

Optimizing the hunt for extraterrestrial high-energy neutrino counterparts

Pouya M. Kouch^{1,2,3,*}, Elina Lindfors^{1,2}, Talvikki Hovatta^{2,3}, Ioannis Liodakis^{4,2,5}, Karri I. I. Koljonen⁶, Kari Nilsson², Jenni Jormanainen^{1,2}, Vandad Fallah Ramazani², and Matthew J. Graham⁷

¹ Department of Physics and Astronomy, University of Turku, FI-20014 Turku, Finland

² Finnish Centre for Astronomy with ESO (FINCA), Quantum, Vesilinnantie 5, University of Turku, FI-20014 Turku, Finland

³ Aalto University Metsähovi Radio Observatory, Metsähovintie 114, FI-02540 Kylmälä, Finland

⁴ Institute of Astrophysics, FORTH, N. Plastira 100, GR-70013 Heraklion, Greece

⁵ NASA Marshall Space Flight Center, Huntsville, AL 35812, USA

⁶ Institutt for Fysikk, Norwegian University of Science and Technology, Høgskloreiringen 5, Trondheim 7491, Norway

⁷ Division of Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena CA91125, USA

Received 3 December 2024 / Accepted 23 February 2025

ABSTRACT

It has been a decade since the IceCube collaboration began detecting high-energy (HE) neutrinos originating from cosmic sources. Despite a few well-known individual associations and numerous phenomenological, observational, and statistical multiwavelength studies, the origin of astrophysical HE neutrinos largely remains a mystery. To date, the most convincing associations link HE neutrinos with active galactic nuclei (AGNs). Consequently, many studies have attempted population-based correlation tests between HE neutrinos and specific AGN subpopulations (such as blazars). While some of the associations are suggestive, no definitive population-based correlation has been established. This could result from either a lack of a population-based correlation or insufficient detection power, given the substantial atmospheric neutrino background. By leveraging blazar variability, we performed spatio-temporal blazar-neutrino correlation tests aimed at enhancing detection power by reliably incorporating temporal information into the statistical analysis. We used simulations to evaluate the detection power of our method under various test strategies. We find that: (1) with sufficiently large source samples, if 20% of astrophysical HE neutrinos originate from blazars, we should robustly observe $\sim 4\sigma$ associations; (2) a counting-based test statistic combined with a top-hat weighting scheme (rather than a Gaussian one) provides the greatest detection power; (3) applying neutrino sample cuts reduces detection power when a weighting scheme is used; and (4) in top-hat-like weighting schemes, low p-values do not occur arbitrarily with an increase in the HE neutrino error region size (any such occurrence is indicative of an underlying blazar–neutrino correlation).

Key words. astroparticle physics – neutrinos – galaxies: active – galaxies: jets – galaxies: statistics

1. Introduction

In 2008, the IceCube Neutrino Observatory began observing the sky for high-energy (TeV–PeV) neutrinos (hereafter “neutrinos”; e.g., Aartsen et al. 2013; IceCube Collaboration et al. 2021). Since then, hundreds of such neutrinos have been detected from cosmic sources (IceCube Collaboration 2013). Despite more than a decade of observations, the origin of these astrophysical neutrinos remains elusive. This uncertainty is partly due to their intrinsic rarity, and partly due to observational challenges posed by the atmospheric neutrino background (e.g., Abbasi et al. 2021a; Troitsky 2021).

Active galactic nuclei (AGNs), powered by supermassive black holes, have long been suspected to be sources of neutrinos (e.g., Mannheim & Biermann 1989; Stecker et al. 1991; Mannheim et al. 1992; Mücke & Protheroe 2001; Becker 2008; Dermer et al. 2014; Stecker 2013; Hooper & Plant 2023). A fraction of AGNs exhibit highly collimated jets of relativistically moving plasma. When the jet of an AGN aligns with our line of sight, the AGN is classified as a blazar. The small viewing angle of the jet causes all emission from the relativistically accelerated plasma to become extremely Doppler-boosted, mak-

ing blazars some of the brightest objects in the Universe (for recent reviews, see, e.g., Böttcher 2019; Blandford et al. 2019; Hovatta & Lindfors 2019).

In 2017, the IceCube Neutrino Observatory detected a high-energy neutrino coming from the direction of the blazar TXS 0506+056, which was in a state of heightened electromagnetic activity ($>3\sigma$ spatio-temporal chance coincidence; IceCube Collaboration et al. 2018). Subsequent analysis of archival data revealed an excess of neutrinos from the direction of TXS 0506+056. Later, the IceCube collaboration identified another AGN, the Seyfert-II NGC 1068, as a source of neutrinos with a 4.2σ significance level (IceCube Collaboration et al. 2022). Additionally, two other blazars, PKS 1424+240 and GB6 J1542+6129, were identified as potential neutrino emitters (Abbasi et al. 2021b).

While many blazar–neutrino correlation studies have focused on individual source associations (e.g., Krauß et al. 2014; Kadler et al. 2016; Righi et al. 2019; Franckowiak et al. 2020; Das et al. 2022), many studies have also investigated the blazar–neutrino connection via population-based statistical analyses (e.g., Aartsen et al. 2020; Plavin et al. 2020; Giommi et al. 2020; Plavin et al. 2021; Smith et al. 2021; Hovatta et al. 2021; Zhou et al. 2021; Bartos et al. 2021; Kun et al. 2022; Buson et al. 2022, 2023; Plavin et al. 2023; Novikova et al.

* Corresponding author; pouya.kouch@utu.fi

2023; Bellenghi et al. 2023; Suray & Troitsky 2024; Lu et al. 2024; Plavin et al. 2024; Kouch et al. 2024), aiming to leverage the abundance and precise sky localization of blazars to enhance detection power. Several of these studies found hints of a potential connection, but none conclusively established a correlation between blazar populations and neutrinos. Therefore, it is crucial to identify the most effective detection strategies, which is the focus of this work.

An important feature of blazars is the stochastic nature of their energy output. Blazars exhibit significant variability across nearly all electromagnetic bands, alternating between quiescent and enhanced emission states (commonly referred to as “flaring” periods). In our previous studies (Hovatta et al. 2021 and Kouch et al. 2024; hereafter H21 and K24, respectively), we focused on flaring periods in the lower energy bands. In the radio band, flares often trace the enhancement of the overall jet activity (e.g., Lähteenmäki & Valtaoja 2003; León-Tavares et al. 2011), suggesting that major flares could correlate with enhanced neutrino production (e.g., Plavin et al. 2020; H21). In the optical band, the variability typically traces the slower radio-band variability as well as the faster X-ray and γ -ray flares (e.g., Lindfors et al. 2016), which requires efficient particle acceleration and cooling (e.g., Oikonomou et al. 2019; Kreter et al. 2020; Stathopoulos et al. 2022). This is the basis for the hypothesis of a temporal connection between flaring periods and neutrino emission.

Incorporating the temporal information of neutrinos into blazar–neutrino correlation schemes is crucial for boosting detection power (e.g., Abbasi et al. 2024), especially given the rather poor sky localization of neutrinos. This temporal connection between blazar flares and neutrino production may also imply a nontemporal effect: if more neutrinos are produced during flares, then a more frequently flaring blazar is more likely to be a neutrino emitter on average. Additionally, since γ -ray-emitting blazars are more variable in both radio and optical bands than γ -ray-dark blazars (e.g., Richards et al. 2011, 2014; Hovatta et al. 2014), we could likewise expect neutrino-emitting blazars to show greater time-averaged variability than neutrino-dark blazars. These are admittedly rather simplistic interpretations of the complex flaring behavior in blazars. Nevertheless, we consider it worthwhile to investigate a possible connection between time-averaged variability and neutrinos (see Sect. 2.2).

In our previous simulations (Liodakis et al. 2022, hereafter L22), we evaluated the feasibility of detecting a significant spatio-temporal blazar–neutrino correlation using available data (~ 1800 blazar radio light curves and 56 IceCube high-energy neutrino events, as in H21). We showed that incorporating the temporal dimension is necessary for establishing such a correlation reliably. We also found that increasing the blazar sample size strengthened the correlation. In this study, in preparation for our upcoming blazar–neutrino correlation study set to utilize ~ 4000 blazar optical light curves, we performed similar simulations with an expanded sample of blazar light curves (~ 4000 ; see Sect. 2.2) and a larger, better-sampled catalog of IceCube high-energy neutrinos (283 events as used in K24, mostly taken from IceCat-1; Abbasi et al. 2023; see our Sect. 2.1). Our focus is to simulate blazars with an underlying spatio-temporal correlation with neutrinos, and then evaluate the effectiveness of different correlation test strategies to establish this induced connection. This allows us to determine the most optimal test strategies for application to observed datasets in our upcoming study.

We ran these tests on five blazar samples: two for sanity checks, two for special cases, and one for a potentially realistic scenario (see Sect. 3.1). In short, we first calculated the

correlation strengths between neutrinos and a population of 4000 random blazars (the first sanity-check sample) across all test strategies. Then, we added a number (tens to hundreds) of spatio-temporally neutrino-associated blazars to this random sample and reevaluated the correlation strengths. We repeated this simulation process 1000 times to determine the fraction of significant detections produced by each test strategy; this allowed us to systematically identify the most effective correlation test strategies for detecting a potential spatio-temporal blazar–neutrino correlation.

In Sect. 2.1 we describe the neutrino data used. In Sect. 2.2 we present the blazar sample and variability measures. In the beginning of Sect. 3, we outline the basics of the spatio-temporal search method applied in H21, K24, and this study. In Sect. 3.1 we explain how we simulated five blazar datasets with varying degrees of spatio-temporal connection to neutrinos. In Sects. 3.2, 3.3, and 3.4 we describe how we set up and combined different correlation test strategies. We present and discuss our results in Sect. 4. Finally, in Sect. 5 we summarize our findings.

2. Data

2.1. High-energy neutrino events

The majority of the neutrino data used in this study are from the IceCube collaboration’s IceCat-1 (Abbasi et al. 2023), which recorded the first catalog of muon-track high-energy neutrino events between May 2011 and December 2022 based on real-time alert selection criteria. Our sample also includes events from an earlier IceCube collaboration study (Abbasi et al. 2022), which analyzed events between May 2009 and December 2018; it adds 16 events due to an earlier start date (when the telescope was in partial configuration) and the inclusion of non-real-time events. Together, our combined dataset termed IceCat1+, contains a total of 283 events. We previously used IceCat1+ in K24, where it is available as an electronic table.

Each neutrino in IceCat1+ has an estimated kinetic energy and a “signalness” (\mathcal{S} ; a measure indicating the likelihood of the neutrino being of astrophysical origin). Additionally, the IceCube collaboration provides maximum-likelihood contours on the sky, resembling ellipsoids, centered on the most probable arrival direction of each neutrino (e.g., Fig. 3 in Abbasi et al. 2023). They report the right ascension (RA) and declination (Dec.) of these arrival directions along with four asymmetric error bars in the RA and Dec directions, representing the 90%-confidence-level contour size. The error bars are determined by bounding the 90%-confidence-level contour within a rectangle. Since maximum-likelihood contours generally resemble ellipsoids with a Gaussian drop-off, it is mathematically convenient to approximate them using the four asymmetric RA and Dec. error bars entirely enclosing the 90%-confidence-level contour area. We refer to this region as the $\geq 90\%$ -likelihood error region (see Fig. 1). We note that each of the four asymmetric error bars is taken as a one-sided, 2σ Gaussian error.

In IceCat1+, the median signalness is $\bar{\mathcal{S}} = 0.429$ and the median area of the $\geq 90\%$ -likelihood error region is $\bar{\Omega} = 6.63 \text{ deg}^2$. For each neutrino, the latter quantity was calculated as

$$\Omega = \frac{\pi}{4} (\alpha^+ \cdot \delta^+ + \alpha^- \cdot \delta^+ + \alpha^- \cdot \delta^- + \alpha^+ \cdot \delta^-), \quad (1)$$

where $\alpha^{+/-}$ and $\delta^{+/-}$ represent the asymmetric 2σ Gaussian error bars in the RA and Dec directions, respectively (see Fig. 1).

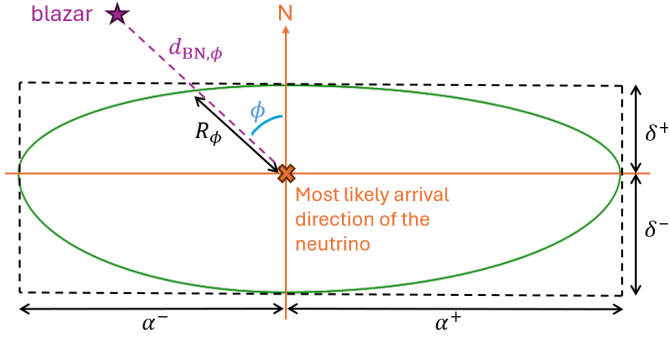


Fig. 1. Spatial orientation of an example blazar–neutrino pair. The blazar is shown as a purple star. The most likely arrival direction of the neutrino is shown by the orange cross. The $\geq 90\%$ -likelihood error region of the neutrino (see Sect. 2.1) is shown by the green ellipsoid. $d_{\text{BN},\phi}$ is the distance between the blazar and the center of the neutrino error ellipsoid, shown by the dotted purple line. R_ϕ is the distance between the edge and center of the ellipsoid along the line that goes through the blazar, shown by the black double-headed arrow oriented at a ϕ angle from the northern axis. $\alpha^{+/-}$ and $\delta^{+/-}$ represent the asymmetric 2σ Gaussian error bars in RA and Dec, respectively.

2.2. Blazars

As mentioned in Sect. 1, our goal is to perform simulations to test the detection power of various strategies in advance of our upcoming blazar–neutrino correlation study, which will utilize ~ 4000 optical light curves. In line with this, we simulated blazar datasets containing around 4000 sources. The simulation details are given in Sect. 3. In this subsection we introduce the blazar population used and their variability measures.

Decades of studies on the variability of blazars in the optical band have revealed that these sources exhibit significant differences in timescale, amplitude, and duty cycle (see Sect. 1). To parameterize both time-averaged and time-resolved variability, we used CAZ (CRTS+ATLAS+ZTF) optical light curves of blazars from the CGRaBS sample, as defined in K24. These light curves combine data from several all-sky surveys and, where available, from dedicated monitoring programs such as the Katzman Automatic Imaging Telescope (KAIT¹; Filippenko et al. 2001; Liodakis et al. 2018) and Tuorla² (Nilsson et al. 2018). The all-sky surveys include the Catalina Real-time Transient Survey (CRTS³; Drake et al. 2009), the Asteroid Terrestrial-impact Last Alert System survey (ATLAS⁴; Tonry et al. 2018), and the Zwicky Transient Facility survey (ZTF⁵; Bellm et al. 2019). We compiled the light curves by performing forced-photometry on the sky coordinates of the CGRaBS blazars. While the CGRaBS sample is complete down to a 4.8 GHz flux density of 65 mJy and radio spectral index $\alpha > -0.5$ where $S \propto \nu^\alpha$, it over-represents low synchrotron peaked blazars, which are brighter than other blazar subclasses in the radio band. Thus, while the CGRaBS sample may not fully represent the entire blazar population, it provides a suitable basis for the general parameterization of their variability.

