum_{i=1}^{n_{\text{perm}}} \mathbb{1}\{S_v \leq s_1^i\}$
14: $\hat{\pi}'_v = \frac{1}{n_{\text{perm}}} \sum_{i=1}^{n_{\text{perm}}} \mathbb{1}\{S'_v \leq s_2^i\}$
15: **end for**
16: ■ **Identify the nodes where the null hypothesis is rejected**
17: Define the set $R_{\text{CTST}} = \{v \in V \,|\, \hat{\pi}_v \leq \frac{\pi^*}{2}$ or $\hat{\pi}'_v \leq \frac{\pi^*}{2}\}$
18: **return** $\{\hat{\pi}_v\}_{v\in V}$, $\{\hat{\pi}'_v\}_{v\in V}$, R_{CTST} =0

⋆ Steps 7 & 8 use the implementation of de la Concha et al. (2024).

Another important point is the role of correlations in graph-structured MTST. Correlated test statistics $\{S_v\}_{v\in V}$ are expected due to possible dependencies between the node-level datasets $\{\mathcal{X}_v\}_{v\in V}$ and by the graph penalization term of the collaborative LRE. Ignoring such correlations in the design of a MT may lead to a loss of power, which is why this question is an active research topic.

To address the two issues, i.e. lacking a closed-form FWER expression and the correlation between test statistics, we propose a permutation test over the vectors $X_{:,j} = (x_{1,j}, \ldots, x_{N,j})^{\mathsf{T}}$ (resp. $X'_{:,j} = (x'_{1,j}, \ldots, x'_{N,j})^{\mathsf{T}}$), that keeps as a block the observations that have the same index, j, across all nodes. The permutation test infers the distribution of the maximum test statistic $S_G = \max_{v\in V} S_v$, and uses it to determine R_{CTST}, achieving this way weak FWER control at the level of the given π^*. This is in fact stated by Theorem 3.2, whose proof is provided in Appendix B. The complete CTST algorithm, when using GRULSIF for graph-based LRE, is provided in Alg. 1.

Theorem 3.2. *Consider Problem 1 and assume the observations* $\mathbf{X} = \{X_v\}_{v\in V}$ *are iid for each node* $v \in V$*, same for and* $\mathbf{X}' = \{X'_v\}_{v\in V}$ *(see Eq. 3). Let* $\dot{\mathbf{X}}$*,* $\dot{\mathbf{X}}'$ *the permuted datasets as described in Alg. 1 and* π^* *a user-defined rate. Let* $F(\cdot\,|\,\mathbf{X}\cup\mathbf{X}')$ *denote the probability distribution of* $S(\dot{\mathbf{X}}\|\dot{\mathbf{X}}') = \max_{v\in V} \hat{PE}^{\alpha}_v(\dot{X}_v\|\dot{X}'_v)$ *given* $\mathbf{X}\cup\mathbf{X}'$ *and let* $\hat{q}(\mathbf{X}\cup\mathbf{X}') = \sup\{s\in\mathbb{R}\,|\,F(s\,|\,\mathbf{X}\cup\mathbf{X}') \le 1-\frac{\pi^*}{2}\}$ *be the point determining the upper* $((1-\frac{\pi^*}{2})\cdot 100)$*-percentile. Then, if* H_{null} *is true, that is* $p_v = q_v$*,* $\forall v \in V$*, then it holds:*

$$\mathbb{P}(S > \hat{q}(\mathbf{X}\cup\mathbf{X}')) \le \frac{\pi^*}{2}. \tag{15}$$

Moreover, when $S' = \max_{v\in V} \hat{PE}^{\alpha}_v(\dot{X}'_v\|\dot{X}_v)$ *is used as a test statistic, then, under* H_{null} *we have:*

$$\mathbb{P}(S > \hat{q}(\mathbf{X}\cup\mathbf{X}')) \quad \textit{or} \quad S' > \hat{q}'(\mathbf{X}\cup\mathbf{X}')) \le \pi^*, \tag{16}$$

which implies $FWER(R_{\text{CTST}}) = \mathbb{P}(FP \ge 1\,|\,\mathrm{H}_{\text{null}}) \le \pi^*$.

CTST without graph structure. By replacing GRULSIF by its POOL variant, which uses the same estimation framework but neutralizes the graph component (de la Concha et al., 2024), we derive the reduced CTST-POOL variant (i.e. $W = \mathbf{0}_{N\times N}$ in Eq. 8). In other words, the nodes are allowed to collaborate only at the level of building the common RKHS. The CTST-POOL variant (henceforth referred to simply as POOL) can be relevant when there is no graph underlying the MT problem.

4 EXPERIMENTS

CTST is evaluated in the context of graph-structured MT, in synthetic and real scenarios. The goal is to show the gains of combining a non-parametric graph-based collaborative estimation of node-level test statistics, with the weak FWER control based on a permutation test with a maximum statistic. Note that CTST is not a direct competitor to methods such as SAHBA, PCT, or TFCE (see Sec. 1). In fact, those can be seen as complementary approaches to CTST, as they could post-process CTST's output, however, studying how to combine these approaches is beyond the scope of this paper. To make fair comparisons and keep the flexibility of non-parametric methods, we restrict our attention to estimation approaches built upon Kernel Methods. Tab. 1 mentions all the compared methods.

Each non-parametric method requires fixing the regularization constants and the hyperparameters of the kernel function; we focus on Gaussian kernels with width parameter σ. LRE-based methods use cross-validation to fix the hyperparameters, while from several works addressing this issue for MMD, we compare against the original MMD-MEDIAN version that is based on the median heuristic (Gretton et al., 2012), and the MMD-MAX method (Sutherland et al., 2017) that computes a score associated with the power of the two-sample test. Details on hyperparameter selection are provided in Appendix C.

For the competitors we follow the traditional MT approach: we first estimate node-level test statistics independently for each node, and then we control for the MCP using a non-parametric resampling test with a maximum statistic (Westfall and Young, 1992), which achieves weak FWER control. The node-level test statistics $\{S_v\}_{v\in V}$ coincide with the notion of dissimilarity measured by each method (ϕ-divergence or MMD), and the distribution of $S_G = \max_{v\in V} S_v$ under the H_{null} is estimated via a permutation test (see Alg. 1). We address the non-symmetricity of the ϕ-divergence-based methods same as we did in Sec. 3.3 for CTST, by comparing both ordered pairs p, q and q, p. Then, given a user-provided threshold rate π^*, we identify the sets of rejected hypotheses $R_{\phi\text{-div}} = \{v \in V\,|\,\hat{\pi}_v < \frac{\pi^*}{2}$ or $\hat{\pi}'_v < \frac{\pi^*}{2}\}$ and $R_{\text{MMD}} = \{v \in V\,|\,\hat{\pi}_v < \pi^*\}$.

Each instance of the four designed fully synthetic scenarios is generated by first generating a random graph and then by defining the scheme of the occurring change over a subset of the nodes:

- *Synth.Ia&b* use a *Stochastic Block Model* (SBM) with 4 clusters, with 25 nodes each (intra-cluster edge probability: 0.5; inter-cluster edge probability: 0.01). Then, the same behavior (change of measure or not) is set cluster-wise for all the nodes in each cluster, $C_1, ..., C_4$.
- *Synth.IIa&b* use a *Grid* graph (GRID) with 100 nodes forming a 10×10 regular tiling. In this case, an ego-network-based scheme is employed, which picks a node u at random, with probability proportional to its node degree, and then considers that only the nodes in u's 2-hop ego-network, denoted simply as $C(u)$, shall experience a change of measure.

4.1 Synthetic experiments

Synthetic scenarios provide by design the set $\mathbf{I}_0^{\text{c}} = V\backslash\mathbf{I}_0$, which is the indexes v's where $p_v \neq q_v$, hence allow the comparison of the power of the different MT approaches. The scenarios detailed in Tab. 2 are similar to those in de la Concha et al. (2024) to satisfy the graph smoothness hypothesis (connected nodes have similar behavior), and to pose various challenges. On the top of each of those scenarios, we build a two-sample test comparing p_v vs. q_v. The two pdfs may differ in terms of mean, shape, covariance, etc. (see the node-level hypotheses in Tab. 2). Moreover, there can be more than one type of change in the same scenario.

We measure the performance of a MT approach along two axes: First, the efficiency of its FWER control,

Table 1: List of competitors. All the methods that are included in our experimental evaluation study for the graph-structured multiple two-sample test problem. 'l-r.' indicates the method that estimates the non-regularized likelihood-ratio ($\alpha = 0$).

Method	Reference	Estimate	Sim. measure	Graph
KLIEP	Sugiyama et al. (2007)	l-r.	KL-divergence	No
LSTT	Sugiyama et al. (2011a)	l-r.	χ^2-divergence	No
RULSIF	Yamada et al. (2013)	relative l.-r.	χ^2-divergence	No
MMD	Gretton et al. (2012)	MMD	MMD	No
POOL	this work	relative l-r.	χ^2-divergence	No
CTST	this work	relative l-r.	χ^2-divergence	Yes

Table 2: Synthetic experiments. The scenarios are defined by the graph structure they employ and the node-level distributions (p_v and q_v) generating the data observations at each node. '•' denotes cases where distributions or their parameters remain unchanged.

		Node-level hypotheses		
Experiment	Location	p_v	vs.	q_v
Synth.Ib SBM 4 clusters	$v \in C_1$	$\mathrm{N}(\mu{=}0, \sigma{=}1)$	vs.	$\mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$
	$v \in C_2 \cup C_3$	$\mathrm{N}(\mu{=}0, \sigma{=}1)$	vs.	•
	$v \in C_4$	$\mathrm{N}(\mu{=}0, \sigma{=}1)$	vs.	$\mathrm{N}(\mu{=}1, \sigma{=}\bullet)$
Synth.Ib SBM 4 clusters	$v \in C_1 \cup C_2$	$\mathrm{N}(\mu{=}(0,0)^\intercal, \Sigma_{1,2}{=}-\frac{4}{5})$	vs.	•
	$v \in C_3$	$\mathrm{N}(\mu{=}(0,0)^\intercal, \Sigma_{1,2}{=}\frac{4}{5})$	vs.	$\mathrm{N}(\mu{=}\bullet, \Sigma_{1,2}{=}0)$
	$v \in C_4$	$\mathrm{N}(\mu{=}(0,0)^\intercal, \Sigma_{1,2}{=}0)$	vs.	$\mathrm{N}(\mu{=}(1,1)^\intercal, \Sigma_{1,2}{=}\bullet)$
Synth.IIa GRID 10×10	$v \in C(u)$	$\mathrm{N}(\mu{=}\mathbf{0}_3, \Sigma_{i,i}{=}1, \Sigma_{1,2}{=}\frac{4}{5}, \Sigma_{3,1}{=}0)$	vs.	$\mathrm{N}(\mu{=}\bullet, \Sigma_{i,i}{=}\bullet, \Sigma_{1,2}{=}0, \Sigma_{3,1}{=}\bullet)$
	$v \notin C(u)$	$\mathrm{N}(\mu{=}\mathbf{0}_3, \Sigma_{i,i}{=}1, \Sigma_{1,2}{=}\frac{4}{5}, \Sigma_{3,1}{=}0)$	vs.	•
Synth.IIb GRID 10×10	$v \in C(u)$	$\mathrm{N}(\mu{=}(0,0)^\intercal, \Sigma{=}10I_2)$	vs.	Gaussian Mixture $\binom{\text{with equal}}{\text{proportion}}$ $\mathrm{N}(\mu_1{=}(\ 0,\ \ 0)^\intercal, \Sigma{=}5I_2)$ $\mathrm{N}(\mu_2{=}(\ 0,\ \ 5)^\intercal, \Sigma{=}5I_2)$ $\mathrm{N}(\mu_3{=}(\ 0, -5)^\intercal, \Sigma{=}5I_2)$ $\mathrm{N}(\mu_4{=}(\ 5,\ \ 0)^\intercal, \Sigma{=}5I_2)$ $\mathrm{N}(\mu_5{=}(-5,\ \ 0)^\intercal, \Sigma{=}5I_2)$
	$v \notin C(u)$	$\mathrm{N}(\mu{=}(0,0)^\intercal, \Sigma{=}10I_2)$	vs.	•

i.e. the probability to occur one or more false positives under the H_{null} of Eq. 14. Second, how informative the estimated node-level π-values are, i.e. whether the low π-values are associated with nodes in $\mathbf{I}_0^c$. From a practitioner's perspective, when comparing a complex system across two different time-stamps or experimental conditions, methods that are robust to false positives (avoid asserting a statistically non-existent difference) are preferred. Second, we measure how accurately the MT procedure identifies the nodes responsible for an observed deviation; this is summarized by the Alternative Free-response Receiver-Operating Characteristic (AFROC) curve (Chakraborty and Winter, 1990). The detailed estimation we used for the AFROC curves is provided in Appendix C.

