

Pulsar Timing methods for evaluating dispersion measure time-series

F. Iraci^{1,2*}, A. Chalumeau^{3,4}, C. Tiburzi¹, J. P. W. Verbiest⁵, A. Possenti¹, G. M. Shaifullah^{3,4}, S. C. Susarla⁶, M. A. Krishnakumar^{7,8,9}, M. T. Lam^{10,11,12}, H. T. Cromartie¹³, M. Kerr¹⁴, and Jean-Mathias Grießmeier^{15,16}

¹ INAF - Osservatorio Astronomico di Cagliari, via della Scienza 5, 09047 Selargius (CA), Italy

² Dipartimento di Fisica, Università di Cagliari, Cittadella Universitaria, I-09042 Monserrato (CA), Italy

³ Dipartimento di Fisica “G. Occhialini”, Università degli Studi di Milano-Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy

⁴ INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy

⁵ Florida Space Institute, University of Central Florida, 12354 Research Parkway, Orlando, FL 32826, USA

⁶ Physics, School of Natural Sciences, Ollscoil na Gaillimhe — University of Galway, University Road, Galway, H91 TK33, Ireland

⁷ Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany

⁸ Fakultät für Physik, Universität Bielefeld, Postfach 100131, 33501 Bielefeld, Germany

⁹ National Centre for Radio Astrophysics, Pune University Campus, Pune 411007, India

¹⁰ SETI Institute, 339 N Bernardo Ave Suite 200, Mountain View, CA 94043, USA

¹¹ School of Physics and Astronomy, Rochester Institute of Technology, Rochester, NY 14623, USA

¹² Laboratory for Multiwavelength Astrophysics, Rochester Institute of Technology, Rochester, NY 14623, USA

¹³ National Research Council Postdoctoral Associate, National Academy of Sciences, Washington, DC 20001, USA resident at Naval Research Laboratory, Washington, DC 20375, USA

¹⁴ Space Science Division, Naval Research Laboratory, Washington, DC 20375–5352, USA

¹⁵ LPC2E - Université d’Orléans / CNRS, 45071 Orléans cedex 2, France

¹⁶ Observatoire Radioastronomique de Nançay (ORN), Observatoire de Paris, Université PSL, Univ Orléans, CNRS, 18330 Nançay, France

ABSTRACT

Context. Radio pulsars can be used for many studies, including the investigation of the ionised interstellar medium and the solar wind via their dispersive effects. These phenomena affect the high-precision timing of pulsars and are among the main sources of noise in experiments searching for low-frequency gravitational waves in pulsar data.

Aims. In this paper, we compare the functionality and reliability of three commonly used schemes to measure temporal variations in interstellar propagation effects in pulsar-timing data.

Methods. We carry out extensive simulations at low observing frequencies (100–200 MHz) by injecting long-term correlated noise processes with power-law spectra and white noise, to evaluate the robustness, accuracy and precision of the following three mitigation methods: epoch-wise (**EW**) measurements of interstellar dispersion; the **DMX** method of simultaneous, piece-wise fits to interstellar dispersion; and **DM GP**, which models dispersion variations through Gaussian processes using a Bayesian analysis method. We then evaluate how reliably the input signals are reconstructed and how the various methods react to the presence of achromatic long-period noise.

Results. All the methods perform well, provided the achromatic long-period noise is modeled for **DMX** and **DM GP**. The most precise method is **DM GP**, followed by **DMX** and **EW**, while the most accurate is **EW**, followed by **DMX** and **DM GP**. We also test different scenarios including simulations of L-band ToAs and realistic DM injection, with no significant variation in the obtained results.

Conclusions. Given the nature of our simulations and our scope, we deem that **EW** is the most reliable method to study the Galactic ionized media. Follow-up works should be conducted to confirm this result via more realistic simulations. We note that **DM GP** and **DMX** seem to be the most performing techniques in removing long-term correlated noise, and hence for gravitational wave studies. However, full simulations of pulsar timing array experiments are needed to support this interpretation.

Key words. pulsar: general, interstellar medium, methods: data analysis

1. Introduction

Pulsars are highly magnetized, rapidly rotating neutron stars that emit beams of radiation, mainly in the radio band, from their magnetic poles. Therefore, pulsars are visible as periodic sources to terrestrial observers. The extremely regular rotation of a pulsar allows to predict the time of arrival (ToA) of its radiation at a ra-

dio observatory and to extract from these data information about the pulsar itself, its environment and perturbing effects through the pulsar timing procedure (e.g., Lorimer & Kramer 2005). The model which contains the parameters necessary to describe the signal propagation, usually referred to as timing model, is used to calculate expected ToAs that are then compared with the observed ones to create and analyze the timing residuals (the difference between the observed ToAs and the ones predicted by

* e-mail: francesco.iraci@inaf.it

the timing model). The parameters in the timing model are iteratively fit to minimize the timing residuals and to achieve a more and more precise description of the analyzed pulsars. The shorter the spin period and the brighter the pulsar, the smaller its weighted root-mean-square (RMS) timing residuals and the more precisely determined are the timing-model parameters. The most rapidly rotating pulsars are known as "millisecond pulsars" (Backer et al. 1982) because they have spin periods on the order of milliseconds. The technique of pulsar timing opens up the study not only of the physical characteristics of pulsars, but also other effects, such as the propagation effects through ionized media (Rickett 1990; Foster & Cordes 1990); the most conspicuous of which is dispersion.