For the time-averaged variability measure, we used fractional variability (F_{var}), which quantifies the overall variability of the flux density over a specific period. For a sequence of data points, $x_i \pm \epsilon_i$, where $i = 1, \dots, N$, fractional variability is defined

¹ <https://w.astro.berkeley.edu/bait/kait.html>

² <https://users.utu.fi/kani/1m/>

³ <https://crts.caltech.edu/>

⁴ <https://fallingstar-data.com/>

⁵ <https://irsa.ipac.caltech.edu/Missions/ztf.html>

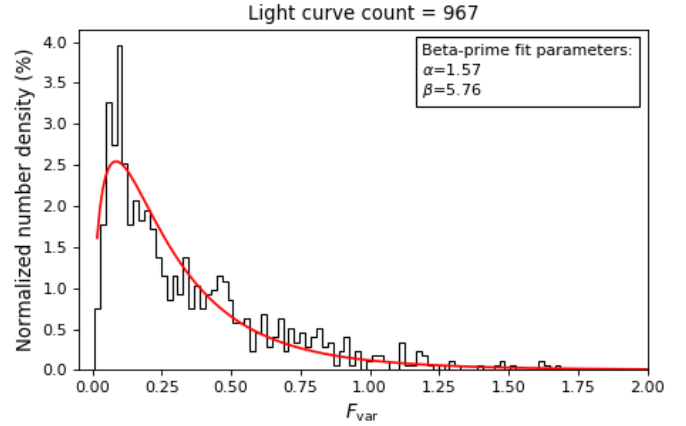


Fig. 2. Global distribution of F_{var} for 967 CAZ light curves of the CGRaBS sample. The red curve shows a Beta-prime parameterization of the distribution. The Beta-prime fit parameters are $\alpha = 1.57$ and $\beta = 5.76$.

as follows:

$$F_{\text{var}} \equiv \sqrt{\frac{S^2 - \bar{\epsilon}^2}{\bar{x}^2}}, \quad (2)$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ is the sample mean, $S^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{x} - x_i)^2$ is the sample variance, and $\bar{\epsilon}^2 = \frac{1}{N} \sum_{i=1}^N \epsilon_i^2$ is the mean square error of the sample (e.g., Gliozzi et al. 2003; Vaughan et al. 2003; MAGIC Collaboration et al. 2024).

For the time-resolved variability measure, we used the activity index (AI), which we define here as the flux density at a given time (S_t) normalized by the median flux density (\bar{S}). We note that this definition differs from that in H21 and K24, where S_t was taken as the average flux density within a time window, while here S_t is calculated instantaneously at time t .

To ensure statistical robustness, we limited our sample to light curves with at least 100 data points, resulting in 967 CAZ light curves. Figure 2 shows the global distribution of F_{var} for these light curves parameterized using a Beta-prime distribution with a probability density function $f(x) = x^{\alpha-1}(1+x)^{-\alpha-\beta}/\mathcal{B}(\alpha,\beta)$ where $\mathcal{B}(\alpha,\beta)$ is the Beta function. The Beta-prime function is suitable for this parameterization as it is continuous and covers the positive real line (e.g., Blinov et al. 2016; Hovatta et al. 2016; Bourguignon et al. 2018). The best-fit Beta-prime parameters for the F_{var} distribution are $\alpha = 1.57$ and $\beta = 5.76$.

Regarding AI, a flux density value around time t is typically chosen (e.g., H21 and K24). However, for computational efficiency, we quantified AI directly from the flux density (S) distribution of each blazar (see Sect. 3.1.1). Thus, for each simulated blazar, we generated a random, yet realistic, flux density distribution.

To achieve this, we found the flux density distribution for all 967 CAZ light curves and parameterize them using a lognormal distribution, with the probability density function $f(x) = (x \sigma \sqrt{2\pi})^{-1} \cdot \exp[-(\ln x - \mu)^2 / (2\sigma^2)]$. The best-fit lognormal “location” and “scale” parameters are referred to as μ_{LN} and σ_{LN} , respectively. For example, Fig. 3 shows the CAZ light curve and the median-normalized flux density distribution for the blazar J0509+0541 (TXS 0506+056) as well as its best-fit lognormal curve. Repeating this process for all light curves yields global distributions for μ_{LN} and σ_{LN} as shown in Fig. 4.

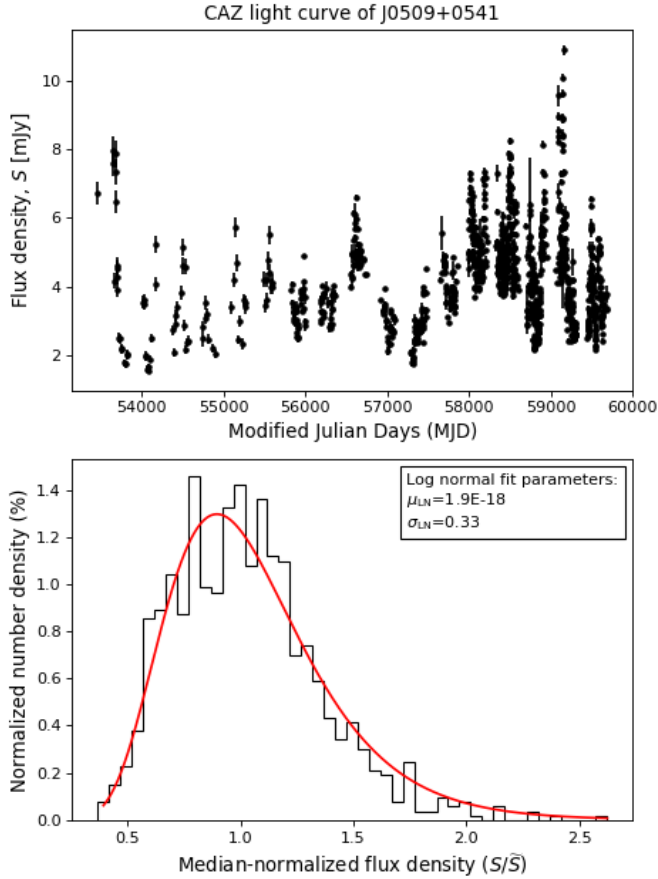


Fig. 3. CAZ optical light curve of the blazar J0509+0541 (also known as TXS 0506+056). *Top*: CAZ light curve. *Bottom*: Corresponding median-normalized flux density (S/\bar{S}) distribution and its best-fit log-normal distribution (shown as the red curve with the parameters $\mu_{LN} \approx 10^{-18}$ and $\sigma_{LN} = 0.33$).

Since the light curves are median-normalized, we expect their μ_{LN} to be around zero. Indeed, the observed global distribution of μ_{LN} (top panel of Fig. 4) is generally close to zero: its median is on the order of 10^{-15} , with 90% of values below 0.052. Notably, this distribution exhibits two peaks at $\mu_{LN} \approx 10^{-17}$ and $\mu_{LN} \approx 10^{-3}$, arising due to the presence of two distinct populations of flux densities in the CAZ light curves⁶. Additionally, $\mu_{LN} \geq 0.1$ in $\sim 10\%$ of the fits, which is likely due to low-quality light curves or poor lognormal fits. Hence, we set μ_{LN} to zero for simplicity. On the other hand, the σ_{LN} global distribution shows a long-tailed Gaussian-like spread, which we fit with a Beta-prime function (see the bottom panel of Fig. 4); the best-fit parameters are $\alpha = 2.02$ and $\beta = 8.97$. Therefore, using $\mu_{LN} = 0$ and selecting a random value from the best-fit Beta-prime distribution for σ_{LN} , we can generate flux density distribution resembling those observed.

⁶ The first half of the CAZ light curves is almost exclusively populated by CRTS light curves, while the other half contains the other surveys combined. This creates two distinct portions with different properties (e.g., cadence, signal-to-noise ratio, average flux density), often resulting in two distinct peaks in global distributions. This is explored further in the CAZ data release paper, Kouch et al. (in prep.).

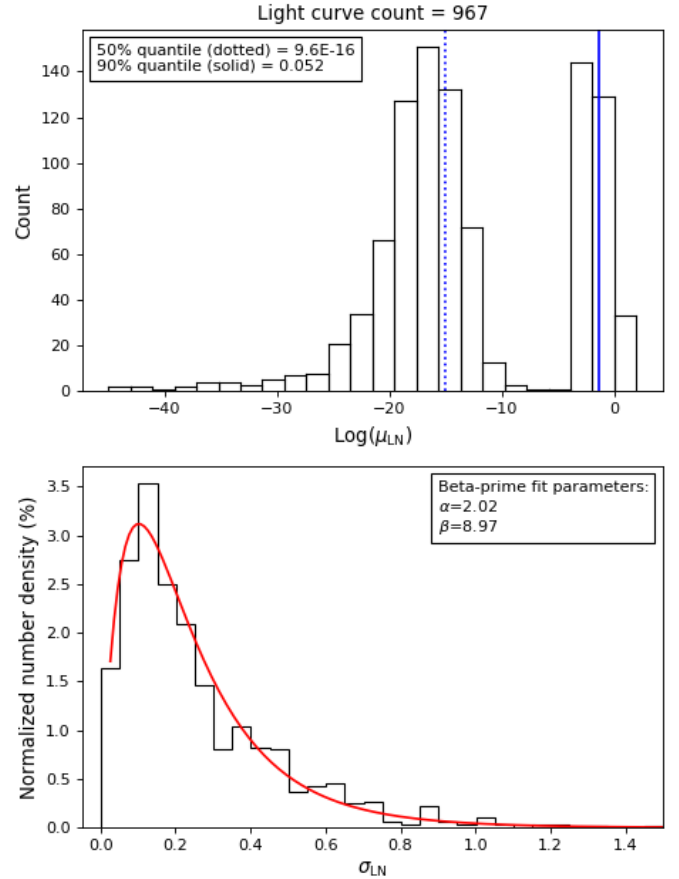


Fig. 4. Global distribution of the lognormal best-fit parameters (μ_{LN} and σ_{LN}) on all 967 CAZ flux density distributions. *Top*: Distribution of μ_{LN} in log-scale. *Bottom*: Distribution of σ_{LN} and its best-fit Beta-prime distribution (shown as the red curve with the parameters $\alpha = 2.02$ and $\beta = 8.97$).

3. Correlation tests

To search for a potential spatio-temporal correlation between blazars and neutrinos, we closely followed the methodology presented in K24, which in turn was based on our original study, H21. This correlation test begins by quantifying one or more global test-statistic (TS) parameters derived from variability measures of the spatially neutrino-associated blazars. It then determines the chance probability of these TS parameters by comparing them to those obtained from repeated Monte Carlo iterations in which the neutrinos are randomly shifted in RA⁷. These chance probabilities are used to test and potentially reject physically motivated null-hypotheses.

In our simulations, we constructed a “spatial-only” TS parameter based on the time-averaged variability measure (F_{var}) and a spatio-temporal TS parameter based on the time-resolved variability measure (AI); these are described in Sect. 2.2. In K24, we searched for a blazar–neutrino association using several test strategies, involving both spatial-only and spatio-temporal TS parameters. Since these searches were performed on real data, with the intent of establishing a correlation, all our significance levels had to be corrected for multiple trials. However, in this study, we simulated blazars with predetermined level of correlation to neutrinos (see Sect. 3.1); this allowed us to assess the

⁷ Randomizing neutrinos must only be done by shifting them in RA because neutrinos are not uniformly distributed in Dec..

detection power of each test strategy in a controlled manner. In Sects. 3.2, 3.3, and 3.4 we describe these strategies.

3.1. Simulating the blazar samples

In this subsection we describe the simulation of the blazar samples used in this paper with varying levels of induced blazar–neutrino correlation. We then evaluate these induced correlations via various test strategies to identify the method with the highest detection power. In Sect. 3.1.1 we explain the generation of a sample of 4000 random blazars with random variability measures, which serves as our initial sanity-check blazar sample. In Sect. 3.1.2 we describe the construction of four partially neutrino-associated blazar samples with one designed to represent a realistic scenario.

3.1.1. Random blazar sample

We generated a random population of 4000 blazars, each assigned with random time-averaged and time-resolved variability measures (F_{var} and AI; see Sect. 2.2). Given the uniform distribution of blazars across the extragalactic sky (e.g., Healey et al. 2007), generating random sky positions is straightforward. We drew 4000 random values uniformly from $0^\circ \leq \alpha < 360^\circ$ for RA (α), and 4000 from $-90^\circ \leq \delta \leq 90^\circ$ for Dec (δ). Using the observed distributions of variability measures (parameterized in Sect. 2.2), we assigned a random variability measure to each blazar. A random F_{var} value was sampled from a Beta-prime distribution with parameters $\alpha = 1.57$ and $\beta = 5.76$, resembling the observed CAZ F_{var} distribution (see Fig. 2).

Next, we generated random AI values for each random blazar. Since AI values are associated with individual neutrino arrival time, we only required 283 AI values for each blazar. Assuming neutrino arrival times are independent⁸ and that no two neutrinos from a single blazar flare are likely to be detected (e.g., Oikonomou et al. 2019), randomly selecting flux density values from a light curve is mathematically equivalent to randomly sampling flux density values from the flux density distribution of the light curve. Thus, for each random blazar, we drew 283 values from a random flux density distribution. These distributions were generated for each blazar using a lognormal distribution with parameters $\mu_{\text{LN}} = 0$ and σ_{LN} randomly sampled from a Beta-prime distribution with parameters $\alpha = 2.02$ and $\beta = 8.97$ (see Fig. 4).

This sample constitutes our “sim-null” dataset (the purple circle in Fig. 5), which is by construction purely random and serves as our initial sanity-check sample. No correlation should exist between the sim-null blazars and IceCat1+ neutrinos, although low p-values may arise due to statistical fluctuations (i.e., Gaussian statistics dictates 2σ , 3σ , and 4σ p-values arising in 4.55%, 0.27%, and $<0.01\%$ of the simulations, respectively).

3.1.2. Partially neutrino-associated blazar samples

Next, we simulate a small set of blazars, each spatio-temporally associated with a neutrino, and add them to the sim-null sample. In total, we construct four partially neutrino-associated blazar samples: one serves as a second sanity check, two are

special-case scenarios, and one represents a potentially realistic scenario.

To construct the first special-case scenario, we identify the best neutrinos in IceCat1+, specifically those with a very high probability of being astrophysical ($S > 0.85$) and well-localization ($\Omega < 1 \text{ deg}^2$). Four neutrinos⁹ meet these criteria. For each, we simulate a blazar centered on its maximum-likelihood arrival direction, with a Gaussian spread that corresponds to the $\geq 90\%$ -likelihood error region (see Sect. 2.1). Subsequently, we assign each blazar a random F_{var} value of at least 0.37, corresponding to the top 30% of the observed F_{var} distribution (see Fig. 2). To assign 283 random AI values to each blazar, we generate four random flux density distributions as described in Sect. 3.1.1 with additional constraint that σ_{LN} is at least 0.1¹⁰. From each distribution, we sample 282 random AI values and one AI value greater than 1.25 (see K24 for justification). Thus, each simulated blazar has a pseudo-random $F_{\text{var}} > 0.37$, 282 random AI values, and one pseudo-random AI > 1.25 (corresponding to the arrival time of the spatially associated neutrino). Adding these four pseudo-random blazars to the sim-null sample yields the “sim-best” sample, comprising 4004 blazars (the orange set in Fig. 1). This allows us to investigate the impact of four reliably reconstructed neutrino-associated blazars on the detection power of different test strategies.