The AUC of the AFROC curves is reported in Tab. 3. The higher the value of the AUC the better, indicating that a method achieved the required FWER level of $\pi^* = 0.05$ and is still able to identify the nodes in $\mathbf{I}_0^c$. The AFROC curves we designed ignore the false positives at nodes $\{v \,|\, v \in R_{\text{MT}} \cap \mathbf{I}_0\}$. For this reason, we report also the AUC of the ROC curves. The interpretation should take AFROC-AUC as the most important criterion, and ROC-AUC rather as a tiebreaker for approaches with similar AFROC-AUC.

Findings. Tab. 3 shows that CTST is more efficient compared to the other methods that disregard the geometry of the problem. The role of the graph becomes stronger as the observations are fewer, and when the difference between p_v and q_v is more subtle. This effect is more evident when comparing CTST to the no-graph variant POOL. An additional advantage of CTST over POOL is that it is robust and consistent when varying the regularization parameter α (see Appendix C).

Table 3: Results on synthetic scenarios. Non-parametric methods applied on multiple two-sample testing over a known graph. Keeping the graph fixed, the AFROC and ROC curves were computed over 1000+1000 experiment instances generated over H_{null} and H_{alt} of Problem 1, respectively. Higher AUC values are better.

		$n = n' = 50$		$n = n' = 100$		$n = n' = 250$	
Experiment	Method	AFROC AUC	ROC AUC	AFROC AUC	ROC AUC	AFROC AUC	ROC AUC
Synth.Ia	CTST α=0.1	**0.50**	**0.93**	**0.66**	**0.99**	**0.99**	**1.00**
	POOL α=0.1	0.28	0.85	0.49	0.93	0.64	0.99
	RULSIF α=0.1	0.18	0.88	0.47	0.95	0.76	**1.00**
	LSTT	0.07	0.84	0.38	0.91	0.23	0.76
	KLIEP	0.00	0.74	0.35	0.89	0.55	**1.00**
	MMD-MEDIAN	0.33	0.82	0.50	0.89	0.54	0.97
	MMD-MAX	0.33	0.82	0.50	0.88	0.54	0.97
Synth.Ib	CTST α=0.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.1	0.72	**1.00**	0.99	**1.00**	**1.00**	**1.00**
	RULSIF α=0.1	0.44	0.97	0.83	0.88	0.94	0.95
	LSTT	0.36	0.94	0.77	0.91	0.96	0.96
	KLIEP	0.33	0.90	0.79	0.94	0.92	0.93
	MMD-MEDIAN	0.48	0.96	0.52	0.99	0.96	**1.00**
	MMD-MAX	0.48	0.96	0.52	0.99	0.96	**1.00**
Synth.IIa	CTST α=0.1	**0.94**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.1	0.18	0.98	0.84	**1.00**	**1.00**	**1.00**
	RULSIF α=0.1	0.01	0.82	0.30	0.99	0.52	0.61
	LSTT	0.00	0.81	0.23	0.83	0.97	**1.00**
	KLIEP	0.00	0.80	0.29	0.91	0.67	0.73
	MMD-MEDIAN	0.00	0.81	0.01	0.95	0.43	**1.00**
	MMD-MAX	0.00	0.82	0.01	0.95	0.39	**1.00**
Synth.IIb	CTST α=0.1	**0.30**	**0.92**	**0.65**	**0.98**	**0.98**	**1.00**
	POOL α=0.1	0.02	0.84	0.12	0.95	0.78	**1.00**
	RULSIF α=0.1	0.01	0.80	0.06	0.92	0.75	**1.00**
	LSTT	0.00	0.78	0.04	0.91	0.66	**1.00**
	KLIEP	0.00	0.79	0.03	0.85	0.63	**1.00**
	MMD-MEDIAN	0.00	0.78	0.05	0.92	0.60	**1.00**
	MMD-MAX	0.00	0.78	0.05	0.92	0.52	**1.00**

4.2 Two-sample testing on real seismic data

We use public seismic data to showcase CTST's potential in performing spatial statistical analysis. However, this should not be interpreted as an attempt to outperform state-of-the-art methods in that field. Geological hazard monitoring systems comprises several stations strategically positioned across a territory to monitor ground noise and shaking. When a seism occurs, it travels through the earth, and this is captured by the monitoring sensors. Stations closer to the epicenter of a seism tend to exhibit higher response to the event, i.e. faster reactions and more pronounced differences in their pre- and post-event data. In this context, a graph-structured multiple two-sample test can be used for assessing the significance of a seismic event, and for identifying the stations and time periods during which each of them got activated.

Data preprocessing: We analyze public data related to two seismic events occurred in New Zealand. Seism A is of magnitude 5.5 in Richter scale, occurred on May 31, 2021. Seism B is a weaker seism of mag-

nitude 2.6, occurred on Oct 2, 2023[2]. The stations are equipped with strong-motion accelerometers that provide 3d signals corresponding to the shaking across three perpendicular directions.

To compare, we analyze the waveforms from 50 seconds before to 50 seconds after the event, at 100 Hz frequency. The preprocessing details are in Appendix C.

Graph structure: We build in two steps a graph representation that accounts for both spatial and temporal similarities between the seismic stations and their signals. The first step is to build an unweighted *spatial graph* $G_{\mathrm{S}} = (V, E, W)$ considering as nodes the stations whose all accelerometers have available data at the analyzed time period. Edges are drawn from each station to its geographical 3-nearest-neighbors. Subsequently, we integrate the temporal dimension by building a *multiplex graph* $G_{\mathrm{S}\times\mathrm{T}}$ over G. We segment the signal before and after the seismic event in 10 time-windows, each containing the same amount of observations. $G_{\mathrm{S}\times\mathrm{T}}$ indexes the nodes of G_{S} by the time-window, $V \times \mathcal{T}$, where $\mathcal{T} = \{1, ..., 10\}$. Two nodes in $G_{\mathrm{S}\times\mathrm{T}}$, (u,t) and (v,t'), are connected: i) if $t = t'$ and $(u,v) \in E$, i.e. they refer to the same time-window and the nodes u and v are connected in the spatial graph G_{S}, ii) or if $u = v$ and $|t' - t| = 1$, i.e. each node $v \in V$ is connected to its 'copies' in the two adjacent time-windows.

The observations are indexed by the node and the time-window they belong, so the pdfs $\{p_{(v,t)}\}_{(v,t)\in V\times T}$ and $\{q_{(v,t)}\}_{(v,t)\in V\times T}$. After the preprocessing, we obtain two samples for each pair (v,t), $X_{(v,t)} = \{x_{((v,t),i)}\}_{i=1}^{100} \sim p_{(v,t)}$ and $X'_{(v,t)} = \{x'_{((v,t),i)}\}_{i=1}^{100} \sim q_{(v,t)}$. We denote by $t = 0$ the beginning of the sample, that is 50 seconds before the seism. Then, the set $X_{(v,1)}$ refers to the first 5 seconds of preprocessed observations after $t = 0$ and $X'_{(v,1)}$ the 5 seconds of preprocessed observations after the event. Under this configuration, a two-sample test aims to identify the pairs (v,t) where $p_{(v,t)} \neq q_{(v,t)}$.

Findings. Fig. 2 concerns Seism A and the results obtained by CTST and some of the kernel-based two-sample tests listed in Tab. 1. The rest of the results are in Appendix C.2.3. The figures highlight the largest cluster of $C_{\mathrm{S}\times\mathrm{T}}$ of $G_{\mathrm{S}\times\mathrm{T}}$ made of nodes whose estimated π-values is smaller than 0.05. All the tested methods detect correctly an occurring seismic event and identify the most sensitive nodes as those closer to the epicenter. However, the methods that do not account for the graph structure, which here encodes the expected spatial and temporal similarities between stations, lead to results where detections seem not informative. For example, looking at the associated π-values, the effect of a seism takes longer to fade out even when it ceases to be visible in the signals. Contrary, CTST recovers most of the nodes closer to the epicenter and follows better the evolution of the seismic event. These findings are compatible with the results found in the synthetic experiments, as the AFROC-AUC and ROC-AUC measures show that CTST is more robust to false alarms, and that it recovers the nodes of interest with a higher confidence when the assumption of graph smoothness of the likelihood-ratios is satisfied.

[2]Data available by the GeoNet project (GNS Science, 1970):
Seism A: https://www.geonet.org.nz/earthquake/2021p405872
Seism B: https://www.geonet.org.nz/earthquake/2023p741652

5 CTST IN PRACTICE

CTST requires three elements adapted to a problem at hand: a data preprocessing pipeline, an RKHS ($\mathbb{H}$), and a graph G to be used for estimation. These elements should satisfy the following points:

1. The observations at all nodes should belong to the same input space $\mathcal{X}$, hence having the same dimension d, and also be of comparable scale across nodes.
2. The RKHS should be rich enough to approximate, as much as we want, each of the relative likelihood-ratios. More concretely, the α-regularization implies that for all $v \in V$, $r_v^\alpha \in \mathcal{L}^2(P_v^\alpha)$, i.e. it belongs to the space of square-integrable functions w.r.t. P_v^α. Moreover, estimating r_v^α by capitalizing on the variational formulation of the χ^2-divergence (Eq. 5) is equivalent to approximating r_v^α in terms of the distance $\mathcal{L}^2(P_v^\alpha)$. It is therefore clear that a RKHS dense in $\mathcal{L}^2(P_v^\alpha)$ (w.r.t. to the norm $\|\cdot\|_{\mathcal{L}^2(P_v^\alpha)}$) will, in theory, be able to approximate r_v^α as much as we desire. In the setting presented in this paper, where $\mathcal{X} \subset \mathbb{R}^d$, it can be shown that translation-invariant kernels (e.g. Gaussian, Laplacian, and the Matérn class of kernels) are universal kernels that can approximate any function in $\mathcal{L}^2(P_v^\alpha)$.
3. For graph regularization to be beneficial, the graph G should depict well how similar the statistical tests are expected to be at adjacent nodes. Practitioners should aim at having a good prior intuition about this and encode it in an adjacency matrix. In many contexts, relevant graph estimates can be provided by the problem itself, as in the examples included in the paper, or in fields like neuroscience.
4. Finally, the input dimension d affects the computational complexity of CTST. In this work, we built the dictionaries using a simple greedy method that minimizes the linear dependency between dictionary elements (Richard et al., 2009). For some kernels, higher values of d result in larger final dictionaries that increase the computational complexity of GRULSIF (see Sec. 3.2). As the permutation test requires multiple estimations, it is important to seek a common representation for the data in as lower as possible dimensional space.