Dispersion is a phenomenon caused by the dependency of a medium's refractive index μ on the frequency ν of the propagating radiation and on its own free-electron density n_e . In particular:

$$\mu = \sqrt{\left(1 - \frac{f_p}{\nu}\right)} \quad (1)$$

with f_p , the plasma frequency, defined as:

$$f_p = \sqrt{\frac{e^2 n_e}{\pi m_e}} \quad (2)$$

where e and m_e are respectively the electron electric charge and mass. The net effect of these two dependencies is an absolute delay in the arrival time of the radiation bundle after propagating through an ionized medium (such as the interstellar medium, which hosts an ionized component, or the solar wind), and a relative delay among the radiation at different frequencies within that bundle. In particular, we have that the delay Δt for radiation with a frequency ν is approximately (assuming $f_p \ll \nu$):

$$\Delta t = \mathcal{D} \frac{DM}{\nu^2} \quad (3)$$

where \mathcal{D} is the dispersion constant (Lorimer & Kramer 2005), and the DM parameter is the dispersion measure, defined as:

$$DM = \int_{LoS} n_e dl \quad (4)$$

where the integral is performed along the line of sight (LoS) and the units are pc cm^{-3} . The DM of a pulsar is one of the parameters included in its timing model.

Equation (3) can be rewritten following Verbiest & Shaifullah (2018):

$$\Delta t = 4.15 \times 10^3 DM \left(\frac{2B}{\nu_c^2}\right) \quad (5)$$

where B is the fractional bandwidth $(\nu_2 - \nu_1)/\nu_c$ and ν_c is the central frequency of the observing band.

From these equations it is easy to infer that i) the DM is a direct measure of the free-electron content along a given line of sight and ii) large delays between radiation emitted at different frequencies of the same bundle are shown over low central observing frequencies and/or large fractional bandwidths. Note that, in practice, an absolute value of DM can never be achieved due to the evolution of pulsar profiles with the observing frequency (e.g., Hassall et al. 2012). To complicate matters further, pulsars are high-velocity objects, following the kick that they

receive during the supernova event (Hobbs et al. 2005). This implies that they move rapidly across the sky, and hence the LoS varies significantly as well. In doing so, different parts of the ionized interstellar medium (IISM) are crossed, with different associated values for n_e , and hence the value of DM changes with time (Lam et al. 2016). simple polynomial models of DM time evolution This is why simple polynomial models of DM time evolution are often included in the timing model. Nevertheless, the turbulence in the IISM cannot be correctly described via derivatives only, and hence there are a large number of studies in literature reporting on the time-series of DM variations and their connection with the physics of plasma in the Galaxy.

Recently, Donner et al. (2020) reported the DM variations for 36 millisecond pulsars observed for about 7 years at low-radio frequency (100-200 MHz) with the LOFAR (the LOw Frequency ARray, see van Haarlem et al. 2013) interferometer in which they reached a very low median DM uncertainty, of the order of 10^{-5} pc/cm^3 , for a significant fraction of the sample. Besides, DM variations were detected when the median DM uncertainty was lower than a few in 10^{-4} pc/cm^3 . Krishnakumar et al. (2021) presented the 1-year long DM time series of four millisecond pulsars observed with uGMRT using BAND3 (400-500 MHz) and BAND5 (1360-1460 MHz). They showed that the DM precision improves up to 10^{-4} pc/cm^3 when combining data from both of the bands. Agazie et al. (2023) and Jones et al. (2017) reported on the DM variations of up to 68 millisecond pulsars with mainly the Green Bank (722-1885 MHz) and the Arecibo (302-2400 MHz) radio telescopes, highlighting on their monotonic trends and the presence of annual IISM signatures for almost 20 of them. Lastly, Keith et al. (2024) calculated the DM time series for almost 600 long-period pulsars observed with the MeerKAT telescope at L-band (896-1671 MHz) and a large fractional bandwidth, reporting a broad linear correlation between the DM and its first derivative.

While DM variations allow for the investigation of Galactic plasma, when they are unaccounted for may become a nuisance in other experiments such as Pulsar Timing Arrays (PTAs, Tiburzi 2018; Verbiest et al. 2021; Taylor 2021), that search for low-frequency gravitational waves (GWs, Maggiore 2018) in pulsar data. In particular, GWs are expected to cause space- and time-correlated, long-term perturbations in pulsar timing residuals, usually characterized by a steep, red power spectrum (Phinney 2001). PTAs search for these effects by cross-correlating the timing residuals of pairs of selected pulsars, to identify the predicted signature following the so-called Hellings & Downs curve in the case of an isotropic GW background (Hellings & Downs 1983). One of the most challenging parts of PTA analyses is the characterization of signals that are not GWs but that also induce long-term, time-correlated structures in the timing residuals (see, e.g., Verbiest & Shaifullah 2018), i.e., that are sources of red noise. The red-noise processes with the highest amplitude have two main contributions. The first is often called timing noise, or spin noise, and refers to rotational instabilities of the targeted pulsar (Hobbs et al. 2004). The second is the aforementioned time variability in the amount of interstellar dispersion affecting the pulsar radiation (hereafter referred to as DM noise). PTAs have historically been focused on high frequency bands ($\sim 1 - 3 \text{ GHz}$), small fractional bandwidths and/or asynchronous timing observations at different frequency bands (see Lam et al. 2015, where they demonstrated that multi-frequency asynchronous observations will never reduce the timing error due to DM noise below 10ns). Therefore, while the DM noise has indeed an impact on the GW search, PTA data have a poor sensitivity towards it (see also Figure 6 in Verbiest & Shaifullah