To construct the second special-case blazar sample, we focus on neutrinos with $0.50 < S < 0.70$ and $5 < \Omega < 10 \text{ deg}^2$. 13 neutrinos¹¹ meet these “mid-range” criteria. Following the same process as above, we simulate 13 pseudo-random blazars with pseudo-random signals associated with these mid-range neutrinos. Adding these to sim-null sample produces the “sim-mid” sample, comprising of 4013 blazars (the green set in Fig. 1). This sample allows us to evaluate how a dozen blazars associated with mid-range neutrino events affect the detection power of various test strategies.

The remaining two simulated blazar samples represent broader scenarios. In one, all astrophysical neutrinos in IceCat1+ are assumed to be emitted by blazars (“sim-S”), which serves as a second sanity check. The other scenario assumes that 20% of astrophysical neutrinos are emitted by blazars (“sim-0.2S”), representing a potentially realistic scenario. The “sim-S” scenario is highly exaggerated as sources like NGC 1068 (IceCube Collaboration et al. 2022) and non-AGN phenomena like tidal disruption events (van Velzen et al. 2024) and microquasars (Koljonen et al. 2023) could be significant contributors to astrophysical neutrinos. Nevertheless, blazars are estimated to account for $\sim 20\%$ of astrophysical neutrinos in the ~ 100 – 1000 TeV range (e.g., Aartsen et al. 2017; Huber 2019; Das et al. 2021; Oikonomou 2022). Therefore, the sim-0.2S scenario represents a potentially realistic scenario.

In sim-0.2S and sim-S, the inclusion of neutrinos is determined by their signalness (S). For each neutrino, we drew a random number uniformly between 0 and 1. If this number exceeded S , the neutrino was excluded from sim-S. Similarly, if the number exceeded $0.2 \times S$, the neutrino was excluded from sim-0.2S. Repeating this for all 283 neutrinos in IceCat1+, we expected around $283 \times \bar{S} = 283 \times 0.429 \approx 121$ neutrinos in sim-S and around $283 \times 0.429 \times 0.2 \approx 24$ neutrinos in sim-0.2S. We

⁹ IC101028X, IC140611A, IC171106A, and IC201007A.

¹⁰ This condition makes it easier to generate above-average AI values, which improves computation speed.

¹¹ IC110902A, IC120515A, IC131108A, IC131124A, IC140101A, IC160225A, IC170626A, IC170704A, IC170819A, IC170923A, IC180417A, IC190730A, and IC201221A.

⁸ In practice this holds, since the current generation of neutrino observatories have rather poor detection rates of astrophysical neutrinos (e.g., for IceCube, this detection rate is on average ~ 1 per month).

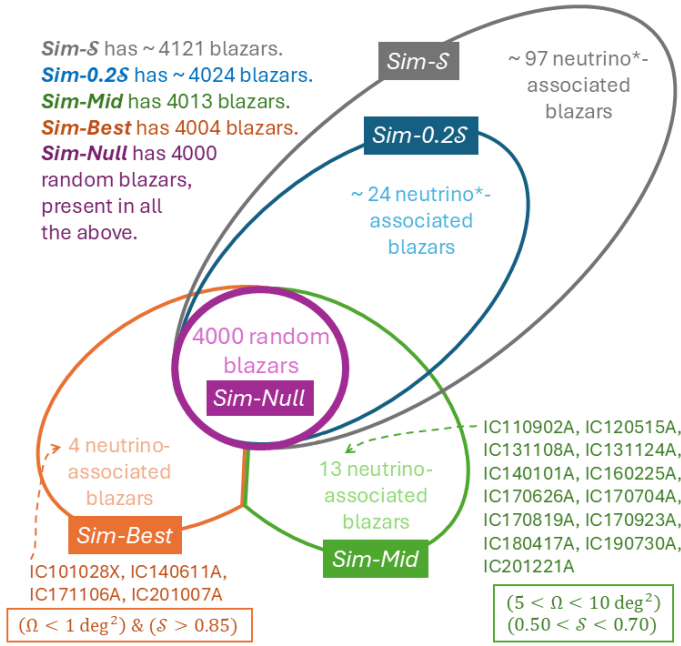


Fig. 5. Venn diagram of the five simulated blazar samples. All contain the 4000 random blazars of sim-null. Sim-best and sim-mid do not overlap, whereas sim-0.2S is contained entirely in sim-S. In sim-best and sim-mid, there are 4 and 13 neutrinos, respectively, which remain the same across all simulations. In sim-S and sim-0.2S, neutrino* indicates their S-based statistical nature.

emphasize that all ~ 24 neutrinos in sim-0.2S are also part of sim-S, leaving ~ 97 neutrinos unique to sim-S.

Following the same procedure as for sim-best and sim-mid, we generated pseudo-random blazars for the accepted neutrinos in sim-S and sim-0.2S. Consequently, sim-S and sim-0.2S comprise approximately 4121 and 4024 blazars, respectively, although the exact numbers vary from due to their statistical nature – this is why we repeated generating these blazar samples 1000 times in this study. In Fig. 5, sim-S is the gray set, and sim-0.2S is the blue set contained within sim-S.

We emphasize that the 4000 random blazars of sim-null, forming the majority of each simulated sample, are identical across a single simulation loop. Therefore, any difference in the correlation strength of test strategies, between the samples in the same loop, reflects the strength of the induced signal, and not random fluctuations in sim-null. Lastly, all blazars in sim-0.2S are included in sim-S. Meanwhile, apart from the 4000 random blazars, sim-best and sim-mid are mutually exclusive, though they may overlap with sim-0.2S and sim-S. In Fig. 5 we show a Venn diagram that visualizes these relations.

3.2. Test-statistic parameters

When a blazar and a neutrino coincide spatially (see Sect. 3.3.1), they are included in the calculations for our correlation test. Depending on the blazar and neutrino sample sizes, there are typically hundreds of such spatial associations. Each spatial association is accompanied by a variability measure (either F_{var} or AI). To construct a single global TS parameter, we can either average a given variability measure or count the occurrences where a given variability measure exceeds a predetermined threshold value. We emphasize that these two distinct methods of calculating the global TS parameter (“averaging” and “counting”) can be applied independently to each variability measure (F_{var} or AI).

The global TS parameters are calculated in both observed and Monte Carlo (random) setups (see Sect. 3). The probability (p) that an observed TS parameter occurs randomly is calculated as

$$p = \frac{M + 1}{N + 1}, \quad (3)$$

where M is the number of random TS parameters greater than or equal to the observed TS parameter, and N is the total number of Monte Carlo iterations (Davison & Hinkley 1997). In this study, $N = 10^4$, meaning the smallest detectable p-value is 0.0001 (i.e., $\sim 4\sigma$ significance level).

When TS parameters are constructed via counting, the pre-determined thresholds for F_{var} and AI are consistent with those used to simulate pseudo-random signals (see Sect. 3.1.2). This means that a blazar located near a neutrino event is simulated to be critically variable ($F_{\text{var}} > 0.37$) and flaring (AI > 1.25) at the arrival time of the neutrino. Therefore, the counting-based test strategies used in this study are designed to perform ideally.

3.3. Handling poorly reconstructed neutrino events

Our correlation test depends on a “spatial” coincidence between a blazar and a neutrino (see more details in Sect. 3.3.1). While the sky coordinates of blazars are typically known down to an arcsec accuracy, the sky error regions for IceCube high-energy neutrino events are generally much larger, often spanning several and sometimes tens of degrees (see Sect. 2.1). Including neutrinos with very large error regions in the analysis without proper treatment can result in spurious associations with many random blazars, reducing detection power.

Previous studies, including H21, addressed this issue by applying neutrino sample cuts, where neutrino events failing to meet predetermined criteria (e.g., error region size) were excluded from the analysis. An alternative approach is to introduce a weighting scheme, as used in K24, that reduces the contribution of poorly reconstructed events to the global TS parameter. In Sect. 3.3.1 we introduce four weighting schemes that are investigated in this study. In Sect. 3.3.2 we explore the impact of implementing neutrino sample cuts. We note that these two techniques can be utilized simultaneously; thus, we systematically studied their individual and combined effects on the detection power of our correlation test.

3.3.1. Weighting schemes

We considered four general weighting schemes in this study: two “unweighted” and two “weighted”. The unweighted schemes, referred to as \emptyset_W , act the control setups where none of the associations are weighted; for these, we used two different neutrino error region sizes (more details given below). On the other hand, the weighted schemes include a “top-hat” weighting approach (W_T) and a “Gaussian” weighting approach (W_G).

The top-hat weighting scheme, introduced in K24, uses the $\geq 90\%$ -likelihood error region of a neutrino (see Sect. 2.1) as its spatial boundary. Here, R_ϕ represents the distance between the center and the edge of the error region at a given phase angle ϕ , while $d_{\text{BN},\phi}$ represents the distance between the blazar and the center of the error region along ϕ (see Fig. 1). The top-hat weight (W_T) is defined as

$$W_T = \begin{cases} 0 & \text{if } d_{\text{BN},\phi} > R_\phi \\ S & \text{if } d_{\text{BN},\phi} \leq R_\phi \text{ and } \Omega \leq \tilde{\Omega} \\ S \cdot \tilde{\Omega} / \Omega & \text{if } d_{\text{BN},\phi} \leq R_\phi \text{ and } \Omega > \tilde{\Omega}, \end{cases} \quad (4)$$

where $\widetilde{\Omega} = 6.63 \text{ deg}^2$ is the global median of Ω (see Sect. 2.1). We note that $W_T \propto \Omega^{-1}$ penalizes poorly localized events when $\Omega > \widetilde{\Omega}$, while $\partial W_T / \partial \Omega = 0$ ensures that well-localized events do not dominate the statistics when $\Omega \leq \widetilde{\Omega}$.

If a blazar falls outside of the neutrino $\geq 90\%$ -likelihood error region (i.e., $d_{\text{BN},\phi} > R_\phi$), it is omitted from the association calculations in the top-hat weighting scheme (i.e., $W_T = 0$). On the other hand, W_T does not penalize associations based on the distance between the center of the neutrino error region and the blazar (i.e., $\partial W_T / \partial d_{\text{BN},\phi} = 0$). In other words, W_T remains the same for both on-center associations ($0 \lesssim d_{\text{BN},\phi} \ll R_\phi$) and off-center ones ($d_{\text{BN},\phi} \lesssim R_\phi$). Thus, if the neutrino error regions are underestimated or systematically shifted from their true centers, genuine associations could be missed. This is a potential caveat of the top-hat weighting scheme.

As an alternative, we introduced a Gaussian weighting scheme that explicitly incorporates \mathcal{S} , Ω , and $d_{\text{BN},\phi}$. It penalizes off-center spatial associations gradually, without requiring a hard “edge”. The Gaussian weight (W_G) is defined as

$$W_G = \mathcal{S} \cdot \frac{\Omega_{\min}}{\Omega} \cdot \exp\left[-\frac{1}{2} \left(\frac{2 d_{\text{BN},\phi}}{R_\phi}\right)^2\right], \quad (5)$$

where $\Omega_{\min} = 0.115 \text{ deg}^2$ (the smallest global Ω) is used as a scaling factor. Figure 1 visualizes the spatial parameters used in Eq. (5).

In the Gaussian weighting scheme all blazar–neutrino pairs are spatially “associated”. However, most associations at substantial distances ($d_{\text{BN},\phi} \gg R_\phi$) yield negligible W_G ; for example, $W_G|_{d=3R} / W_G|_{d=R} = e^{-16} \approx 10^{-7}$ implies that an association occurring at a distance of three times the reported IceCube error bar has a weight roughly seven orders of magnitude smaller than for an event occurring at the reported IceCube error bar. To optimize computational efficiency, we imposed an upper limit: if $d_{\text{BN},\phi} > 3R_\phi$, then $W_G = 0$. The threshold of $3R_\phi$ balances precision with computational efficiency.

Thus, in this study we employed two error region thresholds ($3R_\phi$ and R_ϕ): both are used in the unweighted schemes, while the Gaussian weighting scheme uses $3R_\phi$ and the top-hat weighting scheme uses R_ϕ . We note that the unweighted schemes are top-hat-like, since the associations with blazars outside the error region are assigned a weight of zero, while those inside are given equal weights (i.e., a weight of one).

3.3.2. Neutrino sample cuts

Another method for dealing with the least reliably reconstructed neutrino events is to apply predetermined cuts to the neutrino sample. Leveraging the simulations used in this study, we can access the impact of these cuts on detection power. We used three samples of neutrinos, with one including the full set of 283 neutrinos of the IceCat1+ sample, referred to as the “no-cut” sample. A second sample, referred to as “soft-cut”, is constructed by selecting all neutrinos from IceCat1+ with $\Omega < 50 \text{ deg}^2$, resulting in 245 neutrinos. Finally, we constructed a third sample, referred to as “hard-cut”, by requiring $\Omega < 10 \text{ deg}^2$ and $\mathcal{S} > 0.5$, producing a set of 71 neutrinos.

3.4. Final correlation test strategies

Figure 6 visualizes our strategies for performing a correlation test on a simulated blazar sample. By combining the methods constructing the TS parameters (see Sect. 3.2), the weighting

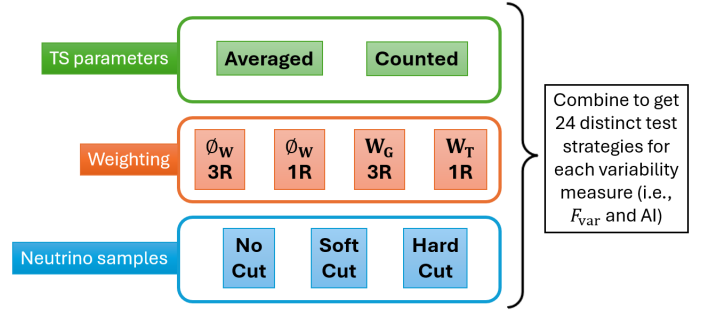


Fig. 6. Visualization of the 24 possible test strategies applicable to the five blazar samples and the two variability measures. There are two options for constructing the global TS parameters (see Sect. 3.2), four for weighting (3R and 1R refer to when $3R_\phi$ and R_ϕ are used as the error region edges, respectively; see Sect. 3.3.1), and three for neutrino sample cuts (see Sect. 3.3.2).

schemes (see Sect. 3.3.1), and the different neutrino samples (see Sect. 3.3.2), we derived $2 \times 4 \times 3 = 24$ unique test strategies. Since these can be applied to each of the five simulated blazar samples (see Fig. 5), we were left with $24 \times 5 = 120$ correlation tests for each of the two variability measures of interest: the time-averaged F_{var} and the time-resolved AI (see Sect. 3). Thus, we obtained a total of $120 \times 2 = 240$ blazar–neutrino correlation strengths (i.e., p-values) in a single simulation step. By investigating the behavior of these p-values over 1000 simulation steps, we then determined the most optimal strategies for reliably detecting a potential blazar–neutrino correlation.

4. Results and discussion

Using 24 different test strategies, we obtained spatio-temporal correlation strengths (i.e., p-values) between a simulated blazar sample and IceCat1+. Since the simulated blazar sample is identical across all tests, the test strategy that results in the smallest p-value is the most optimal (i.e., it offers the greatest detection power). This comparison would not be reliable with only a single simulated blazar sample due to statistical fluctuations. Thus, we repeated the simulation process 1000 times, obtaining 1000 p-values for each test strategy to ensure robust comparisons of the p-values. Notably, since we did not perform hypotheses testing (as the underlying blazar–neutrino correlation is simulated and known), trial correction was unnecessary. The distributions of these p-values are shown in Figs. A.1–A.10, where we also provide the fraction of p-values crossing the 2σ , 3σ , and $\sim 4\sigma$ thresholds along with their mean and standard deviation. Here, we focus on the fraction of p-values crossing the 3σ threshold (i.e., $p < 0.0027$) as a measure of quantifying and comparing the quality of different test strategies; we denote these fractions as $\mathcal{F}_{3\sigma}$. Table 1 summarizes all $5 \times 2 \times 4 \times 3 \times 2 = 240$ values of $\mathcal{F}_{3\sigma}$ (given in %). Below we discuss the implications of these 240 results.