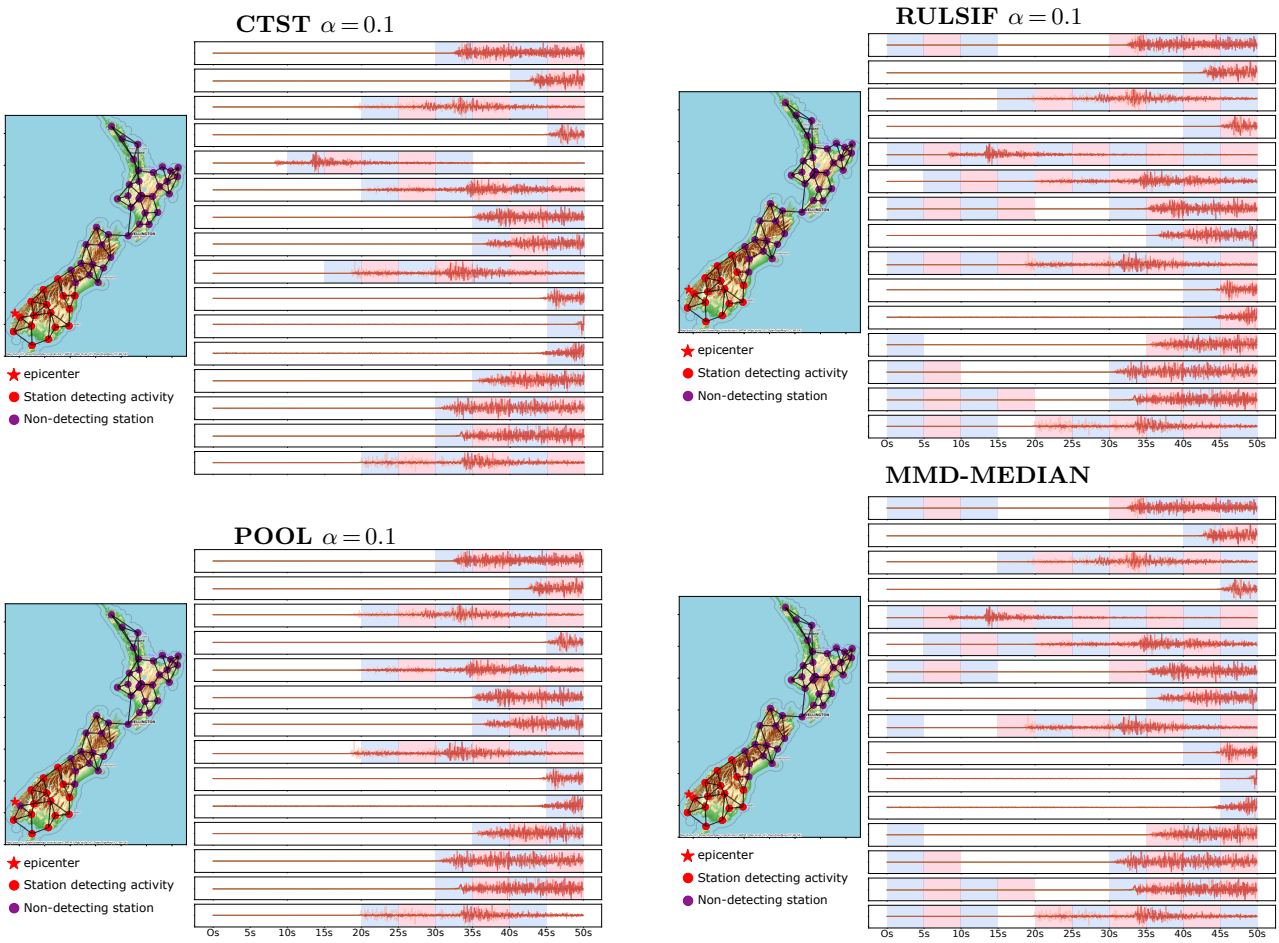


Figure 2: Results of seismic activity detection for Seism A. Four methods are compared. For each method, on the left side, there appear the locations of the stations on the map of New Zealand, connected in a 3NN graph; on the right side, there are the post-event signals associated with the stations (in proximity order to the epicenter) detecting seismic activity in at least one time-window (such time-windows appear in either light blue or pink color). **Left column:** Results of the proposed graph-based CTST and the POOL variant that ignores the graph structure (both with $\alpha = 0.1$). **Right column:** Results of two kernel-based two-sample tests fitted individually and independently at each node.

6 CONCLUSIONS

We introduced CTST, a novel collaborative non-parametric testing framework, designed for multiple two-sample testing over the nodes of a graph. Our approach integrates advances in collaborative likelihood-ratio estimation to compute jointly node-level test statistics and identify null hypotheses to be rejected, under a graph smoothness hypothesis. CTST can deal with scenarios where the data at each node is multivariate, and the probabilistic models are unknown and may differ from node to node. The only two required conditions are that the joint likelihood-ratios can be approximated by the same RKHS, and that adjacent nodes have similar likelihood-ratios with respect to the norm of the RKHS. The merits of this framework are demonstrated in synthetic experiments, and in a real-world use-case concerning seismic activity detection.

Acknowledgments

The authors acknowledge the support of the Industrial Data Analytics and Machine Learning Chair hosted at ENS Paris-Saclay.

References

Agrawal, R. and Horel, T. (2021). Optimal Bounds between f-Divergences and Integral Probability Metrics. *Journal of Machine Learning Research*, 22(128):1–59.

Bach, F. (2024). Sum-of-Squares Relaxations for Information Theory and Variational Inference. *Foundations of Computational Mathematics*.

Bargiotas, I., Kalogeratos, A., Limnios, M., Vidal, P.-P., Ricard, D., and Vayatis, N. (2021). Revealing posturographic profile of patients with parkinsonian syndromes through a novel hypothesis testing framework based on machine learning. *PLOS ONE*, 16(2).

Beck, A. and Tetruashvili, L. (2013). On the convergence of

block coordinate descent type methods. *SIAM Journal on Optimization*, 23:2037–2060.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (2010). ObsPy: A Python Toolbox for Seismology. *Seismological Research Letters*, 81(3):530–533.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Rey-Bellet, L. (2022a). (f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. *Journal of Machine Learning Research*, 23(39):1–70.

Birrell, J., Katsoulakis, M. A., and Pantazis, Y. (2022b). Optimizing Variational Representations of Divergences and Accelerating Their Statistical Estimation. *IEEE Transactions on Information Theory*, 68(7):4553–4572.

Broniatowski, M. and Keziou, A. (2006). Minimization of ϕ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442.

Chakraborty, D. P. and Winter, L. H. (1990). Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, 174(3):873–881.

Chen, Y., Wang, T., and Samworth, R. (2021). High-dimensional, multiscale online changepoint detection. *Journal of Royal Statistical Society, Ser. B., to appear*.

Csiszár, I. (1967). On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 2:329—-339.

de la Concha, A., Kalogeratos, A., and Vayatis, N. (2024). Collaborative likelihood-ratio estimation over graphs.

de la Concha, A., Vayatis, N., and Kalogeratos, A. (2023). Online non-parametric likelihood-ratio estimation by Pearson-divergence functional minimization.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

GNS Science (1970). GeoNet Aotearoa New Zealand Earthquake Catalogue.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Harchaoui, Z., Bach, F., Cappe, O., and Moulines, E. (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine*, 30(4):87–97.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley, New York.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley.

Li, A. and Barber, R. F. (2018). Multiple Testing with the Structure-Adaptive Benjamini–Hochberg Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):45–74.

Li, X., Zhao, T., Arora, R., Liu, H., and Hong, M. (2018). On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, 18(184):1–24.

Lopez-Paz, D. and Oquab, M. (2017). Revisiting classifier two-sample tests. In *Int. Conf. on Learning Representations*.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190.

Nguyen, X., Wainwright, M. J., and Jordan, M. (2008). Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. on Information Theory*, 56(11):5847–5861.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Richard, C., Bermudez, J. C. M., and Honeine, P. (2009). Online prediction of time series data with kernels. *IEEE Trans. on Signal Processing*, 57(3):1058–1067.

Sheldon, D. (2008). Graphical Multi-Task Learning. Technical report, Cornell University.

Smith, S. and Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, 44(1):83–98.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 20.

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011a). Least-squares two-sample test. *Neural Networks*, 24(7):735–751.

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011b). Least-squares two-sample test. *Neural networks : the official journal of the International Neural Network Society*, 24:735–51.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *Int. Conf. on Learning Representations*.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

Westfall, P. H. and Young, S. S. (1992). *Resampling-based multiple testing*. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. (2011). Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, in each theorem, we state the mathematical assumptions of the problem.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, this is for example discussed at the end of Sec. 3.2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, code will be made public available after publication.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, both Theorem 3.1 and Theorem 3.2 state clearly the required assumptions.
   (b) Complete proofs of all theoretical results. Yes, the detailed proofs are provided in Appendix A and B.
   (c) Clear explanations of any assumptions. Yes, the interest of each assumption is described through the text.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, all these elements are available in the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, these elements are discussed in Appendix C.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, all these elements are available in the supplementary material.
   (d) A description of the computing infrastructure used. Not Applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes, we provide the specific ULR to the real-data set used in the paper.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes,
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

A VARIATIONAL FORMULATION OF χ^2-DIVERGENCE

ϕ-divergences are a useful mathematical tool as they offer a unified perspective on various statistical problems. One of the most prominent example is the canonical maximum likelihood estimation procedure, which can be interpreted as minimizing the KL-divergence between the empirical distribution of the observed data and a distribution selected from a parametric family. This duality of ϕ-divergence as a tool to measure statistical similarity and as a tool to build and train probabilistic models is made explicit in the variational representation formulas of ϕ-divergences. Deriving valid variational representations is an active research topic (Broniatowski and Keziou, 2006; Nguyen et al., 2008, 2010; Agrawal and Horel, 2021; Birrell et al., 2022a,b; Bach, 2024) and there are multiple formulations built using different assumptions. Theorem 3.1 is a consequence on the following result of the literature.

Theorem A.1. *(Theorem 4.3 in Broniatowski and Keziou (2006)). Let $\mathcal{F}$ be some class of measurable real-valued functions defined on $\mathcal{X}$, $\mathcal{B}$ the set of all bounded measurable real-valued functions defined on $\mathcal{X}$, and by* $\text{span}(\mathcal{F}\cup\mathcal{B})$ *the set of all linear combinations of the set $\mathcal{F}\cup\mathcal{B}$ and define the set:*

$$\mathcal{M}_{\mathcal{F}} = \left\{ Q \in \mathcal{M} \;\middle|\; \int |f|\, d|Q| < \infty,\ \forall f \in \mathcal{F} \right\}, \tag{17}$$

where $|Q|$ denotes the total variation of the signed finite measure Q and $\mathcal{M}(\mathcal{X})$ refers to the space of all finite signed measures.