2018). This is why PTAs are starting to use observing campaigns obtained with low-frequency, large-bandwidth facilities such as LOFAR and NenuFAR (the New Extension in Nançay Upgrading LOFAR, see Zarka et al. 2012), see e.g. Donner et al. (2020); Tiburzi et al. (2021); Bondonneau et al. (2021).

In this paper we aim to assess the performances in terms of precision and accuracy of three methods to calculate DM variations, that have been used to study Galactic plasma or to model the DM noise impact in PTA data, based on a comprehensive series of simulations. In Section 2 we describe these three approaches and the simulations. In Section 3 we report the simulation results, which are discussed in Section 4. Finally, in Section 5 we draw our conclusions.

We do not assess the impact of GWs on the DM recoveries methods and also we do not check the impact of these schemes on the timing model parameters. These aspects will be tested in future works.

We stress that there are also other effects that might affect pulsar timing at low radio frequencies. With LOFAR observations the ToAs are generated using a frequency-resolved template (see, e.g., Donner et al. 2020; Tiburzi et al. 2021) and hence interstellar scintillation is not affecting the timing solution. Pulse broadening and scattering variations are two other features that might be present in low frequency observations. In this paper we do not take into account these effects and we simulate a more commonly occurring scenario, without frequency-dependent contributions other than DM variations.

2. Methods / Simulations and analysis

2.1. Simulations

To compare the effectiveness of the various PTA methods to calculate and account for DM variations we evaluate their application on simulated data, which are produced using the `libstempo` software package (Vallisneri 2020).

2.1.1. Simulations with `libstempo`

`libstempo` is a python wrapper of the TEMPO2 software package (Edwards et al. 2006) which allows the use of all of the TEMPO2 functionalities within a python-based environment. In particular, we exploit the `libstempo.toasim.fake_pulsar` library to simulate 10 narrow-band ToAs per observing epoch in the most commonly used LOFAR frequency band for pulsar observations, between 100 and 200 MHz with $5 \mu\text{s}$ fixed ToA template-fitting error bars. The simulated observing epochs have a regular fortnightly cadence, and cover a time span of $T_{span} = 3000$ days. To build simulations as close as possible to the typical PTA dataset, we inject stochastic noise in the form of white noise, achromatic (i.e., radio frequency independent) red noise and DM variations.

White noise (WN) reflects instrumental errors and instrumental sensitivity, as well as intrinsic pulse jitter. We model it by considering two parameters: EFAC which is a multiplicative factor that takes into account ToA measurement errors; EQUAD which is added in quadrature and accounts for any other white noise that may be given by profile variations and possible systematic errors. The final ToA uncertainty is then:

$$\sigma_{\text{ToA}} = \sqrt{(\text{EFAC} \cdot \sigma_{\text{temp}})^2 + \text{EQUAD}^2}, \quad (6)$$

Table 1. List of the input noise parameter values adopted for our simulations.

EFAC = 1.2 EQUAD = $2 \cdot 10^{-6}$ s		
$(\gamma_{RN} = 3.7; \gamma_{DM} = 8/3)$	or	$(\gamma_{RN} = 2.5; \gamma_{DM} = 3.2)$
$A_{RN} = -12.3$	$A_{RN} = -12.3$	$A_{RN} = -12.3$
$A_{DM} = -12.6$	$A_{DM} = -13.6$	$A_{DM} = -14.6$
$A_{RN} = -13.3$	$A_{RN} = -13.3$	$A_{RN} = -13.3$
$A_{DM} = -12.6$	$A_{DM} = -13.6$	$A_{DM} = -14.6$
$A_{RN} = -14.3$	$A_{RN} = -14.3$	$A_{RN} = -14.3$
$A_{DM} = -12.6$	$A_{DM} = -13.6$	$A_{DM} = -14.6$

Notes. The WN parameters EFAC and EQUAD are fixed in all the realisations respectively to 1.2 and $2 \mu\text{s}$ and $\sigma_{\text{temp}} = 5 \mu\text{s}$. We use two different pairs of spectral indices: the first and more realistic with a steeper RN ($\gamma_{RN} = 3.7, \gamma_{DM} = 8/3$); the second and opposite case is produced with a flatter RN spectrum ($\gamma_{RN} = 2.5, \gamma_{DM} = 3.2$). For each pair we use 9 different combinations of (A_{RN}, A_{DM}) as reported in each cell.

where σ_{temp} is the template-fitting error and it is given as an input to our simulations¹.