4.1. Sanity checks: sim-null and sim-S

The sim-null blazar sample does not have any simulated blazar–neutrino connection (see Sect. 3.1.1); whereas, in sim-S, all astrophysical neutrinos are simulated as originating from blazars (see Sect. 3.1.2). These samples serve as sanity checks of our simulation setup and correlation test results. As shown in Table 1, under all test strategies, the sim-null results are consistent with random fluctuations occurring within the 3σ limits; the

Table 1. Fraction of simulations crossing the 3σ threshold (i.e., $\mathcal{F}_{3\sigma}$ given in percent). The total number of simulations is 1000.

Sim (1)	TS			F_{var}			AI		
	(2)	W (3)	No-cut (4)	Soft-cut (5)	Hard-cut (6)	No-cut (7)	Soft-cut (8)	Hard-cut (9)	
Sim-null (sanity check)	Averaged	\emptyset_W 3R	0.1%	0.3%	0.1%	0.1%	0.1%	0.4%	
	Averaged	\emptyset_W 1R	0.2%	0.2%	0.2%	0.2%	0.2%	0.5%	
	Counted	\emptyset_W 3R	0.0%	0.1%	0.2%	0.1%	0.4%	0.1%	
	Counted	\emptyset_W 1R	0.2%	0.2%	0.3%	0.1%	0.3%	0.1%	
	Averaged	W_G 3R	0.3%	0.4%	0.3%	0.1%	0.1%	0.0%	
	Averaged	W_T 1R	0.2%	0.1%	0.3%	0.5%	0.4%	0.5%	
	Counted	W_G 3R	0.3%	0.3%	0.3%	0.1%	0.1%	0.1%	
	Counted	W_T 1R	0.3%	0.3%	0.5%	0.3%	0.6%	0.1%	
Sim-best (special-case)	Averaged	\emptyset_W 3R	0.2%	0.4%	0.5%	0.3%	0.1%	0.7%	
	Averaged	\emptyset_W 1R	0.2%	0.6%	0.6%	0.4%	0.2%	0.6%	
	Counted	\emptyset_W 3R	0.0%	0.4%	2.2%	0.4%	1.1%	4.7%	
	Counted	\emptyset_W 1R	0.3%	0.5%	4.7%	0.4%	1.2%	8.8%	
	Averaged	W_G 3R	89.6%	89.1%	84.4%	83.6%	83.8%	83.7%	
	Averaged	W_T 1R	5.9%	6.3%	2.8%	2.2%	2.3%	2.0%	
	Counted	W_G 3R	96.7%	96.7%	96.8%	98.7%	98.7%	98.7%	
	Counted	W_T 1R	8.4%	9.3%	20.1%	24.0%	25.3%	47.4%	
Sim-mid (special-case)	Averaged	\emptyset_W 3R	0.5%	1.7%	3.0%	0.3%	0.2%	1.2%	
	Averaged	\emptyset_W 1R	0.9%	1.8%	1.0%	0.7%	0.5%	0.4%	
	Counted	\emptyset_W 3R	0.8%	2.7%	55.7%	1.5%	9.9%	97.2%	
	Counted	\emptyset_W 1R	2.4%	8.8%	94.4%	4.4%	25.3%	100.0%	
	Averaged	W_G 3R	0.4%	0.4%	0.1%	0.2%	0.2%	0.0%	
	Averaged	W_T 1R	25.2%	17.8%	0.2%	8.9%	5.4%	0.1%	
	Counted	W_G 3R	0.8%	0.8%	0.8%	0.4%	0.4%	0.2%	
	Counted	W_T 1R	46.8%	49.2%	82.3%	87.5%	87.9%	99.7%	
Sim-0.2S (realistic)	Averaged	\emptyset_W 3R	1.8%	4.2%	1.7%	0.6%	0.6%	0.6%	
	Averaged	\emptyset_W 1R	3.9%	7.4%	0.6%	0.5%	0.7%	0.4%	
	Counted	\emptyset_W 3R	5.0%	14.2%	24.8%	16.7%	44.2%	55.6%	
	Counted	\emptyset_W 1R	15.6%	43.2%	55.1%	36.9%	81.7%	80.3%	
	Averaged	W_G 3R	53.6%	50.0%	31.8%	46.8%	42.9%	27.9%	
	Averaged	W_T 1R	58.9%	41.7%	2.1%	26.6%	14.8%	1.0%	
	Counted	W_G 3R	68.7%	68.6%	58.8%	83.6%	83.0%	72.9%	
	Counted	W_T 1R	84.8%	84.8%	69.7%	97.8%	97.5%	88.4%	
Sim-S (sanity check)	Averaged	\emptyset_W 3R	61.7%	80.0%	38.9%	14.6%	25.0%	13.0%	
	Averaged	\emptyset_W 1R	88.2%	95.3%	2.2%	29.0%	29.3%	1.3%	
	Counted	\emptyset_W 3R	99.6%	100.0%	100.0%	100.0%	100.0%	100.0%	
	Counted	\emptyset_W 1R	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
	Averaged	W_G 3R	100.0%	100.0%	98.8%	97.5%	97.8%	95.5%	
	Averaged	W_T 1R	100.0%	100.0%	13.5%	88.9%	77.7%	7.1%	
	Counted	W_G 3R	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
	Counted	W_T 1R	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	

Notes. Column (1) refers to the sample of simulated blazars (see Sect. 3.1). Column (2) refers to the TS parameter (see Sect. 3.2); Averaged and Counted refer to the averaging-based and counting-based TS parameters, respectively. Column (3) refers to the weighting scheme as well as the neutrino error region size (see Sect. 3.3.1); \emptyset_W , W_G , and W_T imply no weighting, Gaussian weighting, and top-hat weighting, respectively; 1R and 3R indicate the neutrino error region edge is R_ϕ and $3R_\phi$, respectively. Columns (4) and (7) refer to the no-cut neutrino subsample, columns (5) and (8) refer to the soft-cut neutrino subsample, and columns (6) and (9) refer to the hard-cut neutrino subsample (see Sect. 3.3.2). In columns (4)–(6), F_{var} is used as the variability measure; while in columns (7)–(9), AI is used as the variability measure.

average $\mathcal{F}_{3\sigma}$ is 0.23% (with a standard deviation of 0.14%), which closely matches the chance probability of obtaining a 3σ result from a completely random process (i.e., $p \approx 0.0027$).

In sim-S, most $\mathcal{F}_{3\sigma}$ are 100% (the average $\mathcal{F}_{3\sigma}$ is 80.3%, with a few test strategies yielding notably lower fractions). This demonstrates two key points: (1) most test strategies can consistently detect 3σ (even $\sim 4\sigma$; see Appendix A) blazar–neutrino associations; and (2) certain test strategies clearly underperform even when detecting the same underlying blazar–neutrino correlation. For example, combining the hard-cut sample with aver-

aging in the top-hat weighting scenario results in poor detection power.

4.2. Realistic scenario: sim-0.2S

In sim-0.2S, 20% of astrophysical neutrinos are simulated to be associated with blazars, a ratio compatible with observations (see Sect. 3.1.2). As shown in Table 1, $\mathcal{F}_{3\sigma}$ ranges from $\sim 98\%$ to $\sim 0\%$, which allows us to systematically disfavor underperforming strategies. Crucially, changing from

sim- S to sim-0.2 S results in reduced $\mathcal{F}_{3\sigma}$ for all test strategies, as expected.

4.2.1. Effect of averaging versus counting

We first compared the average $\mathcal{F}_{3\sigma}$ for tests whose global TS parameter was constructed via averaging to those constructed via counting in the sim-0.2 S scenario. The average $\mathcal{F}_{3\sigma}$ is 17.5% in the former and 59.7% in the latter. This clearly shows that counting consistently outperforms averaging. However, we note that the thresholds used in the counting strategies are ideal (see Sect. 3.2).

4.2.2. Effect of weighting

$\mathcal{F}_{3\sigma}$ increases when changing from unweighted (\emptyset_W 3R and \emptyset_W 1R) test strategies to their respective weighted ones (W_G 3R and W_T 1R). When considering only the counted results in sim-0.2 S , the average $\mathcal{F}_{3\sigma}$ for unweighted tests is 39.4% compared to 79.9% for the weighted ones. This demonstrates that utilizing a weighting scheme universally improves detection power, regardless of the test strategy. Further differences between Gaussian and top-hat weighting schemes are explored in Sect. 4.2.4.

4.2.3. Effect of setting neutrino sample cuts

For weighted and counted tests in sim-0.2 S , $\mathcal{F}_{3\sigma}$ decreases when progressing from the no-cut to soft-cut and then to hard-cut tests (average $\mathcal{F}_{3\sigma}$ is 83.7%, 83.5%, and 72.4%, respectively). This indicates that once weighting is applied, removing neutrinos from the analysis generally reduces detection power. As argued in K24, weighting is preferable for suppressing noise from poorly reconstructed neutrinos. Another caveat of setting hard cuts is increased likelihood of false-negatives when the TS parameter is expected to be a small, whole number (e.g., in counted-AI tests; see Appendix A). Therefore, we recommend using the entire sample of neutrinos in combination with a weighting scheme. However, if sample cuts are necessary for practical reasons (e.g., increasing computational efficiency), they should be kept as minimal as possible.

4.2.4. Weighting scheme: Gaussian versus top-hat

In the optimal test strategy where counting and weighting (with no cuts) are employed, the average $\mathcal{F}_{3\sigma}$ in the Gaussian (W_G 3R) and top-hat (W_T 1R) weighting schemes are 76.2% and 91.3%, respectively. This indicates that the top-hat weighting scheme offers greater detection power than the Gaussian one. Despite the simulated errors being Gaussian, the superior performance of the top-hat scheme can be understood by examining the results of the special-case tests, which are discussed below.

4.2.5. Special-case tests: sim-best and sim-mid

In sim-best, the simulated signal originates from four of the most reliably reconstructed neutrinos; whereas in sim-mid, it comes from 13 mid-range ones (see Sect. 3.1.2). Using Gaussian weighting (in the no-cut and counted test), $\mathcal{F}_{3\sigma}$ for sim-best is 96.7%, while for sim-mid, the same strategy results in $\mathcal{F}_{3\sigma} = 0.8\%$. However, with top-hat weighting, sim-best gives $\mathcal{F}_{3\sigma} = 8.4\%$, while sim-mid gives $\mathcal{F}_{3\sigma} = 46.8\%$. Strikingly, in sim-best, Gaussian weighting outperforms even the sim-0.2 S , which demonstrates its extreme sensitivity to the most

reliably reconstructed events (i.e., only a few well-localized associations can dominate its statistics). This hampers the usefulness of Gaussian weighting in population-based blazar–neutrino correlation tests, because mid-range events – even after accounting for their lower signalness – are more numerous than well-localized neutrinos. For example, there are 220 neutrinos satisfying $0.2 < S < 0.6$ compared to 56 satisfying $0.6 < S < 1.0$; after adjusting for signalness, we expected $220 \times 0.4 = 88$ and $56 \times 0.8 \approx 45$ astrophysical neutrinos in each category to be potentially associated with blazars. While many mid-range neutrinos have rather poor sky localizations (i.e., large Ω), they contain the majority of potential blazar–neutrino associations. Therefore, the ability of top-hat weighting to recover signal from these mid-range events allows it to outperform Gaussian weighting on average. Thus, we conclude that the top-hat weighting scheme is generally preferable in the context of spatio-temporal correlation tests.

4.2.6. Robustness of the top-hat weighting scheme

A potential disadvantage of the top-hat weighting (as compared to Gaussian) is its dependence on the choice of error region size (see Sect. 3.3.1). However, here we demonstrate its robustness against this parameter. We compared the results of the unweighted (\emptyset_W) 1R and 3R tests, whose contributions to the statistics are top-hat-like (see Sect. 3.3.1). We note that: (1) this 1R-to-3R comparison is not possible for W_T as it does not have a 3R counterpart; and (2) this comparison is not applicable to W_G by definition. When changing from \emptyset_W 1R to 3R in sim-null, increasing the neutrino error region size does not arbitrarily induce false-positives. Meanwhile, when a real blazar–neutrino association exists (e.g., sim-0.2 S), some unweighted 3R tests result in 3σ associations. Interestingly, the detection power in the unweighted tests tends to decrease when changing from 1R to 3R, implying that enlarging the error region provides conservative estimate of the correlation strength. In summary, using \emptyset_W as a tracer of top-hat-like weighting schemes shows that: (1) the choice of the error region size does not result in false-positives; and (2) enlarging the error region size yields conservative estimates of the strength of the underlying blazar–neutrino correlation. The latter point is especially useful for correlation tests using IceCube high-energy neutrino events, as these may include unknown systematic errors¹² (e.g., Aartsen et al. 2013; Abbasi et al. 2021c). Consequently, several studies have enlarged the IceCube error regions when searching for blazar–neutrino correlations (e.g., Plavin et al. 2020; H21; K24), the robustness of which is now confirmed by our simulations.

4.2.7. Variability measure: F_{var} versus AI

Finally, we examined the results of the two variability measures, F_{var} and AI. As discussed in Sect. 1, these measures test different physical hypotheses regarding the connection between blazar variability and neutrino emission. When neutrinos are produced during individual flares (as opposed to being produced in highly variable sources), the number of 3σ associations is on average higher. Nevertheless, we emphasize that testing for both hypotheses using the optimal test strategies is feasible, as both yield relatively high fractions of 3σ associations.

¹² This can be due to various factors, which may include: the unknown kinematic angle between the neutrino and the muon, the unknown scattering properties of the ice, and the lack of knowledge about the exact location of the optical modules.

4.3. Implications for H21 and K24

With the identification of counting and top-hat weighting (without neutrino sample cuts) as the most optimal strategy, we now revisit our previous studies (H21 and K24), where we used flux density and AI to construct the TS parameters (in the radio band for H21, and in the radio and optical bands for K24). While we cannot directly comment on average flux density as it is not a variability measure in this study, we can provide insights into AI, although we note that its definition somewhat differed in those studies (see Sect. 2.2). In both H21 and K24, we considered one averaging and two counting scenarios to construct the global AI-based TS parameter. In H21, we employed a cut on the neutrino sample comparable to the hard-cut used in this study (no weighting was utilized). Conversely, in K24, we omitted neutrino sample cuts but introduced and employed top-hat weighting. First, in both studies, the use of counting (rather than averaging) resulted in the strongest associations, aligning with our findings here when a real blazar–neutrino association exists. Second, in hindsight, the search criteria employed in K24 appear to have been the most optimal and resulted in $p = 0.0377$ (before trial correction¹³, under the most directly comparable scenario). This p-value is significantly higher than expected if 20% of astrophysical neutrinos are associated with blazars (in which case our simulations predict a 97.8% chance of $p < 0.0027$). This suggests that the fraction of astrophysical neutrinos from blazars is substantially lower than 20%. However, we strongly caution against over-interpreting this implication. In K24, the analysis was based on 1061 optical light curves whereas in this study we use ~4000 light curves, and as shown in L22, the number of light curves is critically important to the robustness of the results.