Assume that ϕ is differentiable. Then, for all $Q \in \mathcal{M}_{\mathcal{F}}$, such that $\mathcal{D}_\phi(P\|Q)$ is finite and $\phi'(\frac{dQ}{dP})$ belongs to $\text{span}(\mathcal{F}\cup\mathcal{B})$*, $\mathcal{D}_\phi(P\|Q)$ admits the dual representation:*

$$\mathcal{D}_\phi(P\|Q) = \sup_{g \in \text{span}(\mathcal{F}\cup\mathcal{B})} \int g(x')dQ(x') - \int \phi^\star(g)(x)\, dP(x) \tag{18}$$

where $\phi^\star$ denotes the convex conjugate of ϕ. And the function $g^* = \phi'(\frac{dQ}{dP})$ is a dual optimal solution. Furthermore, if ϕ is essentially smooth, then g^* is the unique dual solution P-a.e.

Proof of Theorem 3.1.

The χ^2-divergence refers to the case where $\phi(\zeta) = \frac{(\zeta-1)^2}{2}$, which is strictly convex around 1 and essentialy smooth. Its convex conjugate is given by $\phi^*(s) = \frac{s^2}{2} + s$. Furthermore, as r^α is a bounded function, we can conclude that $PE(P^\alpha\|Q)$ is bounded as well.

As $Q \in \mathcal{P}(\mathcal{X})$, then $|Q| = Q$ and $Q \in \mathcal{M}_{\mathbb{H}}$. To verify the last point, take $f \in \mathbb{H}$, then:

$$\int |f(x')|\, dQ(x') = \int |\langle f, \mathrm{K}(x', \cdot)\rangle_{\mathbb{H}}|\, dQ(x') \leq \kappa \|f\|_{\mathbb{H}} < \infty,$$

where the first equality is given by the representer property of $\mathbb{H}$, and the second one is a consequence of the Cauchy-Schwarz inequality.

The upper-bound on r^α implies that the function $\phi'(r^\alpha(x)) = r^\alpha(x) - 1$ belongs to $\text{span}(\mathbb{H}\cup\mathcal{B})$. Then, all the requirements of Theorem A.1 are satisified and we obtain the following expression:

$$PE(P^\alpha\|Q) = \sup_{g \in \text{span}(\mathbb{H}\cup\mathcal{B})} \int g(x')dQ(x') - \int \left[\frac{g^2(x)}{2} + g(x) \right] dP^\alpha(x), \tag{19}$$

which admits as unique optimal solution $g^* = r^\alpha(x) - 1$. Let us rewrite Eq. 19 in terms of functions of the form $g = f - 1$:

$$\begin{aligned} PE(P^\alpha\|Q) &= \sup_{f \in \text{span}(\mathbb{H}\cup\mathcal{B})} \int (f(x') - 1)dQ(x') - \int \left[\frac{(f-1)^2(y)}{2} + (f(y) - 1) \right] dP^\alpha(y) \\ &= \sup_{f \in \text{span}(\mathbb{H}\cup\mathcal{B})} \int f(x')dQ(x') - \int \frac{f^2(y)}{2}\, dP^\alpha(y) - \frac{1}{2} \\ &\geq \sup_{f \in \mathbb{H}} \int f(x')dQ(x') - \int \frac{f^2(y)}{2}\, dP^\alpha(y) - \frac{1}{2}. \end{aligned} \tag{20}$$

■

B REGARDING CTST AND FWER CONTROL

Bellow, we provide the proof for Theorem 3.2, which validates that CTST achieves weak FWER control.

Proof of Theorem 3.2. We start with the assumption that the observations of $\mathbf{X}$ come from the joint pdf $\mathbf{p}$, whose marginals are the node pdfs $\{p_v\}_{v\in V}$. Same for those observations of $\mathbf{X}'$ collected from the joint pdf $\mathbf{q}$ whose marginals are the node pdfs $\{q_v\}_{v\in V}$. We assume the observations at a specific node v, namely $\{x_{v,i}\}_{\forall v\in V, i=1,...,n}$ and $\{x'_{v,i}\}_{\forall v\in V, i=1,...,n'}$, are iid over the variation of index i. Let us define the set of vectors as $\mathbf{Z}=\{Z_1,Z_2,...,Z_{n+n'}\}$, where $Z_i=X_{:,i}=\{x_{v,i}\}_{v\in V}$ for $i\in\{1,...,n\}$ and $Z_{n+j}=X'_{:,j}=\{x'_{v,j}\}_{v\in V}$ for $j\in\{1,...,n'\}$. Then, under H_{null} and the hypothesis of statistical independence, we have that the probability distribution $\mathbf{p}^z$ is exchangeable, where exchangeability means that for any permutation τ on $\{1,...,n+n'\}$, the permuted set of vectors $\mathbf{Z}_\tau=\{Z_{\tau(1)},Z_{\tau(2)},...,Z_{\tau(n+n')}\}$ follow the same law $\mathbf{p}^z$.

Given a permutation τ we assign the first n elements of $\mathbf{Z}_\tau$ to the set $\dot{\mathbf{X}}$ and the remaining n' to the set $\dot{\mathbf{X}}'$. Denote by $F(\cdot\,|\,\mathbf{X}\cup\mathbf{X}')$ the distribution of the scores $S=\max_{v\in V}\hat{PE}_v^\alpha(\dot{\mathbf{X}}_v\|\dot{\mathbf{X}}'_v)$ conditioned on $\mathbf{X}\cup\mathbf{X}'$, and let $\hat{q}(\mathbf{X}\cup\mathbf{X}')=\sup\{s\in\mathbb{R}\,|\;F(s\,|\,\mathbf{X}\cup\mathbf{X}')\leq 1-\frac{\pi^*}{2}\}$. Then, under H_{null}, the echangeability property implies:

$$\mathbb{P}(S>\hat{q}(\mathbf{X}\cup\mathbf{X}'))\;=\mathbb{E}_{\mathbf{X}\cup\mathbf{X}'}\Big[\mathbb{P}(S>\hat{q}(\mathbf{X}\cup\mathbf{X}')\,|\,\mathbf{X}\cup\mathbf{X}')\Big]\leq\mathbb{E}_{\mathbf{X}\cup\mathbf{X}'}\Big[1-F(\hat{q}(\mathbf{X}\cup\mathbf{X}')\,|\,\mathbf{X}\cup\mathbf{X}')\Big]\leq\frac{\pi^*}{2}. \tag{21}$$

In a similar manner, we can verify that for $S'=\max_{v\in V}\hat{PE}_v^\alpha(\dot{\mathbf{X}}'_v\|\dot{\mathbf{X}}_v)$:

$$\mathbb{P}(S'>\hat{q}'(\mathbf{X}\cup\mathbf{X}'))\leq\frac{\pi^*}{2}. \tag{22}$$

By putting together both inequalities, we can conclude:

$$\mathbb{P}(S>\hat{q}(\mathbf{X}\cup\mathbf{X}')\;\;\text{or}\;\;S'>\hat{q}'(\mathbf{X}\cup\mathbf{X}'))\leq\mathbb{P}(S>\hat{q}(\mathbf{X}\cup\mathbf{X}'))+\mathbb{P}(S'>\hat{q}'(\mathbf{X}\cup\mathbf{X}'))\leq\pi^*. \tag{23}$$

And weak control over FWER comes from:

$$\begin{aligned}FWER(R_{\text{CMT}})&=\mathbb{P}(\{\exists v: S_v>\hat{q}(\mathbf{X}\cup\mathbf{X}')\text{ or }S'_v>\hat{q}'(\mathbf{X}\cup\mathbf{X}')\}|\,\mathrm{H}_{\text{null}})\\&\leq\mathbb{P}(\{\exists v: S_v>\hat{q}(\mathbf{X}\cup\mathbf{X}')\}|\,\mathrm{H}_{\text{null}})+\mathbb{P}(\{\exists v: S'_v>\hat{q}'(\mathbf{X}\cup\mathbf{X}')\}|\,\mathrm{H}_{\text{null}})\\&=\mathbb{P}(S>\hat{q}(\mathbf{X}\cup\mathbf{X}')\,|\,\mathrm{H}_{\text{null}})+\mathbb{P}(S'>\hat{q}'(\mathbf{X}\cup\mathbf{X}')\,|\,\mathrm{H}_{\text{null}})\leq\pi^*.\end{aligned} \tag{24}$$

■

C FURTHER DETAILS ABOUT THE EXPERIMENTS

In this section, we give more details on the implementation of the experimental setting described in the main text. Mainly:

1. More details on how the hyperparameters of GRULSIF and the other methods were chosen.
2. Elements to complement the results on the synthetic scenarios. This includes the way the AFROC and the ROC curves were estimated, and a detailed discussion about on the role of the regularization parameter α used by CTST and POOL.
3. Further details on the real-world example, including the preprocessing pipeline and the figures comparing the different multiple hypothesis testing settings.

C.1 Details regarding hyperparameters selection

For RULSIF and ULSIF algorithms, we follow the descriptions of Sugiyama et al. (2011b) and Yamada et al. (2011), and the hyperapemeters are selected independently for each of the nodes. We run a leave-one-out cross-validation procedure over the parameter associated with the Gaussian kernel and the penalization term γ. The parameter σ is selected from the grid $\{0.6\sigma_{\text{median}},0.8\sigma_{\text{median}},1\sigma_{\text{median}},1.2\sigma_{\text{median}},1.4\sigma_{\text{median}}\}$ where σ_{median} is the parameter σ found via the median heuristic over the observations in X'_v. On the other hand, the penalization parameter γ is optimized from the grid $\{1e^{-5},1e^{-3},0.1,10\}$. The procedure for KLIEP is similar, but we use instead a 5-fold cross-validation procedure.

For MMD median and MMD max, we identify the hyperparameters independently for each of the nodes, we follow the guidelines given in Gretton et al. (2012); Sutherland et al. (2017), respectively.

Finally, for CTST and the POOL algorithms, we apply 5-fold cross-validation to select the hyperparameters σ, γ, and λ using the implementation of de la Concha et al. (2024). Since the POOL approach ignores the graph structure, we fix $\lambda = 1$, and the penalization term related with the norm of each functional f_v will depend only on the parameter γ. In order to select the width σ for the Gaussian kernel, we first compute $\{\sigma_v\}_{v\in V}$ for each node via the median heuristic applied to the observations of X_v (such quantities are available when generating the dictionary), and we define $\sigma_{\min} = \text{argmin}\{\sigma_v\}_{v\in V}$, $\sigma_{\text{median}} = \text{median}\{\sigma_v\}_{v\in V}$ and $\sigma_{\max} = \text{argmax}\{\sigma_v\}_{v\in V}$; we then chose the final parameter from the set $\{\sigma_{\min}, \frac{1}{2}(\sigma_{\min}+\sigma_{\text{median}}), \sigma_{\text{median}}, \frac{1}{2}(\sigma_{\max}+\sigma_{\text{median}}), \sigma_{\max}\}$. γ is selected from the set $\{1e^{-5}, 1e^{-3}, 0.1, 1\}/c$, where $c = \sqrt{\min(n,n')}$ for POOL and $c = \sqrt{\min(n,n')}$ and $c = \sqrt{\min(n,n')N}$ for CTST. Finally, we identify the optimal λ^* from the set $\{1e^{-3}, 1e^{-2}, 0.1\cdot\frac{1}{c}, 1, 10\}/\sqrt{\min(n,n')N}$.