Achromatic red noise (RN) and DM variations are both time correlated noise processes which are modeled with a Fourier basis of N_f coefficients and a power-law power spectrum (EPTA Collaboration et al. 2023a). We distinguish between chromatic and achromatic noise processes using a chromatic index α respectively equal to 2 and 0. The time delay induced on a ToA with frequency ν at an epoch t is then:

$$t_{\text{delay}}(t) = \sum_{i=1}^{N_f} \sqrt{\mathcal{P}_i} F_i \left(\frac{\nu}{\nu_{\text{ref}}} \right)^{-\alpha}. \quad (7)$$

\mathcal{P}_i is the powerlaw power spectral density with hyperparameters \hat{A} , the normalized amplitude at the frequency of 1 yr^{-1} , and γ , the spectral index. It is written as follow:

$$\mathcal{P}_i = \frac{\hat{A}_n^2}{12\pi^2 T_{span}} \text{yr}^3 \left(\frac{f_i}{1 \text{ yr}^{-1}} \right)^{-\gamma_n} \quad (8)$$

with $n = \text{RN or DM}$ and we will refer to the amplitude as $A_n = \log_{10}(\hat{A}_n)$. F is the matrix of the cosine functions at each Fourier frequency $f_i = i/N_f$ and time t :

$$F_i(t) = \sqrt{2} \cos(2\pi f_i t + \phi), \quad (9)$$

with ϕ being a random phase drawn from $\mathcal{U}(0, 2\pi)$. The radio frequency is ν and the reference frequency ν_{ref} is set to 1.4 GHz. The time series of the injected noise signal will then be a sum over a finite number N_f of cosine functions (see second and third row of Figure 1 for an example). For both RN and DM noise we consider $N_f = 30$ components. This corresponds to a frequency of about 1.15×10^{-7} Hz which is close to the Nyquist one that, given our 14 days of cadence, is at around 2×10^{-7} Hz.

2.1.2. Simulation parameters

Table 1 shows all the parameters adopted for our simulations, with σ_{temp} , EFAC and EQUAD being fixed respectively to $5 \mu\text{s}$, 1.2 and $2 \mu\text{s}$ for all the cases. Concerning the noise spectral indices, we first assume a realistic spectral index for DM variations

¹ We use the same convention of EPTA Collaboration et al. (2023a) for σ_{ToA} , however some PTA collaborations use the TEMPO2 definition (see Verbiest et al. 2016).

from the Kolmogorov turbulence theory (Armstrong et al. 1995) ($\gamma_{DM} = 8/3$) and a steeper one for the RN ($\gamma_{RN} = 3.7$). Secondly, we consider a different pair of spectral indices: $\gamma_{DM} = 3.2$ and $\gamma_{RN} = 2.5$. For each pair of spectral indices we have studied three fixed values of amplitude of the power spectrum for both of the noise processes, producing a total of nine datasets per pair of spectral indices. A_{RN} and A_{DM} have been chosen within the range of values reported by EPTA Collaboration et al. (2023a).

For each specific set of A_{RN} and A_{DM} and pair of spectral indices, we generate a total of 100 timing residuals realizations. In each realization, the seed for the injected RN is kept constant to maintain the same RN signature across all timing residuals, while the DM noise seed varies with each iteration, resulting in distinct DM time series for all 100 realizations.

Figure 1 clarifies our simulation process by reporting three different scenarios, each of them with a different A_{DM} that increases going from left to right. The first row reports the overall timing residuals; the second row shows the delay due to the injected RN; the third row illustrates the DM variations introduced in our dataset. The last row represents the power spectrum of the noise processes involved in the three scenarios. Note that the first and third rows of each column reports only one of the 100 realizations that have been created for that set of parameters.

2.2. DM recovery methods

Each generated series of timing residuals is analysed with the following methods.

2.2.1. Epoch-wise DM

Epoch-wise DM (**EW**, e.g. Donner et al. 2019; Tiburzi et al. 2019) models a DM value and a constant offset C via TEMPO2 over the parsed ToAs for each simulated observing epoch. In particular the fit functional form is:

$$\Delta t = \frac{DM}{\nu^2} + C \quad (10)$$

where the parameters are the dispersion measure DM, and C , a constant offset that is supposed to absorb any unaccounted achromatic process. The observing frequency ν corresponds in our simulations to an array of 10 values in the LOFAR frequency range 110-190 MHz and Δt is the time delay contained in the measured ToAs. This **EW** analysis returns a pair of DM and its corresponding 1σ uncertainty given by the fit for each observation. The advantage of **EW** is that it simultaneously accounts for the DM and RN effects on the set of parsed ToAs, that corresponds to the ToA set for a single observing epoch. Hence, we do not need to separately model the RN.

2.2.2. DMX

DMX is a method developed by the NANOGrav collaboration (e.g., Agazie et al. 2023, sec 4.1) which performs a piece-wise constant fit of a DM value across an a-priori specified temporal window. In our simulations, each window is such that it contains the 10 ToAs of a single observing epoch so that, at the end of the procedure, we obtain as many **DMX** parameters as the number of observing epochs. In this way **DMX** is extremely similar to Epoch-wise, with two important differences: it uses the PINT software (Luo et al. 2021) suite to perform the fit, and it does not model the constant offset C . This means that fitting for the **DMX** parameters does not account for any RN process in the

data and, in order to do so, the procedure to calculate an unbiased DM time series is divided in three steps: a first fit for the **DMX** parameters; a RN modeling round (while correcting for the initial **DMX** measures) with the enterprise software suite (Ellis et al. 2020); then a second **DMX** fit. In the end, also **DMX** will provide a DM value, with the corresponding uncertainty, for each observation.