5. Summary and conclusions

In this study we simulated five different blazar samples with varying degrees of spatio-temporal correlation to high-energy neutrino events from IceCat1+ (see Sects. 2.1, 2.2, and 3.1). Two of these blazar samples served as sanity checks, two address special-case scenarios, and one represents a potentially realistic test. We then investigated the detection power of our spatio-temporal correlation test using 24 different strategies (see Sect. 3.4). Namely, we considered the impact of constructing TS parameters using time-averaged and time-resolved variability measures, employing both averaging-based and counting-based methods (see Sect. 3.2). We also investigated the effects of using unweighted or weighted schemes (top-hat and Gaussian; see Sect. 3.3.1) and considered the implications of applying cuts to the neutrino sample (see Sect. 3.3.2). The results and their implications are discussed in Sect. 4. Below, we summarize the most important findings:

- Both time-averaged and time-resolved variability measures can achieve $\sim 4\sigma$ blazar–neutrino associations.
- When calculating the global TS parameters, counting-based methods outperform averaging-based approaches.
- Weighing down the associations involving poorly reconstructed events is more effective than applying arbitrary cuts to the neutrino sample.
- Top-hat weighting is preferable to Gaussian weighting, as it is less sensitive to the most reliably reconstructed events and better recovers signals from the more numerous mid-range events.

- When utilizing top-hat-like weighting schemes, enlarging the neutrino error region does not arbitrarily increase the number of false-positives, and such enlargements can help in estimating the underlying correlation conservatively.

We highlight two caveats associated with this study: (1) while we consider signalness to ideally represent the likelihood that an IceCube high-energy neutrino event is of astrophysical origin, in practice, it is only a rough estimate; and (2) the counting scheme is constructed to be optimal (i.e., the neutrino-associated blazars are considered highly variable), and thus the conclusion that the counting-based approach is universally better may not hold as strongly in practice. Nevertheless, these simulations provide a potential method for further constraining the maximum fraction of astrophysical neutrinos originating from blazars.

In general, this kind of a priori fine-tuning of test strategies is essential for addressing the blazar–neutrino connection without introducing unnecessary trials when analyzing real data, especially with the advent of next-generation scientific instruments. For example, the presented methods are particularly useful for determining whether specific subpopulations of blazars are stronger neutrino emitters (e.g., certain high synchrotron peaked blazars, as suggested by Padovani et al. 2016).

The next generation of neutrino observatories such as IceCube-Gen2 (IceCube-Gen2 Collaboration et al. 2014), KM3NeT (Adrián-Martínez et al. 2016), Baikal-GVD (Shoibonov & Baikal Collaboration 2019), RNO-G (Aguilar et al. 2021), and GRAND (Fang et al. 2017) are under development. Additionally, upcoming all-sky surveys, such as the Vera C. Rubin Observatory’s Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al. 2009), the Cosmic Microwave Background Stage 4 project (CMB-S4; Abazajian et al. 2022), and the Simons Observatory (SO; Ade et al. 2019), will revolutionize transient astronomy and provide an unprecedented number of light curves with unparalleled cadence for variable sources (several of which are potential neutrino factories, e.g., blazars, AGNs in general, tidal disruption events, etc.). Moreover, there are already plans to study the central engines of neutrino-emitting blazars down to (sub)parsec-scale resolutions with the next generation Event Horizon Telescope (ngEHT; e.g., Lico et al. 2023; Kovalev et al. 2023). Therefore, studies similar to those presented here are critical for developing the optimal test strategies needed to effectively analyze the wealth of data from these future observatories.

Acknowledgements. We thank the anonymous referee for their highly insightful and constructive comments that helped improve and clarify the paper. P.K. was supported by Academy of Finland projects 346071 and 345899. E.L. was supported by Academy of Finland projects 317636, 320045, and 346071. T.H. was supported by Academy of Finland projects 317383, 320085, 322535, and 345899. J.J. was supported by Academy of Finland projects 320085 and 345899. K.K. acknowledges support from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 101002352). Based on observations obtained with the Samuel Oschin Telescope 48-inch and the 60-inch Telescope at the Palomar Observatory as part of the Zwicky Transient Facility project. ZTF is supported by the National Science Foundation under Grant No. AST-2034437 and a collaboration including Caltech, IPAC, the Weizmann Institute for Science, the Oskar Klein Center at Stockholm University, the University of Maryland, Deutsches Elektronen-Synchrotron and Humboldt University, the TANGO Consortium of Taiwan, the University of Wisconsin at Milwaukee, Trinity College Dublin, Lawrence Livermore National Laboratories, and IN2P3, France. Operations are conducted by COO, IPAC, and UW. The ZTF forced-photometry service was funded under the Heising-Simons Foundation grant #12540303 (PI: M.J.Graham). This work has made use of data from the Asteroid Terrestrial-impact Last Alert System (ATLAS) project. The Asteroid Terrestrial-impact Last Alert System (ATLAS) project is primarily funded to search for near

¹³ In K24, we did not obtain individual post-trial p-values, only a global post-trial p-value, which is not directly comparable to this test strategy.

earth asteroids through NASA grants NN12AR55G, 80NSSC18K0284, and 80NSSC18K1575; byproducts of the NEO smarch include images and catalogs from the survey area. This work was partially funded by Kepler/K2 grant J1944/80NSSC19K0112 and HST GO-15889, and STFC grants ST/T000198/1 and ST/S006109/1. The ATLAS science products have been made possible through the contributions of the University of Hawaii Institute for Astronomy, the Queen's University Belfast, the Space Telescope Science Institute, the South African Astronomical Observatory, and The Millennium Institute of Astrophysics (MAS), Chile. This work has made use of data from the Joan Oró Telescope (TJO) of the Montsec Observatory (OdM), which is owned by the Catalan Government and operated by the Institute for Space Studies of Catalonia (IEEC).

References

- Aartsen, M. G., Abbasi, R., Abdou, Y., et al. 2013, *Phys. Rev. Lett.*, **111**, 021103
- Aartsen, M. G., Abraham, K., Ackermann, M., et al. 2017, *ApJ*, **835**, 45
- Aartsen, M. G., Ackermann, M., Adams, J., et al. 2020, *Phys. Rev. Lett.*, **124**, 051103
- Abazajian, K., Abdulghafour, A., Addison, G. E., et al. 2022, ArXiv e-prints [arXiv:2203.08024]
- Abbasi, R., Ackermann, M., Adams, J., et al. 2021a, *Phys. Rev. D*, **104**, 022002
- Abbasi, R., Ackermann, M., Adams, J., et al. 2021b, *ApJ*, **920**, L45
- Abbasi, R., Ackermann, M., Adams, J., et al. 2021c, in *Proceedings of 37th International Cosmic Ray Conference - PoS(ICRC2021)*, 395, 1045
- Abbasi, R., Ackermann, M., Adams, J., et al. 2022, *ApJ*, **928**, 50
- Abbasi, R., Ackermann, M., Adams, J., et al. 2023, *ApJS*, **269**, 25
- Abbasi, R., Ackermann, M., Adams, J., et al. 2024, *ApJ*, **973**, 97
- Ade, P., Aguirre, J., Ahmed, Z., et al. 2019, *JCAP*, **2019**, 056
- Adrián-Martínez, S., Ageron, M., Aharonian, F., et al. 2016, *J. Phys. G Nucl. Phys.*, **43**, 084001
- Aguilar, J. A., Allison, P., Beatty, J. J., et al. 2021, *J. Instrum.*, **16**, P03025
- Shoibonov, B., & Baikal Collaboration 2019, *J. Phys. Conf. Ser.*, **1263**, 012005
- Bartos, I., Veske, D., Kowalski, M., Márka, Z., & Márka, S. 2021, *ApJ*, **921**, 45
- Becker, J. K. 2008, *Phys. Rep.*, **458**, 173
- Bellenghi, C., Padovani, P., Resconi, E., & Giommi, P. 2023, *ApJ*, **955**, L32
- Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, *PASP*, **131**, 018002
- Blandford, R., Meier, D., & Readhead, A. 2019, *ARA&A*, **57**, 467
- Blinov, D., Pavlidou, V., Papadakis, I. E., et al. 2016, *MNRAS*, **457**, 2252
- Böttcher, M. 2019, *Galaxies*, **7**, 20
- Bourguignon, M., Santos-Neto, M., & de Castro, M. 2018, ArXiv e-prints [arXiv:1804.07734]
- Buson, S., Tramacere, A., Pfeiffer, L., et al. 2022, *ApJ*, **933**, L43
- Buson, S., Tramacere, A., Oswald, L., et al. 2023, ArXiv e-prints [arXiv:2305.11263]
- Das, S., Gupta, N., & Razzaque, S. 2021, *ApJ*, **910**, 100
- Das, S., Gupta, N., & Razzaque, S. 2022, *A&A*, **668**, A146
- Davison, A. C., & Hinkley, D. V. 1997, in *Bootstrap Methods and their Application*, *Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press)
- Dermer, C. D., Murase, K., & Inoue, Y. 2014, *J. High Energy Astrophys.*, **3**, 29
- Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. 2009, *ApJ*, **696**, 870
- Fang, K., Alvarez-Muniz, J., Alves Batista, R., et al. 2017, in *35th International Cosmic Ray Conference (ICRC2017)*, 301, 996
- Filippenko, A. V., Li, W. D., Treffers, R. R., & Modjaz, M. 2001, in *Small Telescope Astronomy on Global Scales*, eds. B. Paczynski, W. P. Chen, & C. Lemme, *IAU Colloq.*, **183**, *ASP Conf. Ser.*, **246**, 121
- Franckowiak, A., Garrappa, S., Paliya, V., et al. 2020, *ApJ*, **893**, 162
- Giommi, P., Glauch, T., Padovani, P., et al. 2020, *MNRAS*, **497**, 865
- Gliozzi, M., Sambruna, R. M., & Eracleous, M. 2003, *ApJ*, **584**, 176
- Healey, S. E., Romani, R. W., Taylor, G. B., et al. 2007, *ApJS*, **171**, 61
- Hooper, D., & Plant, K. 2023, *Phys. Rev. Lett.*, **131**, 231001
- Hovatta, T., & Lindfors, E. 2019, *New Astron. Rev.*, **87**, 101541
- Hovatta, T., Pavlidou, V., King, O. G., et al. 2014, *MNRAS*, **439**, 690
- Hovatta, T., Lindfors, E., Blinov, D., et al. 2016, *A&A*, **596**, A78
- Hovatta, T., Lindfors, E., Kiehlmann, S., et al. 2021, *A&A*, **650**, A83
- Huber, M. 2019, in *36th International Cosmic Ray Conference (ICRC2019)*, **36**, 916
- IceCube Collaboration 2013, *Science*, **342**, 1242856
- IceCube Collaboration (Aartsen, M. G., et al.) 2018, *Science*, **361**, 147
- IceCube Collaboration (Abbasi, R., et al.) 2021, ArXiv e-prints [arXiv:2101.09836]
- IceCube Collaboration (Abbasi, R., et al.) 2022, *Science*, **378**, 538
- IceCube-Gen2 Collaboration (Aartsen, M. G., et al.) 2014, ArXiv e-prints [arXiv:1412.5106]
- Kadler, M., Krauß, F., Mannheim, K., et al. 2016, *Nat. Phys.*, **12**, 807
- Koljonen, K. I. I., Satalecka, K., Lindfors, E. J., & Liodakis, I. 2023, *MNRAS*, **524**, L89
- Kouch, P. M., Lindfors, E., Hovatta, T., et al. 2024, *A&A*, **690**, A111
- Kovalev, Y. Y., Plavin, A. V., Pushkarev, A. B., & Troitsky, S. V. 2023, *Galaxies*, **11**, 84
- Krauß, F., Kadler, M., Mannheim, K., et al. 2014, *A&A*, **566**, L7
- Kreter, M., Kadler, M., Krauß, F., et al. 2020, *ApJ*, **902**, 133
- Kun, E., Bartos, I., Becker Tjus, J., et al. 2022, *ApJ*, **934**, 180
- Lähteenmäki, A., & Valtaoja, E. 2003, *ApJ*, **590**, 95
- León-Tavares, J., Valtaoja, E., Tornikoski, M., Lähteenmäki, A., & Nieppola, E. 2011, *A&A*, **532**, A146
- Lico, R., Jorstad, S. G., Marscher, A. P., et al. 2023, *Galaxies*, **11**, 17
- Lindfors, E. J., Hovatta, T., Nilsson, K., et al. 2016, *A&A*, **593**, A98
- Lioudakis, I., Romani, R. W., Filippenko, A. V., et al. 2018, *MNRAS*, **480**, 5517
- Lioudakis, I., Hovatta, T., Pavlidou, V., et al. 2022, *A&A*, **666**, A36
- LSST Science Collaboration (Abell, P. A., et al.) 2009, ArXiv e-prints [arXiv:0912.0201]
- Lu, M. X., Liang, Y. F., Ouyang, X., Li, R. L., & Wang, X. G. 2024, ArXiv e-prints [arXiv:2404.19730]
- MAGIC Collaboration (Abe, S., et al.) 2024, *A&A*, **685**, A117
- Mannheim, K., & Biermann, P. L. 1989, *A&A*, **221**, 211
- Mannheim, K., Stanev, T., & Biermann, P. L. 1992, *A&A*, **260**, L1
- Mücke, A., & Protheroe, R. J. 2001, *Astropart. Phys.*, **15**, 121
- Nilsson, K., Lindfors, E., Takalo, L. O., et al. 2018, *A&A*, **620**, A185
- Novikova, P., Shishkina, E., & Blinov, D. 2023, *MNRAS*, **526**, 347
- Oikonomou, F. 2022, in *37th International Cosmic Ray Conference*, 30
- Oikonomou, F., Murase, K., Padovani, P., Resconi, E., & Mészáros, P. 2019, *MNRAS*, **489**, 4347
- Padovani, P., Resconi, E., Giommi, P., Arsioli, B., & Chang, Y. L. 2016, *MNRAS*, **457**, 3582
- Plavin, A., Kovalev, Y. Y., Kovalev, Y. A., & Troitsky, S. 2020, *ApJ*, **894**, 101
- Plavin, A. V., Kovalev, Y. Y., Kovalev, Y. A., & Troitsky, S. V. 2021, *ApJ*, **908**, 157
- Plavin, A. V., Kovalev, Y. Y., Kovalev, Y. A., & Troitsky, S. V. 2023, *MNRAS*, **523**, 1799
- Plavin, A. V., Burenin, R. A., Kovalev, Y. Y., et al. 2024, *JCAP*, **2024**, 133
- Richards, J. L., Max-Moerbeck, W., Pavlidou, V., et al. 2011, *ApJS*, **194**, 29
- Richards, J. L., Hovatta, T., Max-Moerbeck, W., et al. 2014, *MNRAS*, **438**, 3058
- Righi, C., Tavecchio, F., & Pacciani, L. 2019, *MNRAS*, **484**, 2067
- Smith, D., Hooper, D., & Viereg, A. 2021, *JCAP*, **2021**, 031
- Stathopoulos, S. I., Petropoulou, M., Giommi, P., et al. 2022, *MNRAS*, **510**, 4063
- Stecker, F. W. 2013, *Phys. Rev. D*, **88**, 047301
- Stecker, F. W., Done, C., Salamon, M. H., & Sommers, P. 1991, *Phys. Rev. Lett.*, **66**, 2697
- Suray, A., & Troitsky, S. 2024, *MNRAS*, **527**, L26
- Tonry, J. L., Denneau, L., Heinze, A. N., et al. 2018, *PASP*, **130**, 064505
- Troitsky, S. V. 2021, *Physics Uspekhi*, **64**, 1261
- van Velzen, S., Stein, R., Gilfanov, M., et al. 2024, *MNRAS*, **529**, 2559
- Vaughan, S., Edelson, R., Warwick, R. S., & Uttley, P. 2003, *MNRAS*, **345**, 1271
- Zhou, B., Kamionkowski, M., & Liang, Y.-F. 2021, *Phys. Rev. D*, **103**, 123018

Appendix A: Distributions of the simulation p-values

In this appendix we provide the p-value distributions of the 240 different correlation tests studied in this paper. Each of the 240 distributions contains 1000 p-values corresponding to 1000 simulation steps. Within each simulation step, we obtained a p-value by running a spatio-temporal correlation test on a sample of high-energy neutrino events and blazars (see Sect. 3).