C.2 Details regarding synthetic scenarios

C.2.1 AFROC and ROC curves

The *Alternative Free-response Receiver Operating Characteristic* (AFROC) curve is an important tool in the context of multiple hypothesis testing, especially in fields where the practitioner seeks a decision to a global problem while requiring correct localization for true positive events (Chakraborty and Winter, 1990). In our context, AFROC allow us to quantify to which extent the compared methods achieve *Family-wise False Positive Rate* (FWER) control under the null hypothesis that all nodes $p_v = q_v$ (see Sec. 3.4 and the H_{null} in Eq. 3), while still being sensitive enough to identify those nodes where $p_v \neq q_v$ (H_{alt}). For each of the synthetic experiments described in Tab. 3, the given input graph G according to the scenario being studied is kept fixed (see Tab. 2), and then the axis of the AFROC curves for the experiments are estimated as follows:

1. Generate 1000 synthetic experiment instances, where for all nodes $p_v = q_v$ and the graph is fixed (Null-instances).
2. Generate 1000 synthetic experiment instances that satisfy the associated schema (Tab. 2-3) (Alternative-instances).
3. For each of the Null-instances and Alternative-instances compute the node-level tests statistics associated to the MTST method used. We refer to the output of this step as processed-Null-instances and processed-Alternative-instances.
4. Threshold the processed-Null-instances and processed-Alternative-instances at the full range of possible threshold values *thd* (bigger than 0 value for the methods being tested), and compute the FWER and the true positive rate (TPR):
   - **FWER (x-axis)** For each threshold level, compute the fraction of processed-Null-instances where there was a least one node whose value was bigger than the fixed *thd*.
   - **TPR (y-axis)** For each threshold value, for each of the Alternative-instances compute the fraction of nodes where $p_v \neq q_v$ whose associated test statistic was bigger than *thd*. The reported TPR is the average TPR estimated over all the Alternative-instances.
5. Finally, we compute the AUC from the resulting curve limited to values of FWER in $[0.00, 0.05]$, which are the values of interest for a test of significance level 0.05.

The higher the value of the AUC of the AFROC curve, the more efficient the analyzed algorithm. We divide the result by 0.05 in order to scale the result and keep the same interpretation as for a classical AUC result.

Notice that AFROC ignores the nodes in the Alternative-instances where $p_v = q_v$ whose associated π-value is small (false rejections), thus the Null hypothesis is incorrectly rejected. To quantify how well a method differentiates the nodes that should be rejected, we estimate as well the usual ROC curves from the processed-Alternative instances and compute the associated AUC. The interpretation of the results should take AFROC-AUC as the most important criterion, and ROC-AUC rather as a tiebreaker for approaches with similar AFROC-AUC.

Finally, recall that, in a given study, the graph is not a random variable but it is rather a given fixed element, which justifies why we do not vary this element in the analysis above.

Table 4: Results on synthetic scenarios with variable regularization parameter α. Non-parametric methods applied on multiple two-sample testing over a known graph. Keeping the graph fixed, the AFROC and ROC curves were computed over 1000+1000 experiment instances generated over H_{null} and H_{alt} of Problem 1, respectively. Higher AUC values are better.

		$n=n'=50$		$n=n'=100$		$n=n'=250$	
Experiment	**Method**	**AFROC AUC**	**ROC AUC**	**AFROC AUC**	**ROC AUC**	**AFROC AUC**	**ROC AUC**
Synth.Ia	CTST α=0.01	**0.57**	0.90	**0.76**	0.96	**1.00**	**1.00**
	POOL α=0.01	0.13	0.81	0.24	0.94	0.86	**1.00**
	CTST α=0.1	0.50	**0.93**	0.66	**0.99**	0.99	**1.00**
	POOL α=0.1	0.28	0.84	0.49	0.93	0.64	0.99
	CTST α=0.5	**0.57**	0.92	0.72	0.97	0.98	**1.00**
	POOL α=0.5	0.27	0.87	0.53	0.96	0.87	**1.00**
Synth.Ib	CTST α=0.01	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.01	0.53	**1.00**	0.91	**1.00**	**1.00**	**1.00**
	CTST α=0.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.1	0.72	**1.00**	0.99	**1.00**	**1.00**	**1.00**
	CTST α=0.5	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.5	0.41	0.99	0.85	**1.00**	**1.00**	**1.00**
Synth.IIa	CTST α=0.01	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.01	0.14	0.96	0.72	**1.00**	**1.00**	**1.00**
	CTST α=0.1	0.94	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.1	0.18	0.98	0.84	**1.00**	**1.00**	**1.00**
	CTST α=0.5	0.98	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	POOL α=0.5	0.04	0.89	0.43	0.99	**1.00**	**1.00**
Synth.IIb	CTST α=0.01	0.18	**0.94**	0.43	**0.99**	**1.00**	**1.00**
	POOL α=0.01	0.02	0.83	0.00	0.73	0.43	0.99
	CTST α=0.1	**0.30**	0.92	**0.65**	0.98	0.98	**1.00**
	POOL α=0.1	0.02	0.84	0.12	0.95	0.78	**1.00**
	CTST α=0.5	0.06	0.89	0.52	**0.99**	0.97	**1.00**
	POOL α=0.5	0.04	0.84	0.07	0.91	0.60	0.99

C.2.2 The role of α

In this section, we discuss the role of parameter α in the graph-structured MTST problem. We retain the same set of experiments described in Sec. 4 to compare the role of α in CTST that integrates the graph structure, as well as in the POOL variant that does not consider the graph. The comparison relies on the AFROC-AUC and ROC-AUC measures. Results are summarized in Tab. 4.

As explained in the main text, the role of α is to upper-bound the relative likelihood-ratios r_v^α, thereby preventing convergence issues in terms of sample size and numerical instability. In previous works, such as those in Yamada et al. (2011); de la Concha et al. (2023, 2024), the role of α has been made explicit as a component that controls the speed of convergence of the LRE based on the Pearson's χ^2-divergence. The conclusion drawn by those papers is consistent: a higher value of α will lead to a faster convergence rate. Nevertheless, a high level of α will hinder to quantify the difference between p_v and q_v via the quantity $PE^\alpha(p_v\|q_v)$. In the limit case, that is $\alpha=1$, $PE^\alpha(p_v\|q_v)=0$, meaning these measures fail to differentiate p_v and q_v regardless of the form those pdfs. Therefore, there exists a trade-off: the stability associated with high values of α versus the sensibility of $PE^\alpha(p_v\|q_v)$ in distinguishing between p_v and q_v. This trade-off becomes more relevant when $PE^\alpha(p_v\|q_v)$ is to be used as a test statistic to carry out hypothesis testing and detection tasks.

Findings. Tab. 4 compares CTST and POOL with $\alpha\in\{0.01,0.1,0.5\}$. The first notable observation is that CTST outperforms consistently POOL regardless of the value of α being used. This finding highlights the predominant role of the graph component over that of α, particularly when α is set in a range of meaningful values. The second observation is that POOL's performance appears more sensitive to the values of α, it shows lower stability, especially when there are fewer observations. In contrast, CTST is more robust with respect to this parameter. This can be attributed to the graph-based regularization term that enforces the relative likelihood-ratios estimates to be close in the RKHS, which translates to point-wise similarity as well (see Eq. 6).

Tab. 4 does not provide a clear guideline for choosing the optimal parameter α for CM2ST. In the main text, we fix $\alpha=0.1$ because it yielded the best results for POOL, and we generally recommend using a value of $\alpha<0.5$ when deploying CTST.

C.2.3 Further details regarding the application of two-sample testing on real seismic data

In this section, we provide more details on the preprocessing pipeline to derive the results described in Sec. 4.2, and the additional figures showing the performance of the different methods.

Data preprocessing. As mentioned in the main text, we analyze waveforms that correspond to two seismic events that occurred in New Zealand. Seism A is of magnitude 5.5, while Seism B is magnitude 2.6. These seismic events are part of the publicly available dataset provided by GeoNet. We used the Python package ObsPy to access the data (Beyreuther et al., 2010).

To study the evolution of seismic activity associated to these events, we retrieve the waveforms from 50 seconds before to 50 seconds after the event. These waveforms correspond to the measurements provided by strong-motion accelerometers that monitor shaking in three perpendicular directions. Therefore, here the input space is $\mathcal{X} \subseteq \mathbb{R}^3$. In each of the scenarios, we limit our attention to stations that had recorded observations for all the three directions during all the analyzed time period.

There are three main characteristic that are required to implement CTST in practice:

1. The relative likelihood-ratios $\{r_v^\alpha\}_{v\in V}$ are expected to be approximated by the same RKHS.
2. The FWER control of CTST (see Theorem 3.2) requires that the observations $\mathbf{X} = \{X_v\}_{v\in V} = \{x_{v,1},...,x_{v,n}\}_{v\in V}$ are iid for each node v, and the same for and $\mathbf{X}' = \{X'_v\}_{v\in V} = \{x'_{v,1},...,x'_{v,n'}\}_{v\in V}$.
3. The vector-valued function $\mathbf{r}^\alpha = (r_1,...,r_N)$ is expected to be smooth with respect to the graph G, i.e. $\|r_u - r_v\|_{\mathbb{H}} < \epsilon$ for connected nodes.

The preprocessing aims to transform and prepare the data so they satisfy these conditions.

We follow the preprocessing pipeline described in Chen et al. (2021) with the toolbox for Seismology ObsPy. The preprocessing is performed independently for each station and independently for each direction. We start by steps that are considered to be standard in seismology: we remove the linear trend and we apply a 2-16 bandpass filter. To reduce the temporal dependency, we compute a root mean square amplitude envelope, then we fit an autoregressive model of order 1, and we keep the residuals from this model. The output is standardized so that it has zero mean and unit variance. To make the data comparable between stations, we divide the output by its maximum value.

Our objective is to provide a visualization that captures the evolution of the seismic event using the measurements available at each station. To this end, we use a graph-structured MTST to identify the specific moments and locations (stations) where the seismic activity appeared to be statistically significant. In this context, $v \in V$ denotes that station v belongs to the set of stations V. To define the statistical test, we need to identify the samples $\mathbf{X}$ and $\mathbf{X}'$ (see Eq. 3) that should be compared across the spatial and temporal dimensions. We denote by τ the time-stamp of the seismic event, then we consider the preprocessed observations in two time frames: $[\tau-50,...,\tau)$ and $[\tau,...,\tau+50)$, i.e. from 50 seconds before τ to 50 seconds after τ. These periods are segmented into 10 time-windows ($\mathcal{T} = \{1,...,10\}$, each of 5 seconds duration) made of 100 prepossessed observations in each of them. According to our notation, $X_{v,1}$ is the first 100 observations at station v after $\tau-50$, while $X'_{v,1}$ denotes the first 100 observations post-event (τ). Following the same logic, $X_{v,2}$ has the observations after $\tau-45$ at station v, while $X'_{v,1}$ denotes the first 100 observations after $\tau+5$. This segmentation yields two samples for each location-time pair $(v,t) \in V\times\mathcal{T}$, $X_{(v,t)} = \{x_{((v,t),i)}\}_{i=1}^{100} \sim p_{(v,t)}$ and $X'_{(v,t)} = \{x'_{((v,t),i)}\}_{i=1}^{100} \sim q_{(v,t)}$. Then, the MTST compares the pdfs $\{p_{(v,t)}\}_{(v,t)\in V\times\mathcal{T}}$ and $\{q_{(v,t)}\}_{(v,t)\in V\times\mathcal{T}}$. Alternatives can be implemented for defining different observations to consider from $\{p_{(v,t)}\}_{(v,t)\in V\times\mathcal{T}}$ to be used to compare with the post-event alternative.