2.2.3. DM GP

The EPTA and PPTA collaborations use a fully Bayesian-based noise-analysis approach carried out with the enterprise software suite, able to describe the WN with the EFAC and EQUAD parameters (described in Section 2.1.1), and the achromatic red noise (RN) and DM variations, which are modeled as stationary Gaussian Processes (GPs, see details in appendix A and van Haasteren & Vallisneri 2014), while the Bayesian inference is performed via the Markov-Chain Monte Carlo sampler PTMCMCSampler (Ellis & Van Haasteren 2017). In this approach, the time delay induced on a ToA with radio frequency ν at an epoch t is written as:

$$\delta t(t) = \sum_{i=1}^{N_f} [a_i \sin(2\pi t f_i) + b_i \cos(2\pi t f_i)] \left(\frac{\nu}{\nu_{ref}} \right)^{-\alpha}, \quad (11)$$

where ν_{ref} , α and f_i are the same as in Equation (7). The number of frequency components we use is the same of the injection: $N_f = 30$ for both RN and DM. The weights a_i and b_i follow a multivariate Gaussian distribution with zero mean and covariance matrix Σ_{jk} that in the frequency domain is given by:

$$\Sigma_{jk} = P_L(f_j) \frac{\delta_{jk}}{T_{span}}, \quad (12)$$

with $j, k = 1 \dots N_f$, the Kronecker delta δ_{jk} and the power spectrum P_L :

$$P_L(f) = \frac{A^2 \text{yr}^3}{12\pi^2} \left(\frac{f}{1\text{yr}^{-1}} \right)^{-\gamma}. \quad (13)$$

By marginalizing over the a_i and b_i parameters, this analysis yields the posterior distributions for EFAC and EQUAD, and for the hyperparameters A and γ of RN and DM power spectra. The posteriors describing the DM power-spectrum allow to reconstruct the DM noise via the LaForge software suite (Hazboun 2020), as a sum over a set of N_f Fourier components as described in appendix A.

In particular, we give as input to LaForge a large number N of random posterior samples of the noise parameters, and we obtain as output N different time-domain realizations of the time delays corresponding to the same noise process. These delays are then converted into DM values by inverting Equation (3), hence yielding a set of N DM time series (i.e., a DM probability distribution for each epoch). In order to compare the **DM GP** output with the results of the other two methods we compute the mean DM and standard deviation σ_{DM} at each epoch, hence obtaining one reference value per observation.

3. Results

We present here the results of the analysis with each method. A total of 3 main methods have been considered (**EW**, **DMX**, **DM GP**), and two additional ones in which the RN is not modeled