Figures A.1 to A.10 show all the 240 p-value distributions of this study. These are divided into ten figures. At the top of each figure, a colored text card shows the variability measure and the blazar sample used in obtaining their respective p-value distributions. The odd numbered figures (A.1, A.3, A.5, A.7, and A.9) use F_{var} as the variability measure, while even numbered figures (A.2, A.4, A.6, A.8, and A.10) use AI. The blazar samples are as follows:

- Sim-null sample (purple text card; Figs. A.1 and A.2)
- Sim-best sample (orange text card; Figs. A.3 and A.4)
- Sim-mid sample (green text card; Figs. A.5 and A.6)
- Sim-0.2S sample (blue text card; Figs. A.7 and A.8)
- Sim-S sample (gray text card; Figs. A.9 and A.10)

In each of the ten figures, there are 24 subplots. These are labeled from (a) to (x) as shown in their top right text box. Each of these subplots shows the p-value distribution of a specific test strategy, which is labeled by the text in the top left box (more details below). The bottom right text box in each subplot gives information about the statistics of the p-value distribution: $>2\sigma$, $>3\sigma$, and $\geq 4\sigma$ refer to the fraction (in %) of p-values that cross the 2σ ($p < 0.0455$), 3σ ($p < 0.0027$), and $\sim 4\sigma$ ($p = 0.0001$; the smallest possible p-value obtainable from our tests) thresholds, respectively. We note that the fraction labeled as $>3\sigma$ is the same as $\mathcal{F}_{3\sigma}$, which is also reported in Table 1 and used in the discussions of Sect. 4. The last two entries in the bottom right text box of each subplot are μ and σ , which are the mean and standard deviation of the p-value distributions, respectively. μ and σ are visualized on each distribution using vertical lines (the solid line is μ and the dotted lines are $\mu \pm \sigma$, as long as mathematically appropriate, i.e., $0 \leq \mu \pm \sigma \leq 1$).

As mentioned above, the top left text box contains information about the specific test strategy to which the distribution pertains. The neutrino sample cut (see Sect. 3.3.2) results are shown as follows:

- No-cut sample: left subplots (a, d, g, j, m, p, s, and v)
- Soft-cut sample: middle subplots (b, e, h, k, n, q, t, and w)
- Hard-cut sample: right subplots (c, f, i, l, o, r, u, and x)

The unweighted and weighted test (see Sect. 3.3.1) results are shown as follows:

- \emptyset_W 3R: rows one and three; subplots (a, b, c, g, h, and i)
- \emptyset_W 1R: rows two and four; subplots (d, e, f, j, k, and l)
- W_G 3R: rows five and seven; subplots (m, n, o, s, t, and u)
- W_T 1R: rows six and eight; subplots (p, q, r, v, w, and x)

Results of the averaging- and counting-based tests (denoted as "Avg" and "Cnt", respectively; see Sect. 3.2) are as follows:

- Avg: rows one, two, five, and six; subplots (a–f and m–r)
- Cnt: rows three, four, seven, and eight; subplots (g–l and s–x)

Regarding the general shape and behavior of the p-value distributions, it is apparent that most of them are well behaved apart from a few oddities that we discuss below. Firstly, from the p-value distributions in the sim-null case (Figs. A.1 and A.2), we see that almost all the p-values are nearly uniformly spread across $0 \leq p \leq 1$ with the mean p-value (μ) being around 0.5 and the standard deviation (σ) being around 0.3. Moreover, the fractions represented by $>2\sigma$ and $>3\sigma$ are around 5% and 0.3%,

respectively. These clearly demonstrate that most of the p-value distributions in the sim-null case are near-Gaussian random distributions — as one would expect from such a random process. However, there are a few nonuniform distributions, namely those seen in subplots (l) and (x) of Fig. A.2. These occur when counting those AI values that surpass a predetermined threshold.

The most noticeable nonuniform feature is the peak of p-values at $p = 1$. This is an artifact of how Eq. 3 is defined: when the observed TS value is exactly zero, $M = N$, which leads to $p = 1$. It turns out that counting AI values that surpass a threshold as well as setting a hard limit on the sample of neutrinos leads to many simulations where the global TS parameter ends up being a small, whole number (e.g., 0, 1, 2, etc.). This is because the hard-cut subsample of neutrinos ignores 75% of the whole IceCat1+ neutrino sample and only consists of 71 spatially well-constrained neutrinos, which drastically reduces the number of blazar–neutrino spatial associations in the 1R test strategies (for both \emptyset_W and W_T). Additionally, when using the counted TS parameters, only those associations with AI above a threshold contribute to the global TS parameter. Unfortunately, Eq. 3 does not perform ideally under such a circumstance when the TS parameters are small, whole numbers, leading to clumping of p-values at specific intervals. However, while these p-value distributions do exhibit nonuniformity and an overall abnormal shape (as compared to the rest of the sim-null distributions), they do not result in smaller than average p-values. In fact, the nonuniformity appears to push these p-value distributions to the higher end (i.e., $p \rightarrow 1$), possibly increasing the number of false-negatives; critically, we emphasize that the number of false-positives remains unaffected. Nevertheless, this is a caveat of setting a hard cut on the neutrino sample, especially when combined with a counting-based strategy that leads to small, whole number TS parameters. Interestingly, in the weighted scenarios, the distribution is more uniform, which can be interpreted as yet another reason why weighting is better than setting neutrino sample cuts. Finally, we point out that such p-value clumping is also somewhat present in the sim-best case (with no weighting employed) when setting a hard-cut on the neutrinos and counting the AI (see subplot l of Fig. A.4) for a similar reason as above, since its TS parameters are also small, whole numbers.

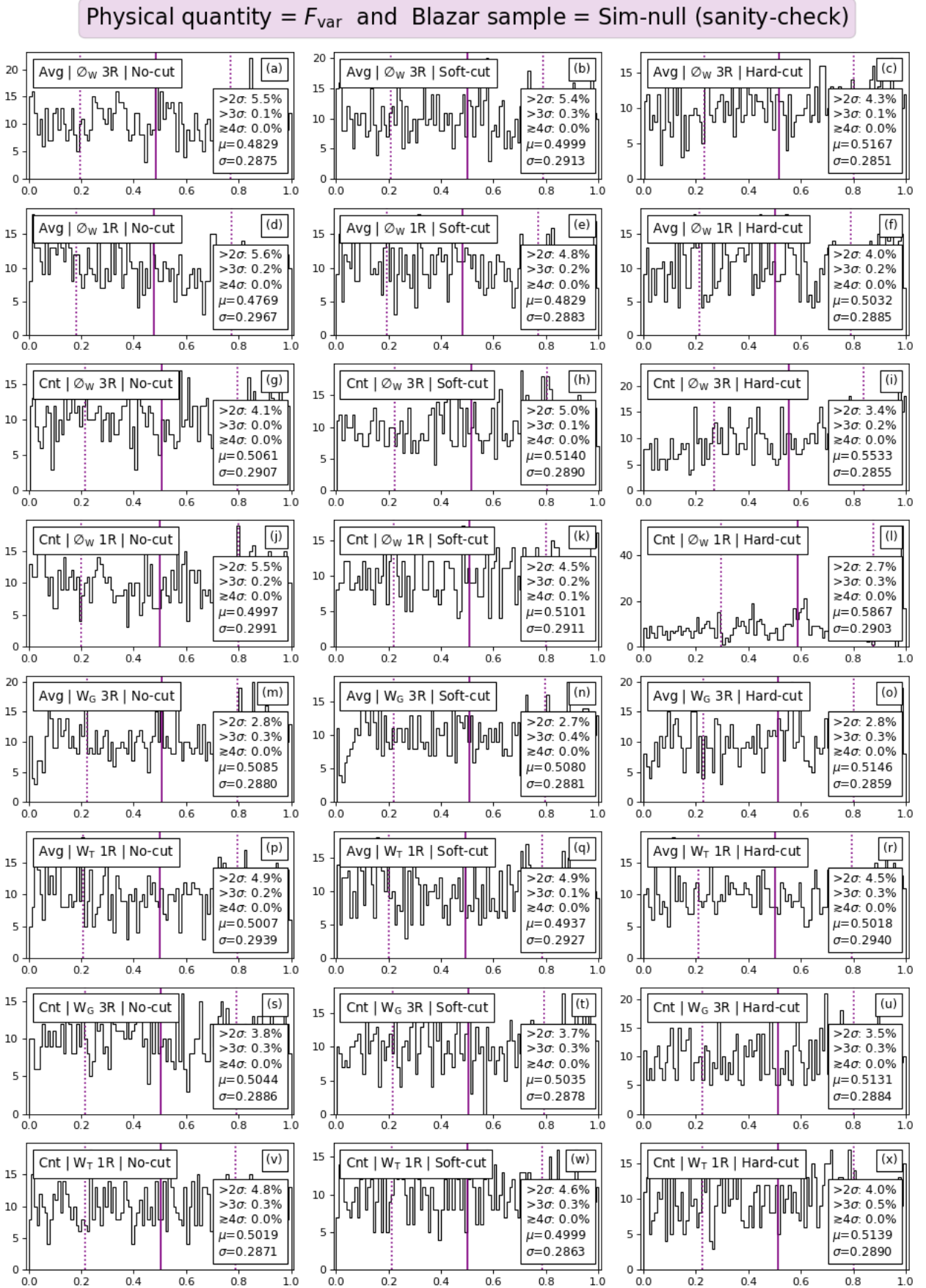


Fig. A.1. P-value distributions when using the time-averaged F_{var} and the sim-null blazars. For more details, see the text in Appendix A.

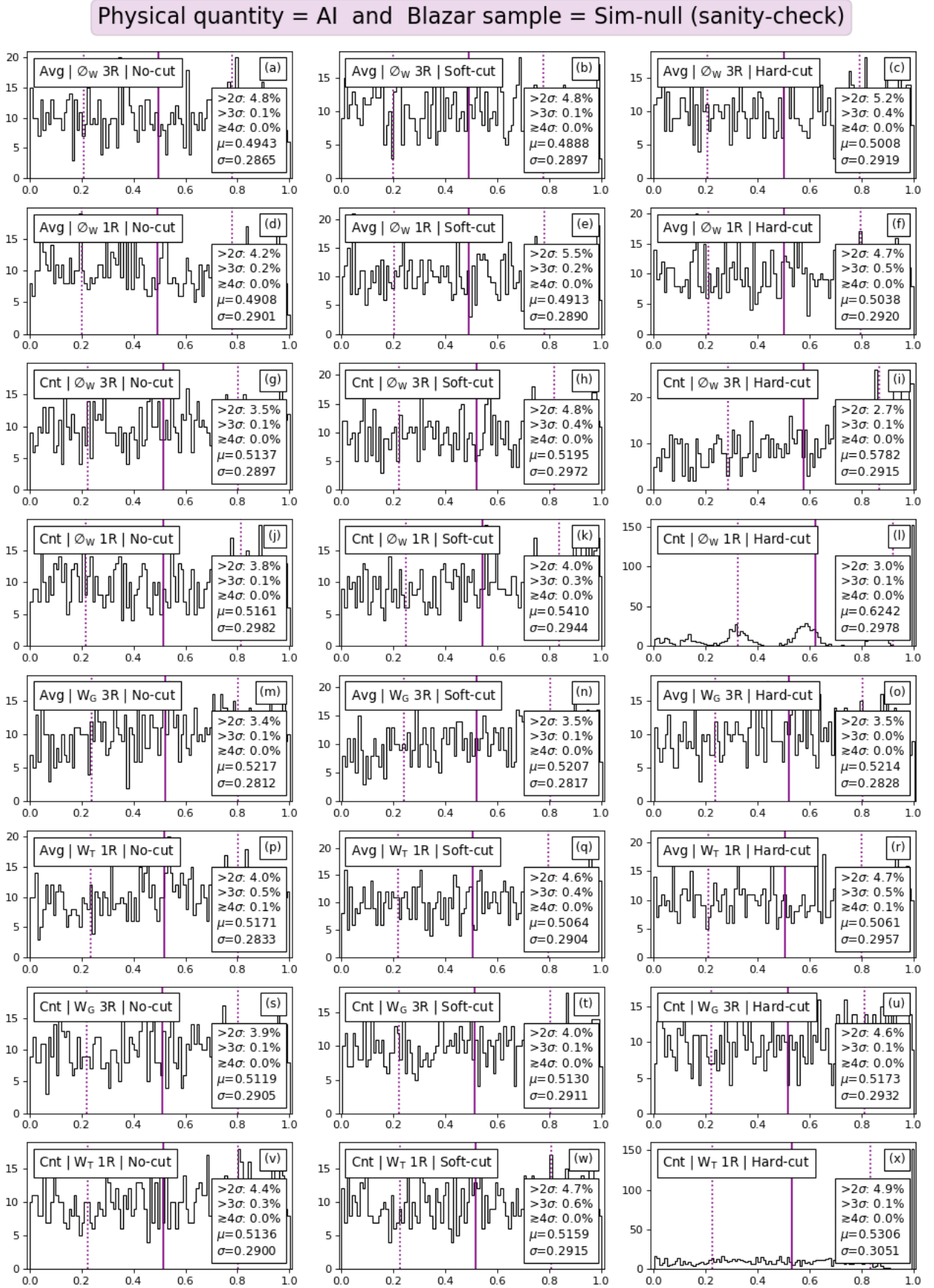


Fig. A.2. P-value distributions when using the time-resolved AI and the sim-null blazars. For more details, see the text in Appendix A.