The sets $\mathbf{X} = \{X_{(v,t)}\}_{v\in V,t\in\mathcal{T}}$ and $\mathbf{X}' = \{X'_{(v,t)}\}_{v\in V,t\in\mathcal{T}}$ represent the observations available at the graph $G_{\mathrm{S}\times\mathrm{T}}$ whose nodes represent a position in space and in time. As in the general graph-structure hypothesis testing problem, $G_{\mathrm{S}\times\mathrm{T}}$ encodes the expected similarity between the results of the test. To encode the fact that close stations are expected to affect each other (recall the *first law of Geography* from Sec. 1), we generate an unweighted spatial graph $G_{\mathrm{S}} = (V,E,W)$ where the nodes represent the geographical positions of the seismic stations and the edges are computed in order to form a 3-nearest neighbors graph. To account for the temporal component similarity expected from the propagation of the seismic waves through the earth, we build an unweighted multiplex network $G_{\mathrm{S}\times\mathrm{T}} = (V_T, E_T, W_T)$ on top of G_{S}. The set of nodes is then the pair $(v,t) \in V_T := V\times\mathcal{T}$, where V denotes the set of nodes of G_{S}. Two nodes in $G_{\mathrm{S}\times\mathrm{T}}$, (u,t) and (v,t'), are connected: i) if $t = t'$ and $(u,v) \in E$, i.e. they refer to the same time-window and the nodes u and v are connected in the spatial graph G_{S}, ii) or if $u = v$

and $|t'-t|=1$, i.e. each node $v \in V$ is connected to its 'copies' in the two adjacent time-windows.

The implementation details of the statistical methods being compared are the same as in the synthetic scenarios, which include the hyperparameters selection related to the estimation of the non-parametric test statistics and the way the permutation test is run.

Findings. Fig. 3-6 and Fig. 7-11 illustrate the output of the graph-structured MTST applied to the waveforms related to each for the two seismic events, Seism A and Seism B. The figures highlight the biggest connected component $C_{\mathrm{S}\times\mathrm{T}}$ made of pairs $(v,t) \in V_T$ that were identified as statistically significant by the method being used. In this application, we called a pair to be statistically significant if its π-value is smaller than 0.05.

We try to show in the figures both dimensions of the test. The graph on the left highlights in red the stations $v \in V$ which were elements of the biggest connected component C_T for at least one time-window, That is, there exist $t \in \mathcal{T} = \{1,...,10\}$ such that $(v,t) \in C_T$. The epicenter is marked by a red star. The time-series at the left show both preprocessed data samples $X_{(v,t)} = \{x_{((v,t),i)}\}_{i=1}^{100} \sim p_{(v,t)}$ (green time-series) and $X'_{(v,t)} = \{x'_{((v,t),i)}\}_{i=1}^{100} \sim q_{(v,t)}$ (red time-series) for the highlighted stations. The periods that were considered statistically significant are delineated by blue/pink colors (we use two colors to differentiate adjacent time-windows where the test rejected the $\mathrm{H}_{\mathrm{null},v}$ hypothesis).

The first thing to notice is that all methods identified the stations that were closer to the epicenter as locations where there was statistically significant evidence of a change. Algorithms that neglect the graph component tend to identify a larger number of time-windows. Upon closer inspection, we can see that many of the identified time-windows appear to be false positives, lacking in global relevance or consistency with the expected evolution of the sesmic activity. Intuitively, for a short time-period around an even (here we analyze 100 seconds overall), we expect the event to alter the behavior of the measurements during consecutive time-windows, and this effect will vanish with time. This pattern is not evident in methods that disregard the spatial and temporal similarity. In contrast, CTST identifies correctly nodes close to the epicenter, and captures the evolution of the seismic activity in a more consistent way. The results are consistent with those obtained from synthetic experiments, where CTST demonstrated superior performance in terms of the AFROC-AUC. This performance also indicates the effective weak FWER control and higher sensitivity in pinpointing the nodes where $p_v \neq q_v$.

From the practitioners' perspective, MT is usually an initial exploratory tool, where tests identified as statistically significant are further inspected with further analyzis. In this sense, false positives may translate to a high cost, since they may lead to the allocation of resources towards the wrong direction. Thus, the accuracy of identifying nodes where $p_v \neq q_v$ is not just a statistical concern, but also a practical one, directly impacting the efficiency and effectiveness of subsequent research efforts.

Figure 3: Seism A in New Zealand (1 of 4).

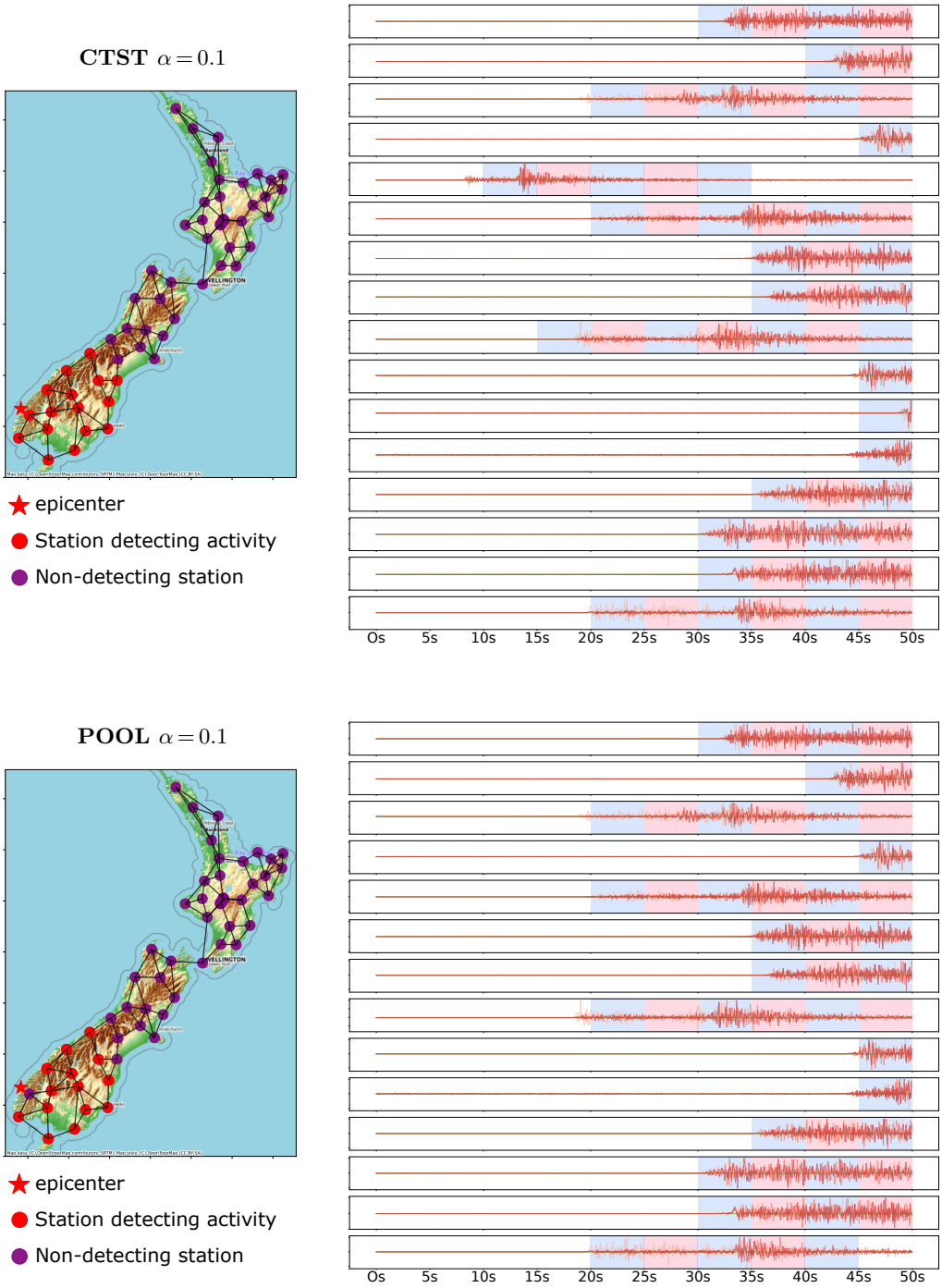


Figure 4: Seism A in New Zealand (2 of 4).

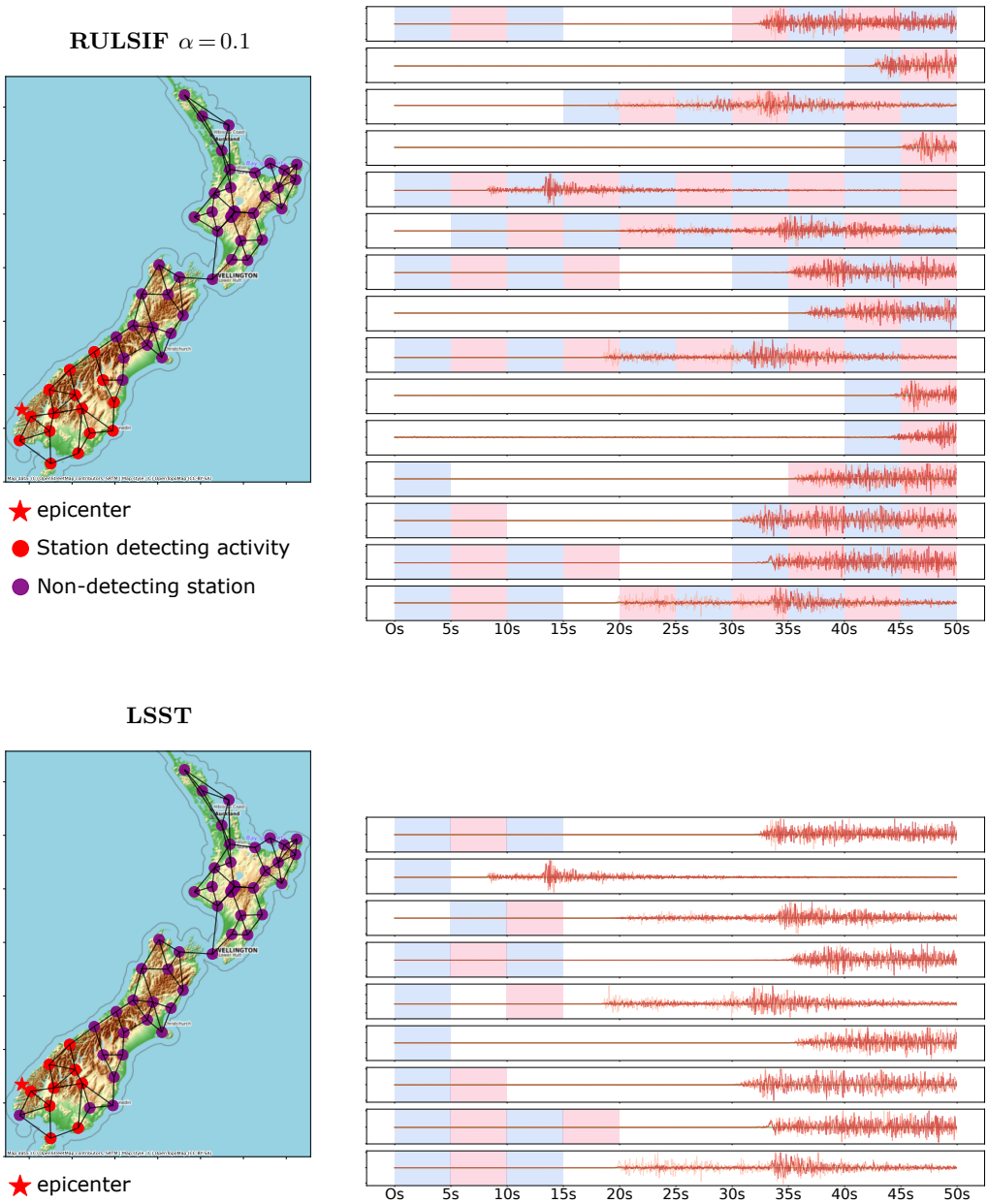


Figure 5: Seism A in New Zealand (3 of 4).