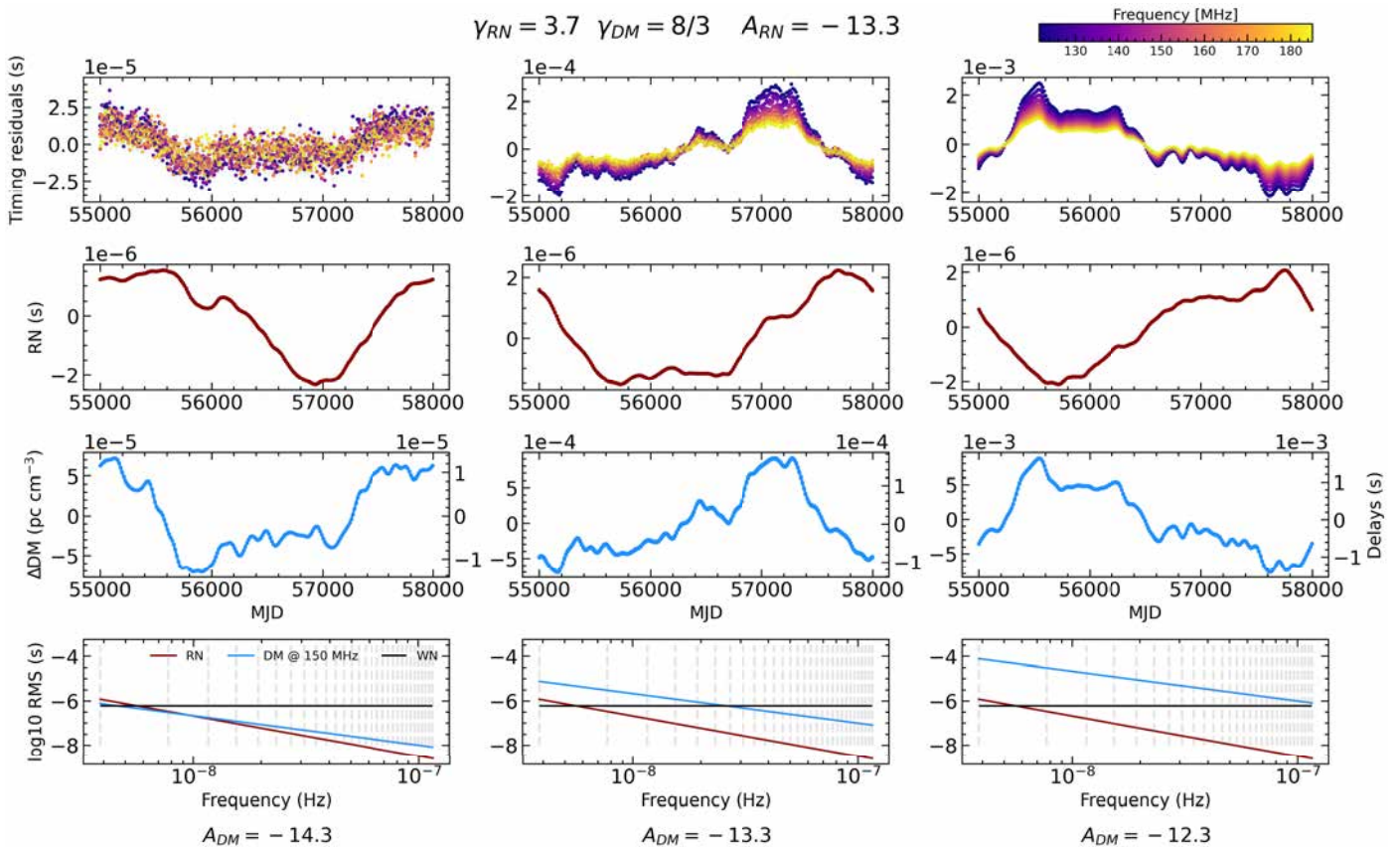


Fig. 1. Simulations carried out with injected values of γ_{RN} , A_{RN} and γ_{DM} of, respectively, 3.7, -13.3 and $8/3$. The columns are associated with increasing values of A_{DM} , left to right (as indicated at the bottom). First row: One out of the 100 simulated timing residuals with a colormap associated to the observing frequency (as reported in the top right corner). Second row: Injected RN signal. Third row: One out of the 100 different injections of the DM time series. On the right y-axis we report the corresponding time delay at the reference frequency of 150 MHz. Fourth row: Power spectra of the injected noise processes: WN (black), RN (red), DM variations at LOFAR central frequency of the band: ~ 150 MHz (light blue). Grey dashed vertical lines are the 30 Fourier frequency bins.

(**DMX no RN, DM GP no RN**)²: In Figure 2 are shown the DM time series recovered with the three main methods on the same noise realization accompanied by the corresponding DM residuals obtained by subtracting the recovered and the injected DMs. An example of the DM residuals that can be obtained by combining together all the 100 realizations generated per each simulation is shown in Figure 3, with the bottom panel representing the RN signal injected in all realizations. In Tables 2 and 3 we report the mean error on the DM estimates for each method and for each set of parameters. Figures 4 and 5 report DM residual histograms summarizing the results obtained for all the simulations per pair of spectral indices. In particular, the values reported in the histograms are obtained by normalizing the DM residuals by their errorbars, in order to have insights about their Gaussianity by fitting them via a Gaussian function. The χ^2 of these fits are given in Tables 4 and 5.

We present also our results on the achromatic noise recovery in Appendix B.

4. Discussion

4.1. Absorption of red noise in the DM reconstruction

The presence of both RN and DM noise in our initial dataset allows us to compare the ability of each method to correctly

² RN modeling is an inherent part of **EW**, so no separate analysis is possible.

identify the chromatic noise processes when also achromatic signals are present. The DM residuals reported in the top panel of Figure 3 show that each method that models RN does not display any evident structure. On the other hand, **DM GP no RN** and **DMX no RN** show evident corruptions in the DM recovery, whose structure matches the injected RN signal (bottom panel). The same conclusion can be drawn by looking at the first column of Figures 4 and 5, in which the distributions of the normalized DM residuals calculated for those two methods are not symmetric and not centered around zero. This indicates that modeling by **DM GP** and **DMX** is not inherently restricted to modeling frequency-dependent signals, but is susceptible to absorbing frequency-independent signals as well, especially if those are not being independently modeled in the analysis. This is true as long as the RN signal is sufficiently ‘loud’ to rise above the WN level. When this condition is not satisfied, or when the power spectrum of the RN exceeds the WN power in by only a few frequency bins, then the DM recovery is not as much affected by the absence of RN modeling. This can be easily seen in the second and third column of Figures 4 and 5, where the distributions are very similar to each other because of the low amplitude of the RN (-13.3 and -14.3 respectively).