Physical quantity = F_{var} and Blazar sample = Sim-best (special-case)

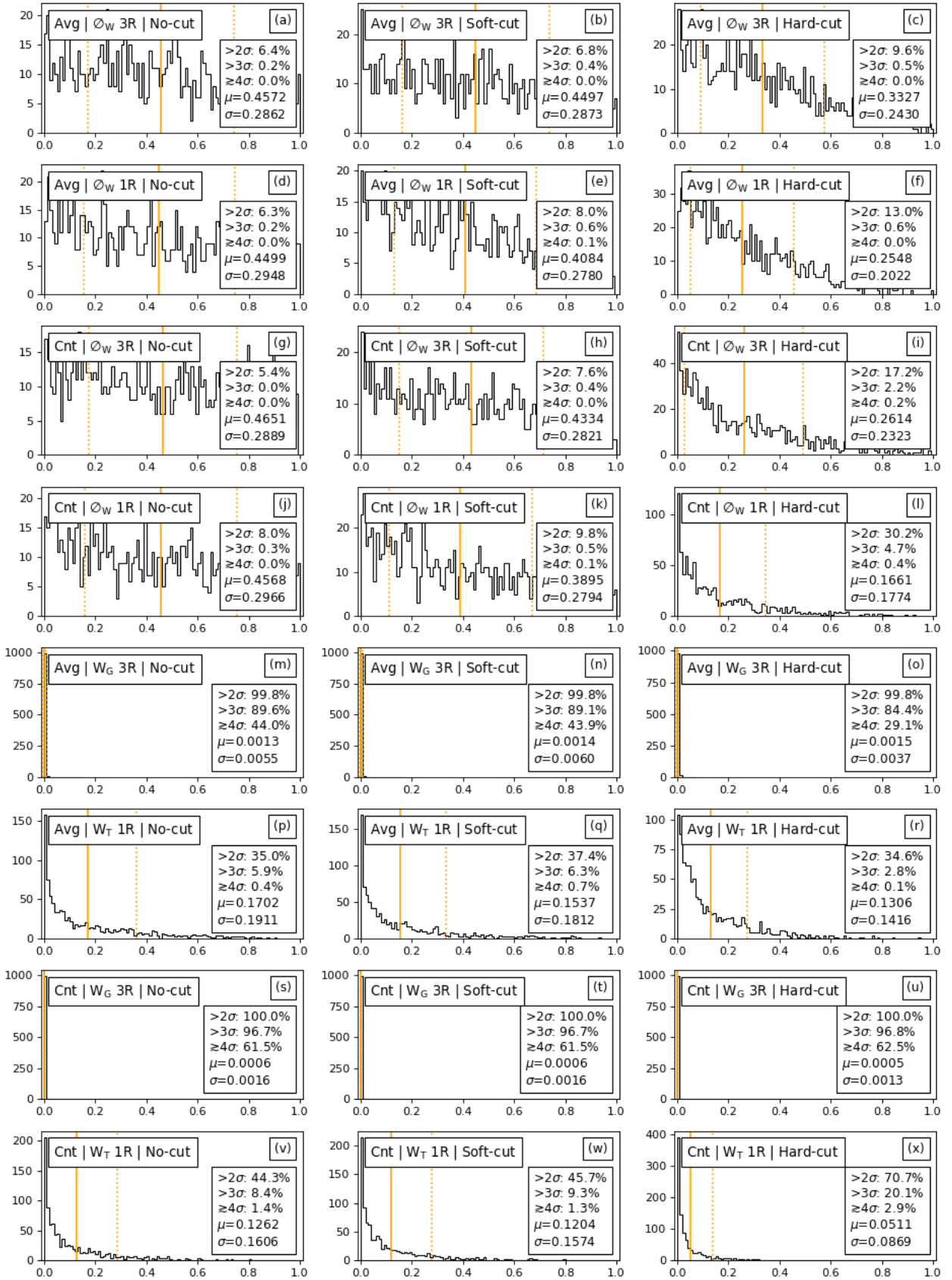


Fig. A.3. P-value distributions when using the time-averaged F_{var} and the sim-best blazars. For more details, see the text in Appendix A.

Physical quantity = AI and Blazar sample = Sim-best (special-case)

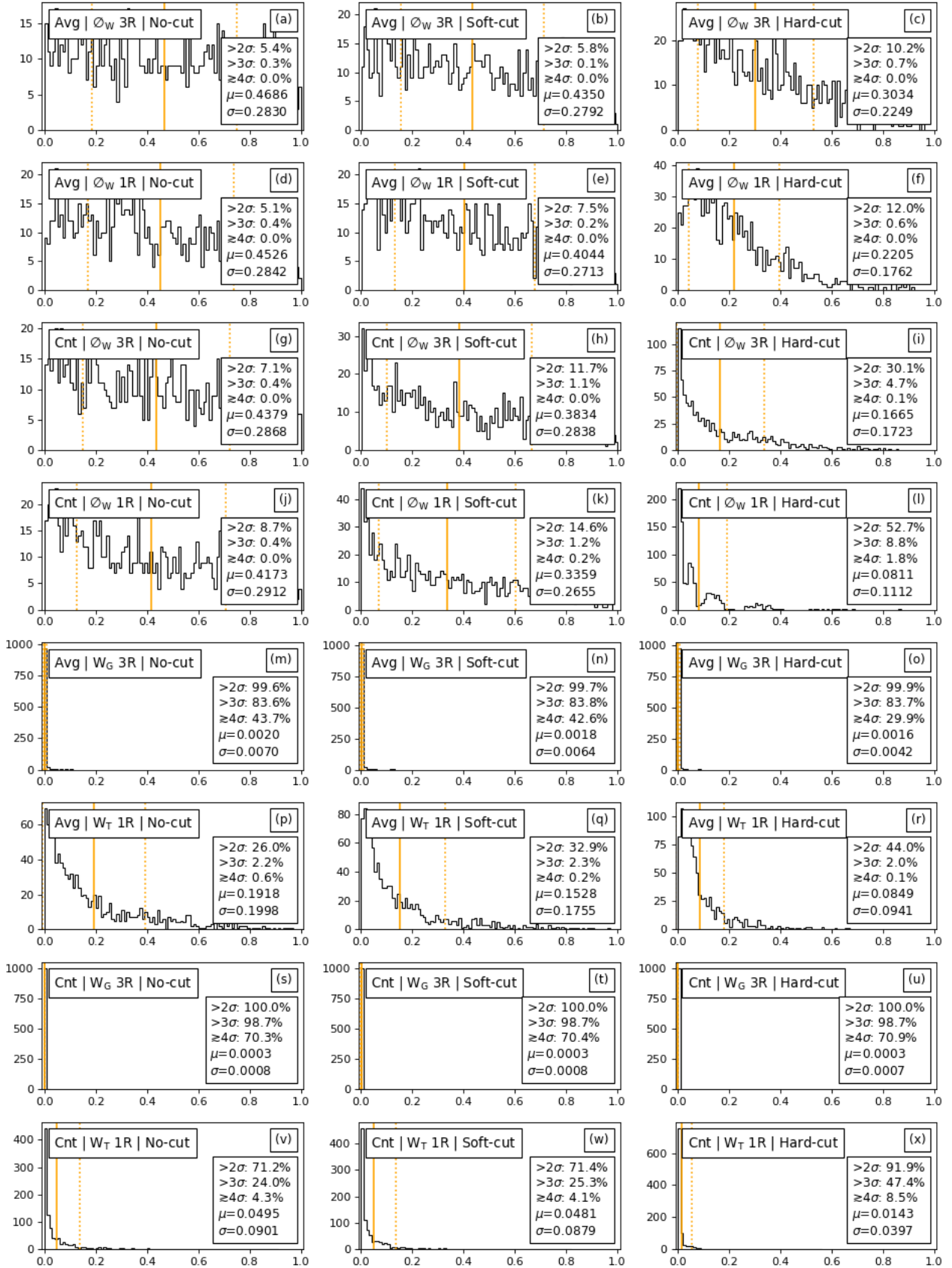


Fig. A.4. P-value distributions when using the time-resolved AI and the sim-best blazars. For more details, see the text in Appendix A.

Physical quantity = F_{var} and Blazar sample = Sim-mid (special-case)

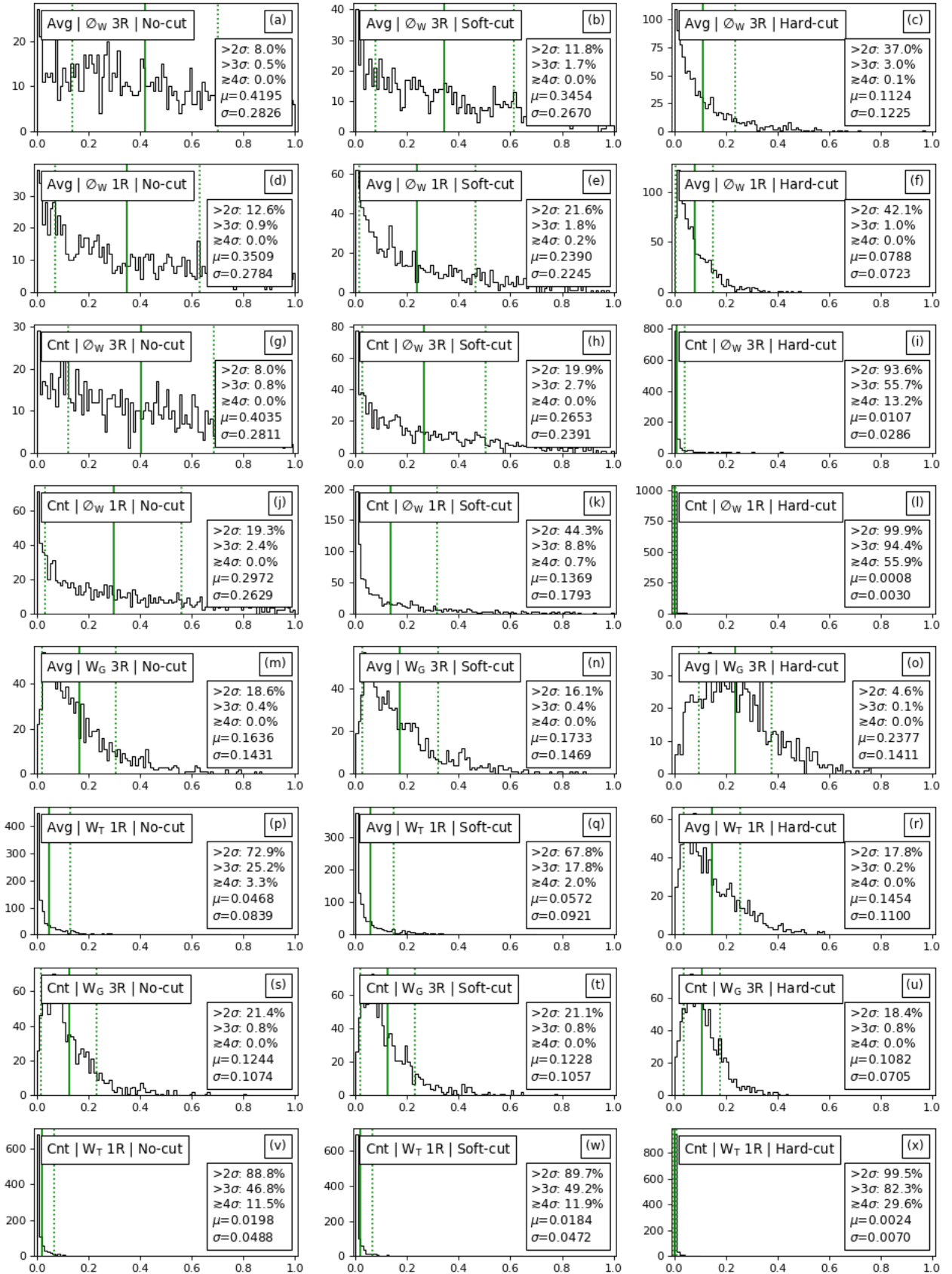


Fig. A.5. P-value distributions when using the time-averaged F_{var} and the sim-mid blazars. For more details, see the text in Appendix A.

Physical quantity = AI and Blazar sample = Sim-mid (special-case)

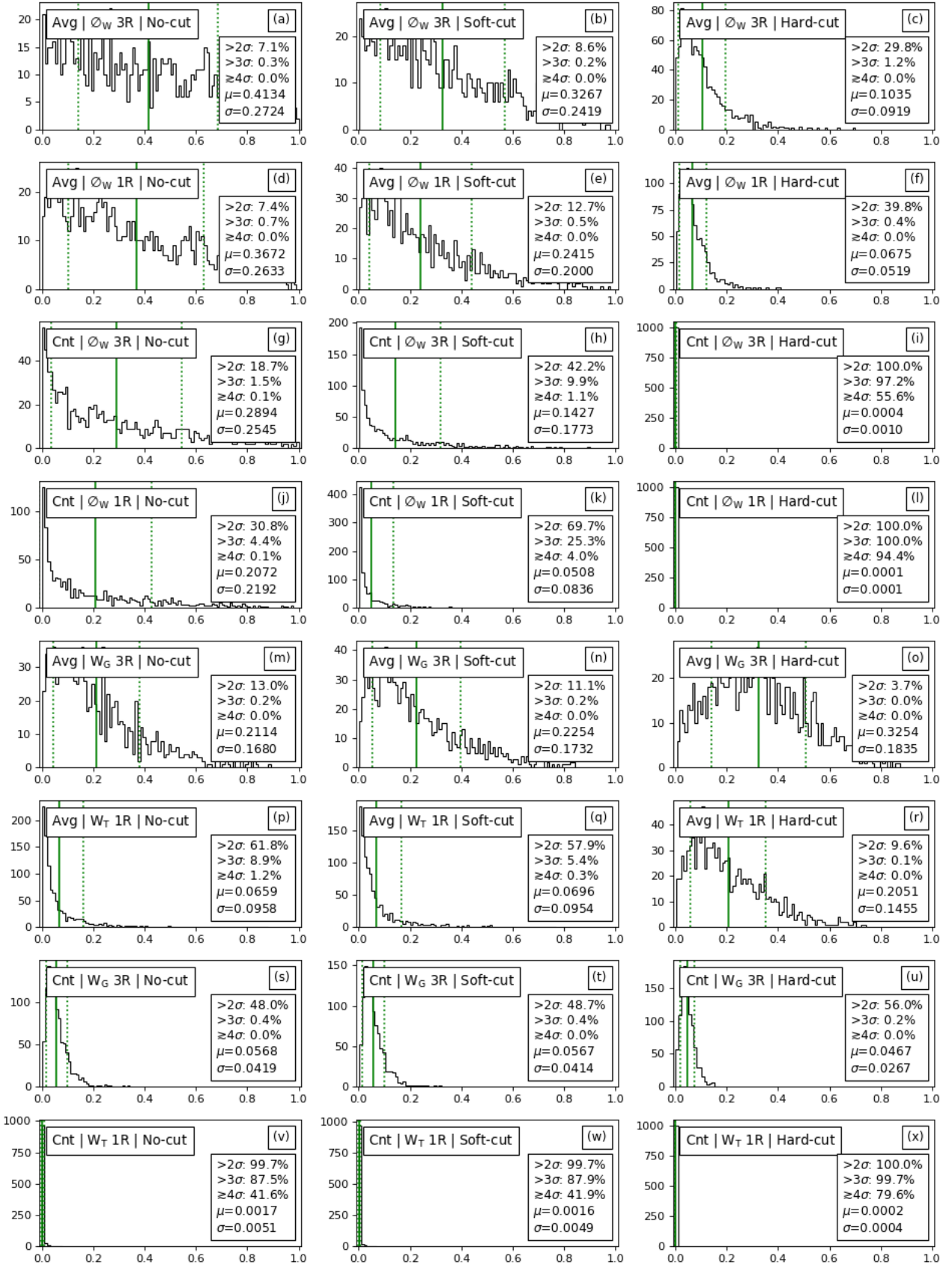
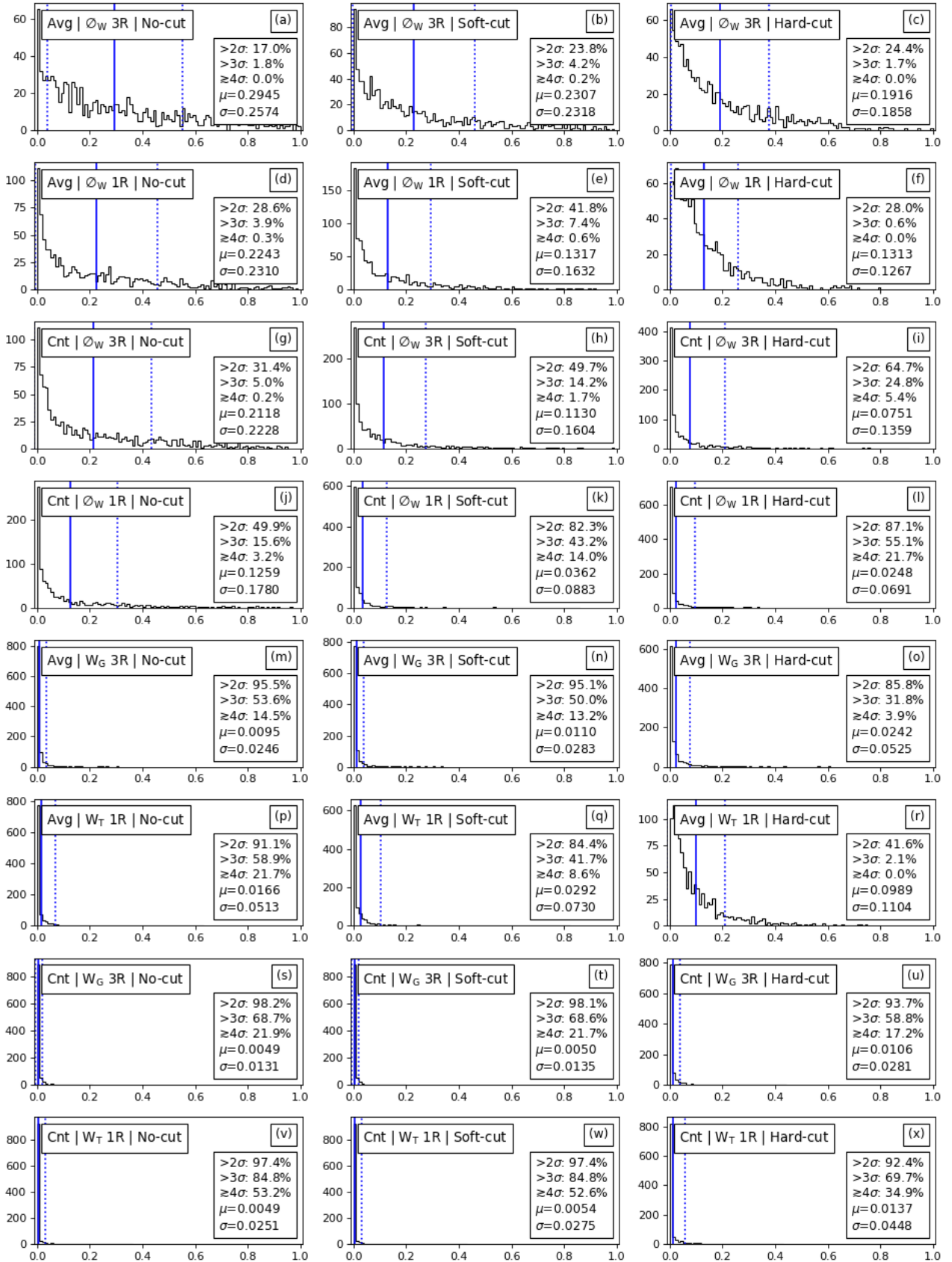