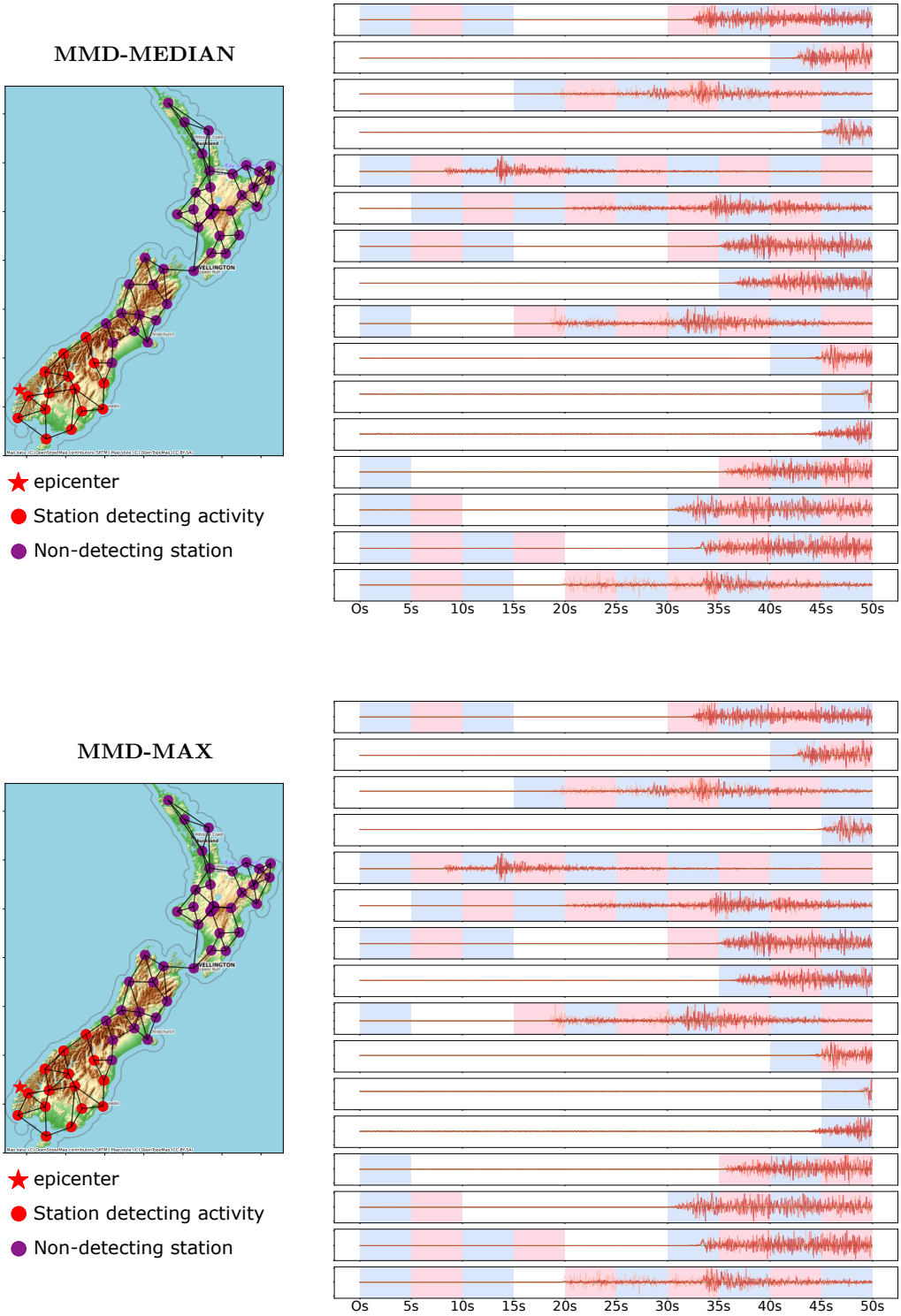


Figure 6: Seism A in New Zealand (4 of 4).

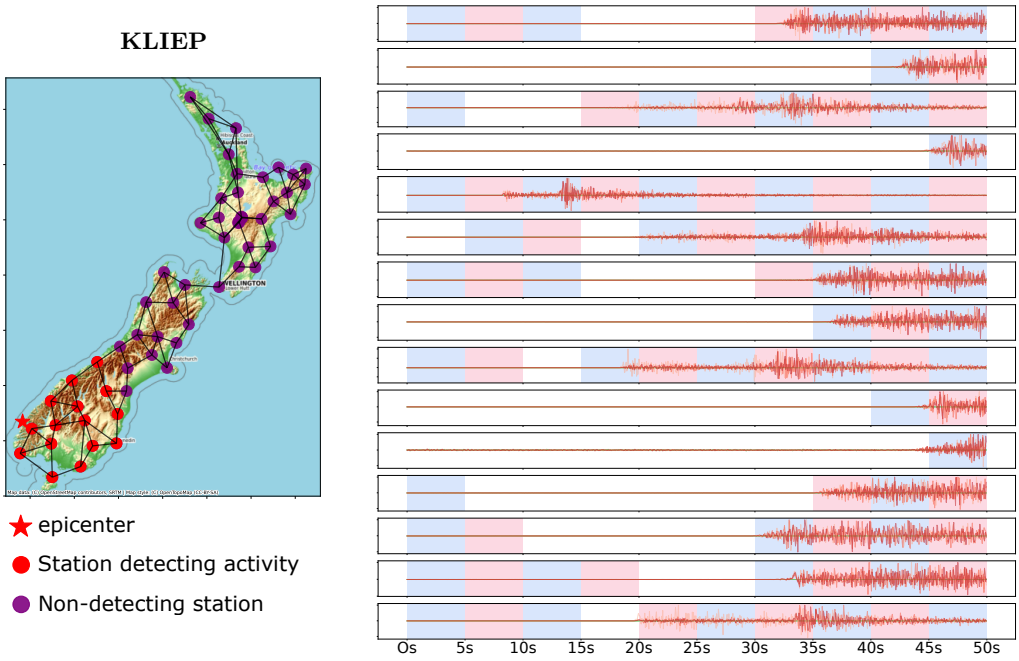
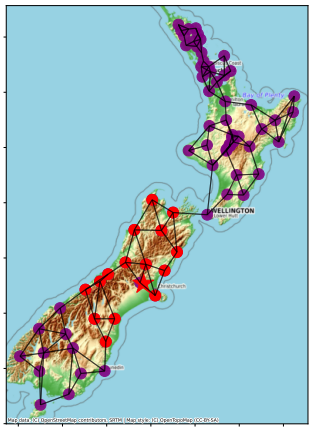
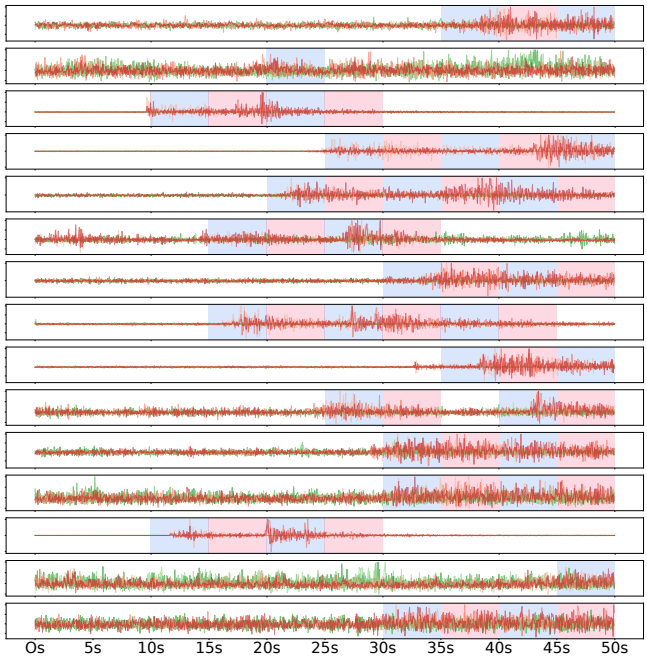


Figure 7: Seism B in New Zealand (1 of 6).

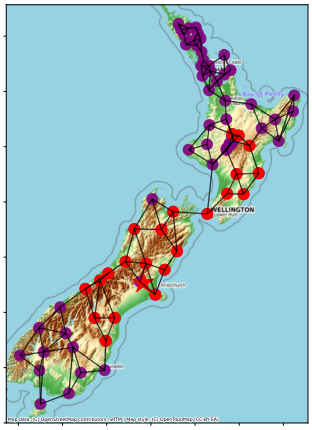
CTST $\alpha = 0.1$



★ epicenter
● Station detecting activity
● Non-detecting station



POOL $\alpha = 0.1$



★ epicenter
● Station detecting activity
● Non-detecting station

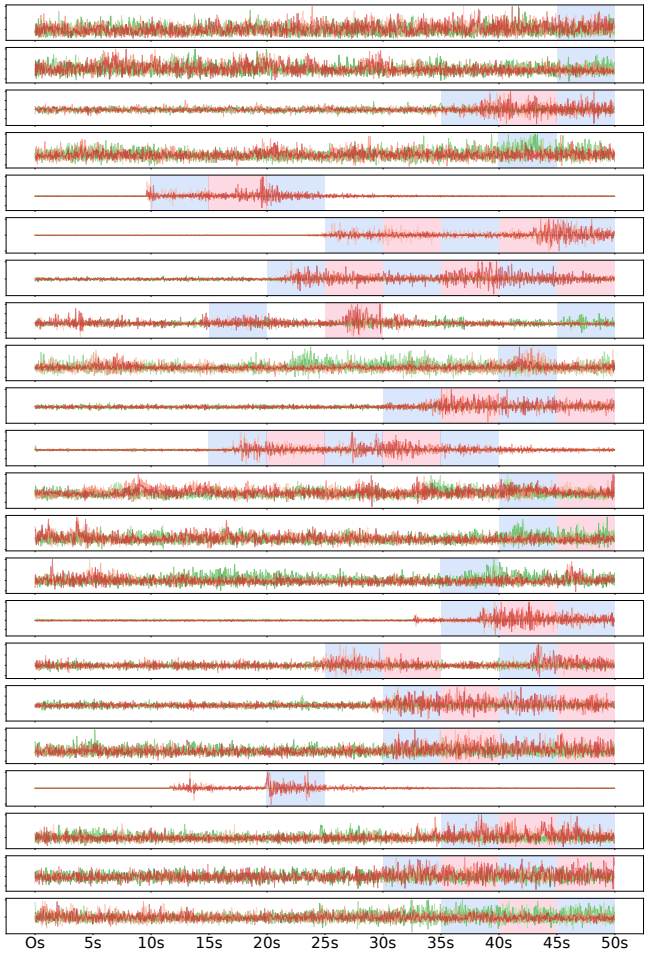
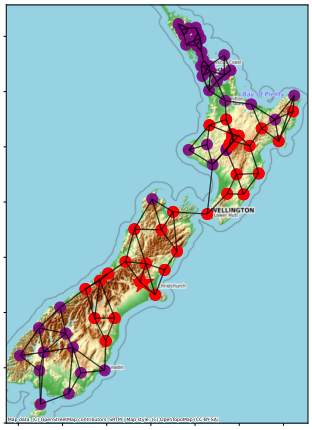


Figure 8: Seism B in New Zealand (2 of 6).