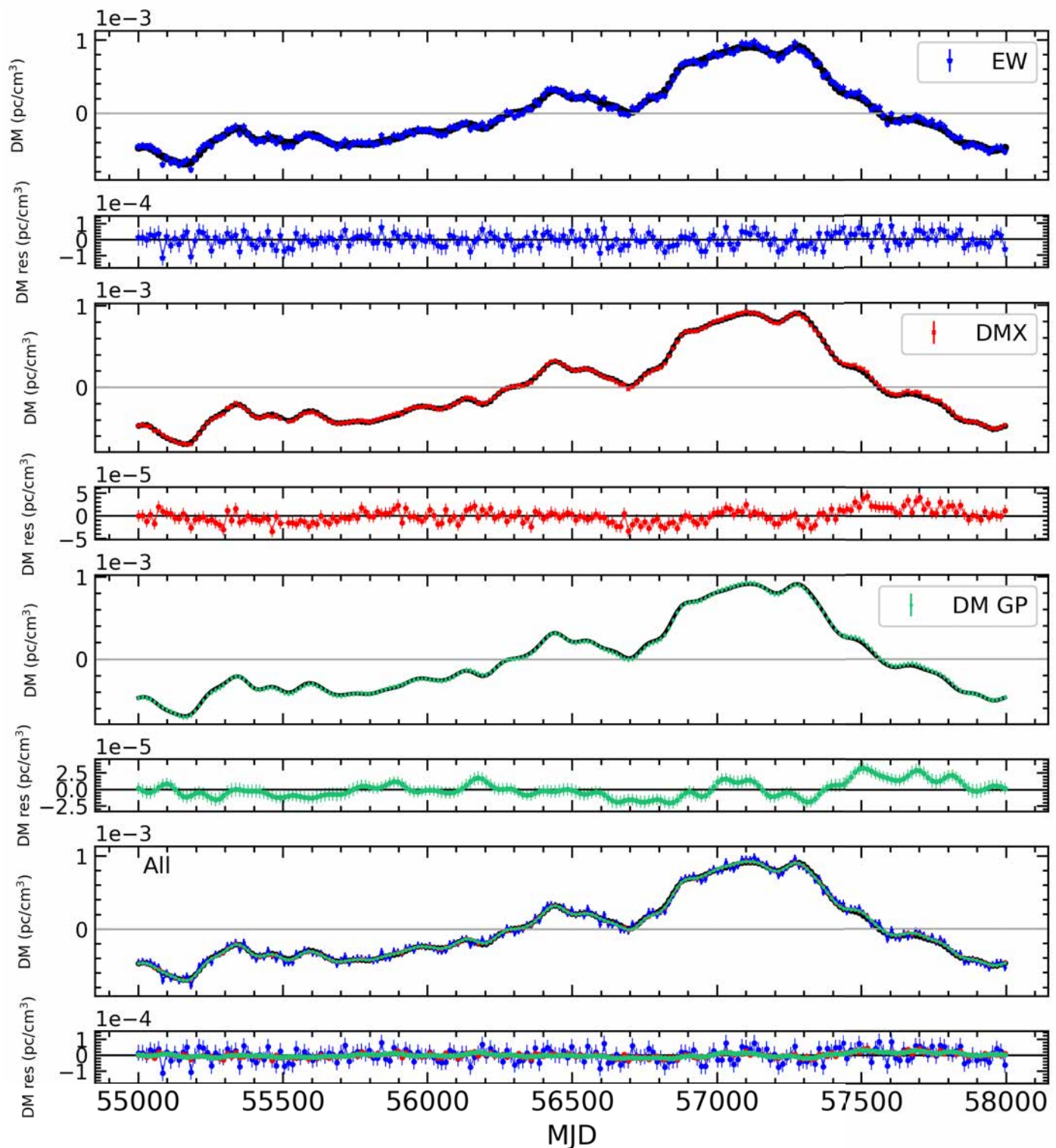


Fig. 2. DM time series and DM residuals as recovered by **EW** (blue, 1st and 2nd panel), **DMX** (red, 3rd and 4th panel) and **DM GP** (green, 5th and 6th panel). The black line is the injected DM time series. In the last two panels we overplot, respectively, the DM time series and the DM residuals for the three methods. The DM and RN injected parameters for this realization are: $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$, $A_{RN} = -12.6$, $A_{DM} = -13.3$

4.2. Comparison of the methods based on precision and accuracy

We base the comparison among the methods on both the yielded precision and accuracy. We define precision as the uncertainty on the measured DM. The method which gives measurements with the lowest uncertainties and hence is the most precise is

DM GP, followed by **DMX** and then **EW**. This can be inferred from Tables 2 and 3, where we report the mean uncertainty on the DM recoveries per each method and set of parameters. The mean uncertainties of **DM GP** are around one order of magnitude smaller than the **EW** ones, while **DMX** sits in the middle of the two. When the noise model lacks the RN component (**DMX no RN**, **DM GP no RN**) and hence we fit the residuals only with