Fig. A.6. P-value distributions when using the time-resolved AI and the sim-mid blazars. For more details, see the text in Appendix A.

Physical quantity = F_{var} and Blazar sample = Sim-0.2S (realistic)

 Fig. A.7. P-value distributions when using the time-averaged F_{var} and the sim-0.2S blazars. For more details, see the text in Appendix A.

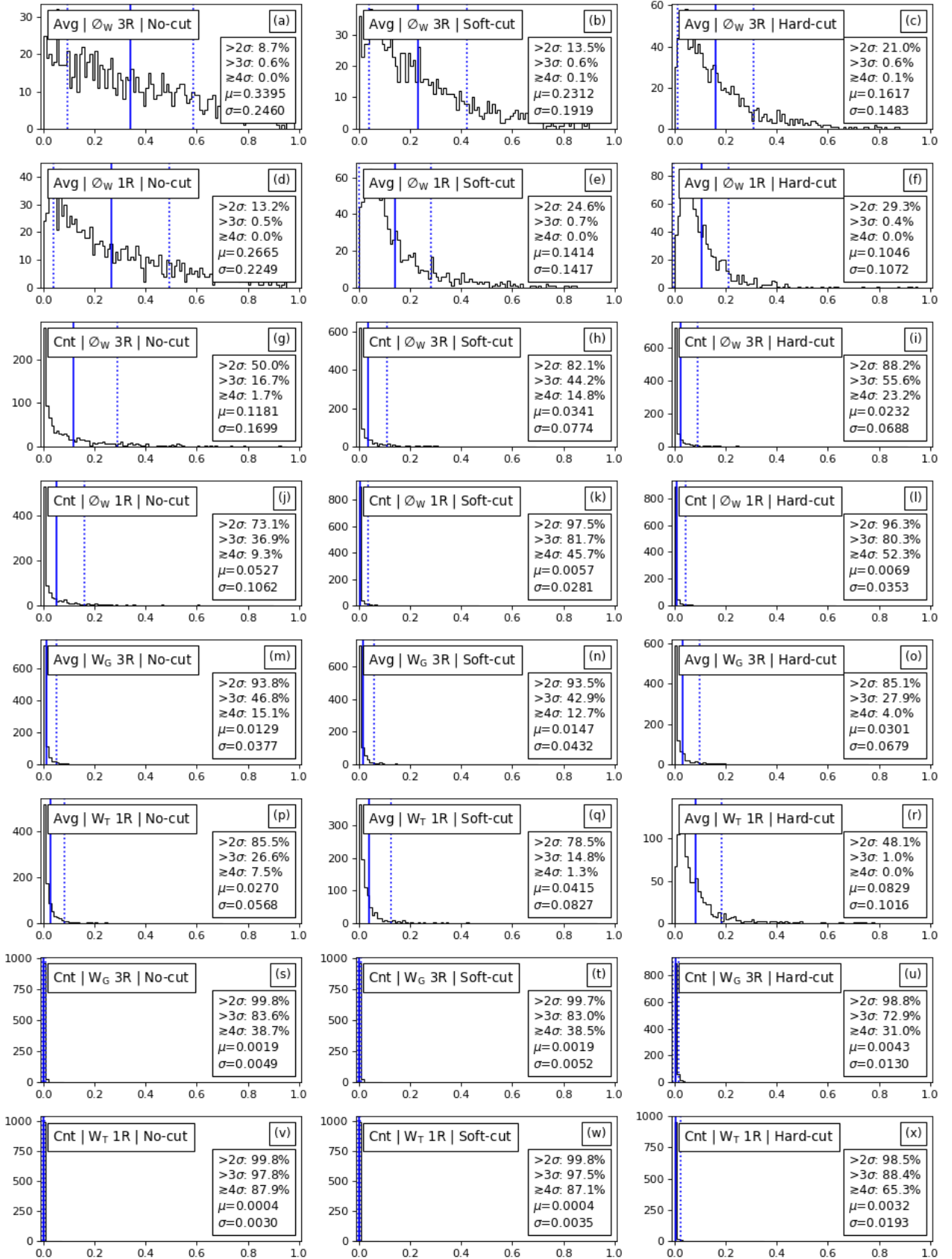
Physical quantity = AI and Blazar sample = Sim-0.2*S* (realistic)

Fig. A.8. P-value distributions when using the time-resolved AI and the sim-0.2*S* blazars. For more details, see the text in Appendix A.

Physical quantity = F_{var} and Blazar sample = Sim-S (sanity-check)

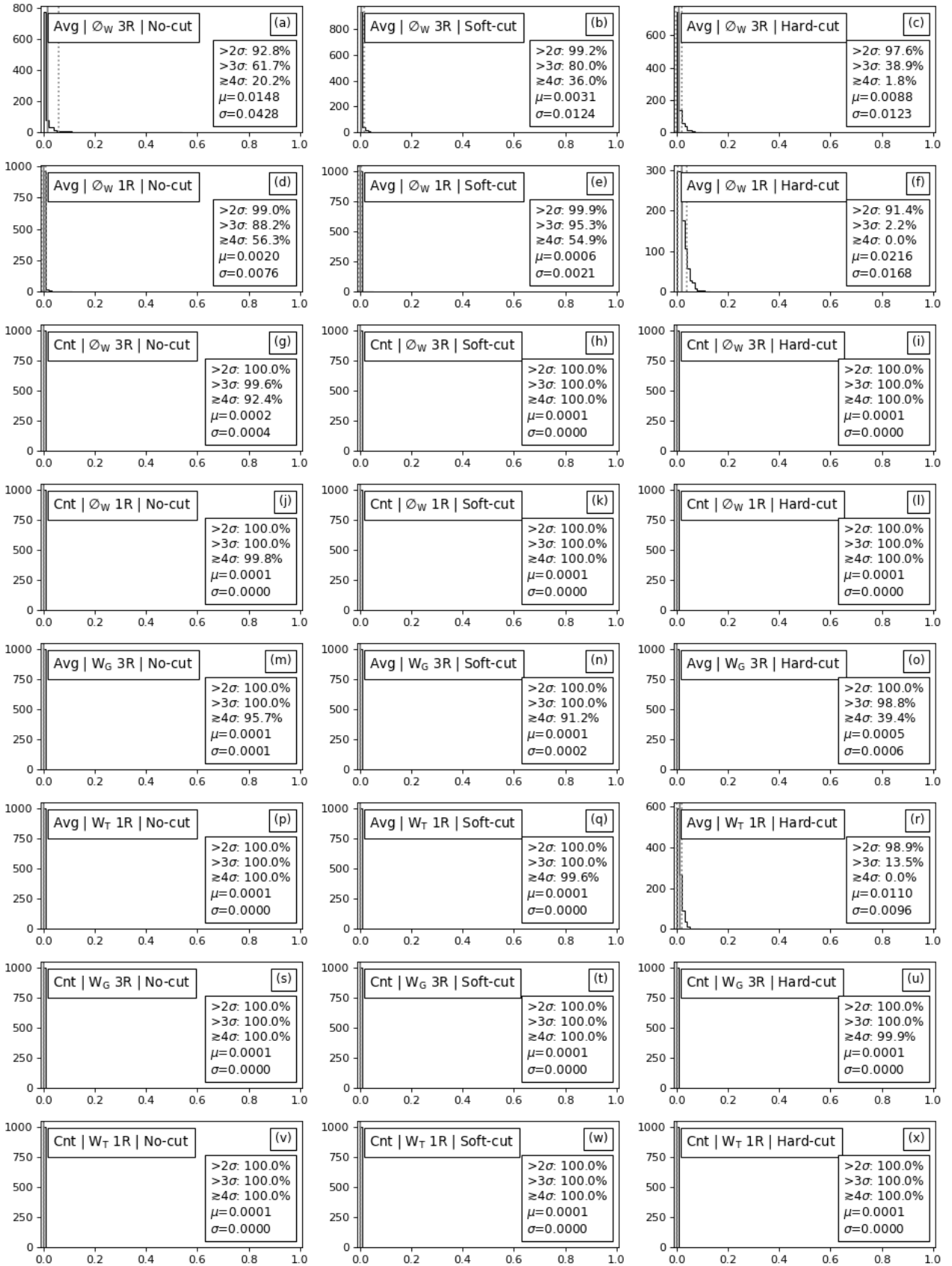


Fig. A.9. P-value distributions when using the time-averaged F_{var} and the sim-S blazars. For more details, see the text in Appendix A.

Physical quantity = AI and Blazar sample = Sim-S (sanity-check)

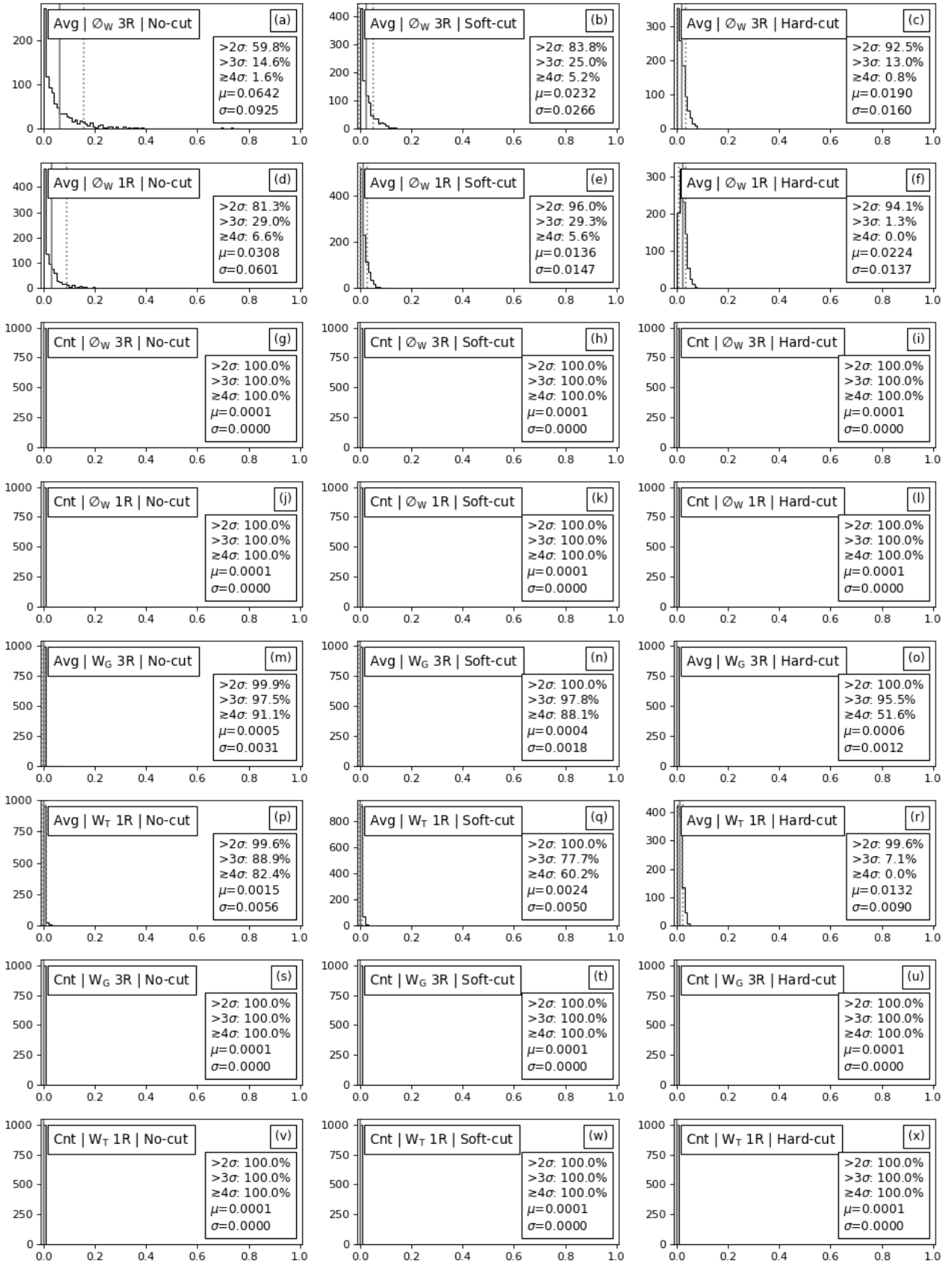


Fig. A.10. P-value distributions when using the time-resolved AI and the sim-S blazars. For more details, see the text in Appendix A.