RULSIF $\alpha = 0.1$



epicenter

Station detecting activity

Non-detecting station

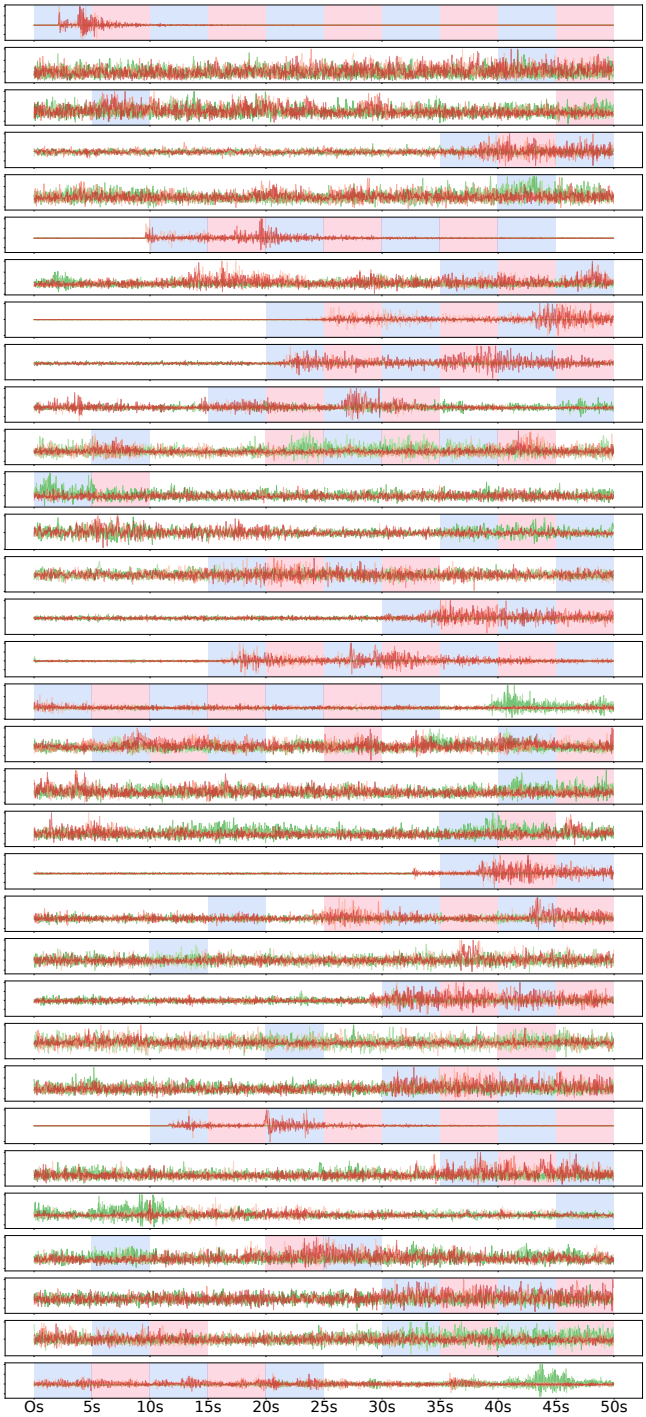
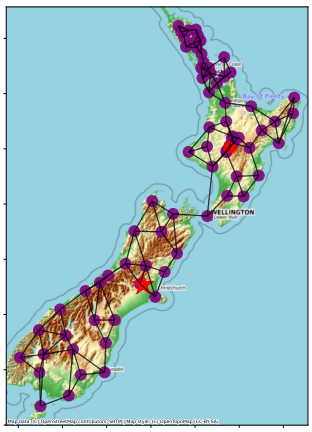


Figure 9: Seism B in New Zealand (3 of 6).

LSST



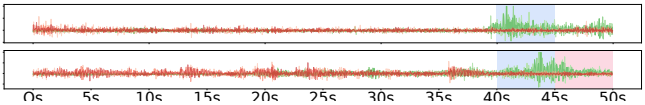


Figure 10: Seism B in New Zealand (4 of 6).

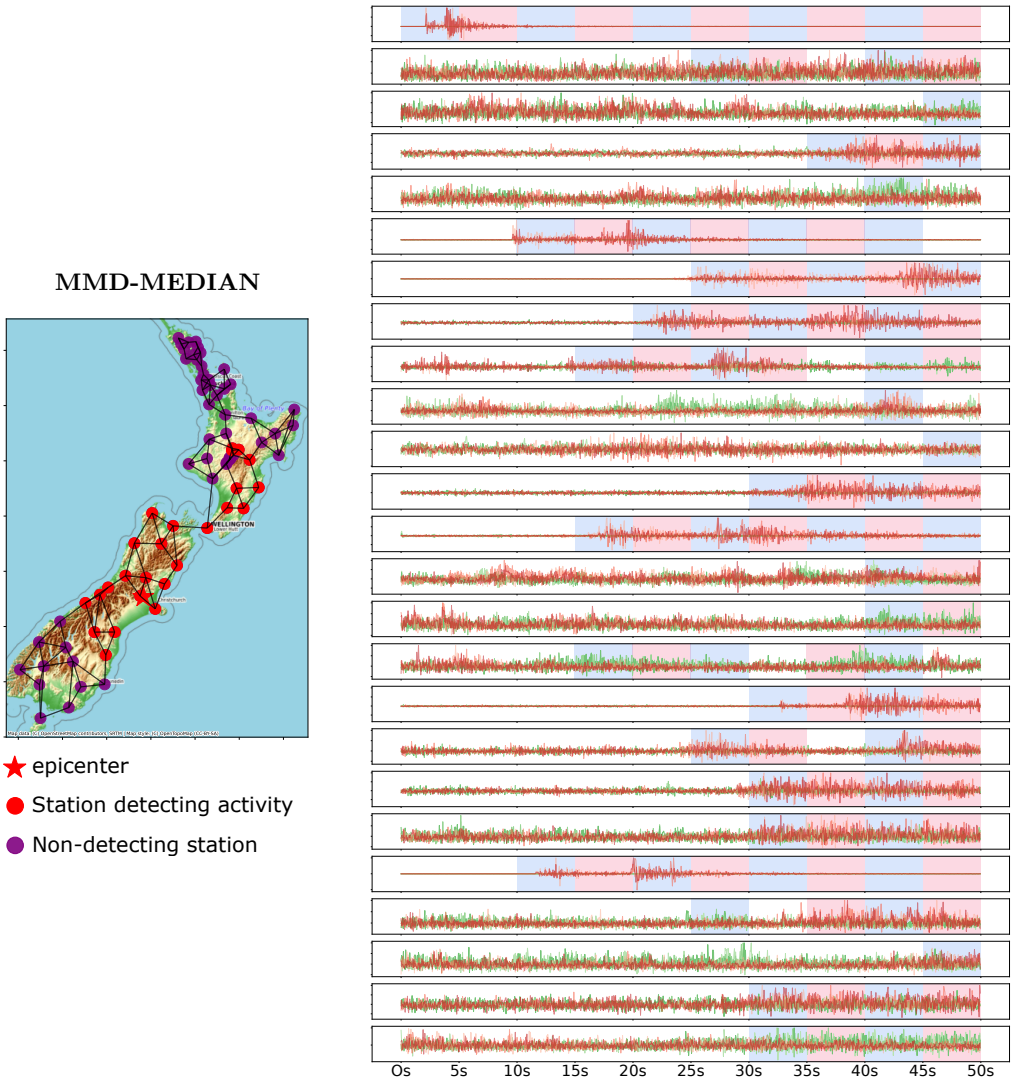


Figure 11: Seism B in New Zealand (5 of 6).

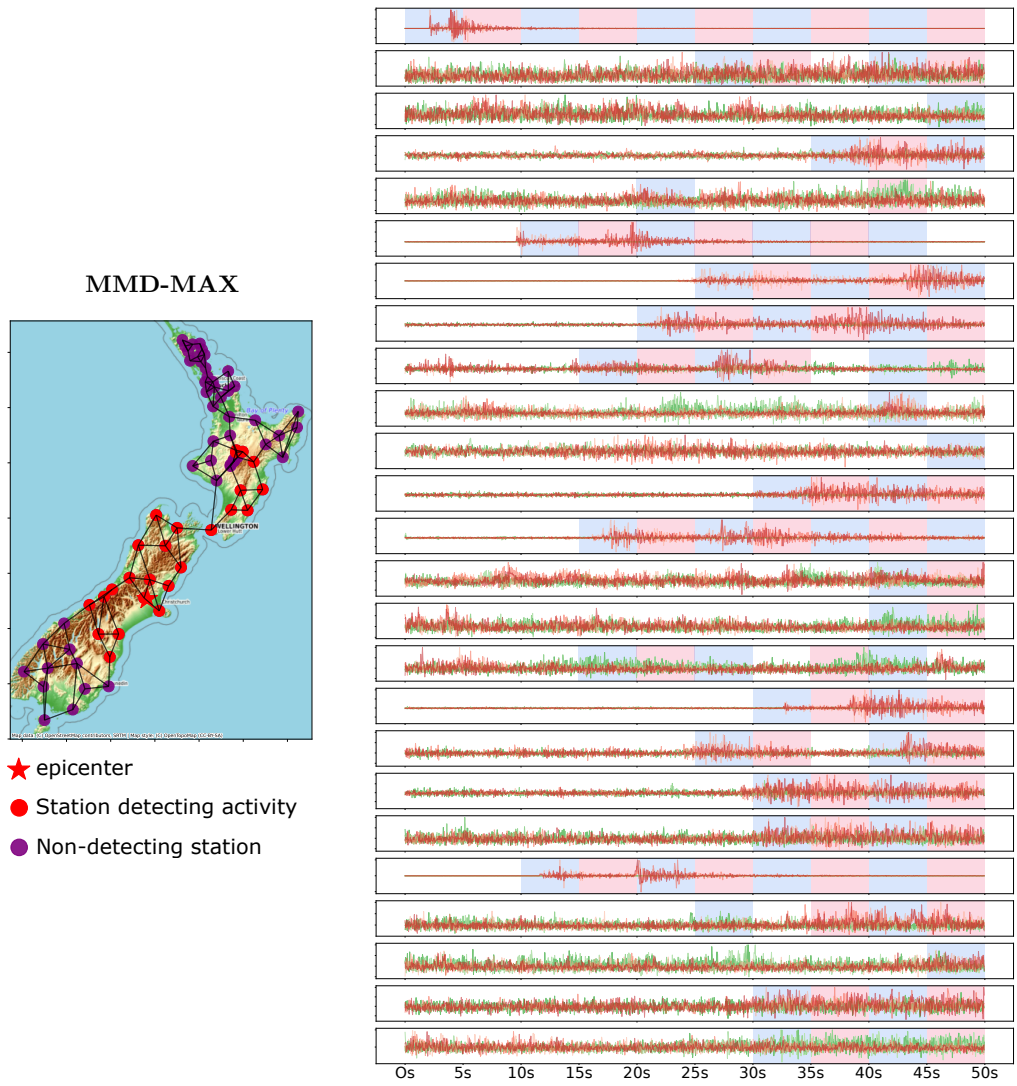


Figure 12: Seism B in New Zealand (6 of 6).