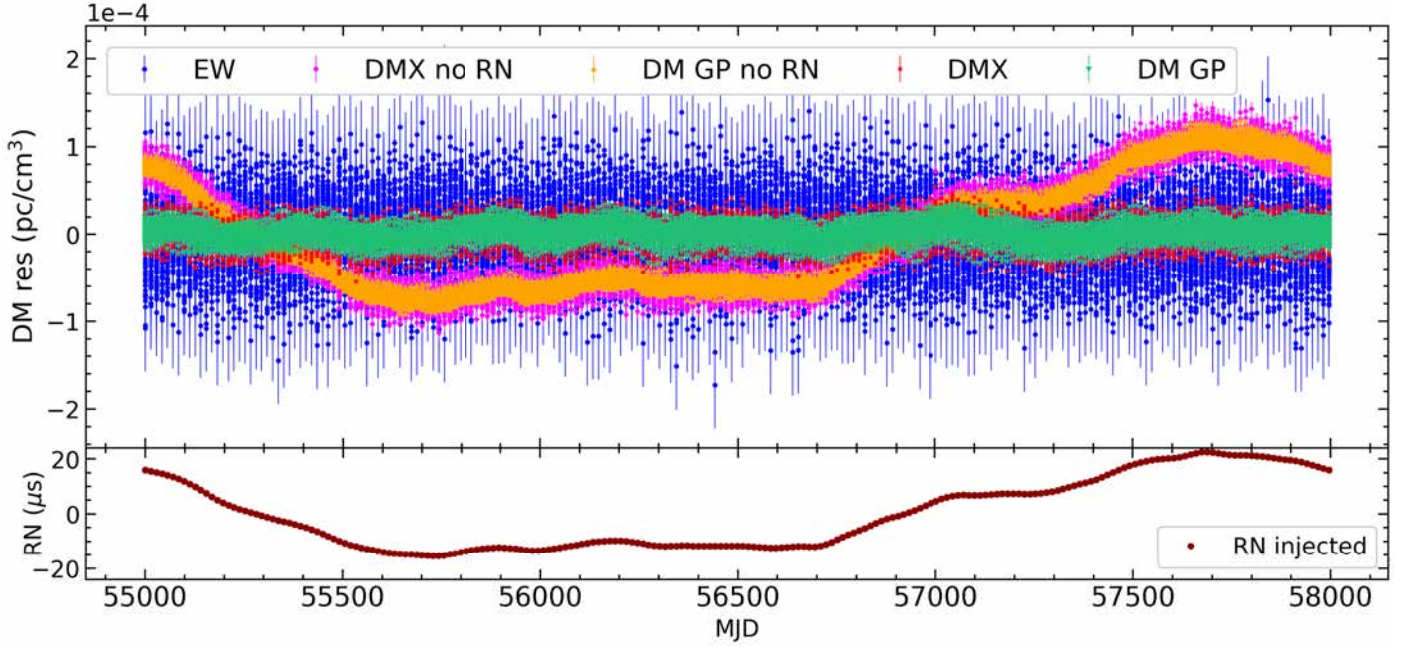


Fig. 3. Top panel: DM residuals of the 100 realization for the five considered methods: **EW**(blue), **DMX no RN** (magenta), **DM GP no RN** (orange), **DMX** (red), **DM GP** (green) for set of input parameters: $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$, $A_{RN} = -12.6$, $A_{DM} = -13.3$. Bottom panel: RN signal injected (the same for all of the 100 realizations. See Section 2)

Table 2. Mean uncertainty on the DM measurements for each method and pair of (A_{RN}, A_{DM}) values in the case of $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$.

A_{RN}	$\langle\sigma_{DM}\rangle \cdot 10^{-5}$ (pc cm $^{-3}$)					A_{DM}
	EW	DMX	DM GP	DMX no RN	DM GP no RN	
-12.6	5.0	1.3	1.0	0.8	0.6	-12.3
-13.6	5.0	1.1	0.7	0.8	0.5	-12.3
-14.6	5.0	1.0	0.6	0.8	0.5	-12.3
-12.6	5.0	1.3	1.0	0.8	0.6	-13.3
-13.6	5.0	1.1	0.6	0.8	0.5	-13.3
-14.6	5.0	1.0	0.7	0.8	0.5	-13.3
-12.6	5.0	1.3	0.8	0.8	0.4	-14.3
-13.6	5.0	1.1	0.5	0.8	0.4	-14.3
-14.6	5.0	1.1	0.4	0.8	0.4	-14.3

Table 3. Mean uncertainty on the DM measurements for each method and pair of (A_{RN}, A_{DM}) in the case of $\gamma_{RN} = 2.5$, $\gamma_{DM} = 3.2$.

A_{RN}	$\langle\sigma_{DM}\rangle \cdot 10^{-5}$ (pc cm $^{-3}$)					A_{DM}
	EW	DMX	DM GP	DMX no RN	DM GP no RN	
-12.6	5.0	1.3	0.9	0.8	0.6	-12.3
-13.6	5.0	1.1	0.6	0.8	0.5	-12.3
-14.6	5.0	1.0	0.6	0.8	0.5	-12.3
-12.6	5.0	1.3	0.9	0.8	0.5	-13.3
-13.6	5.0	1.0	0.6	0.8	0.5	-13.3
-14.6	5.0	1.0	0.6	0.8	0.5	-13.3
-12.6	5.0	1.3	0.7	0.8	0.4	-14.3
-13.6	5.0	1.0	0.4	0.8	0.4	-14.3
-14.6	5.0	1.0	0.4	0.8	0.4	-14.3

the DM and WN, the mean DM uncertainty is systematically lower than for the full noise model which includes WN, RN, DM. This is expected because without fitting for the RN we are reducing the number of parameters in the model. The behavior of the DM residuals when we consider all the 100 simulated realization together is shown in Figure 3 for all of the five methods.

The least precise method, **EW**, is also associated to the largest spread in the residuals, while **DM GP** (with RN modeled) produces a much smaller deviation of the residuals around zero.

While precision informs about the uncertainty level, it does not report how far the data points are from the true value. We define accuracy as the level of agreement between a test result

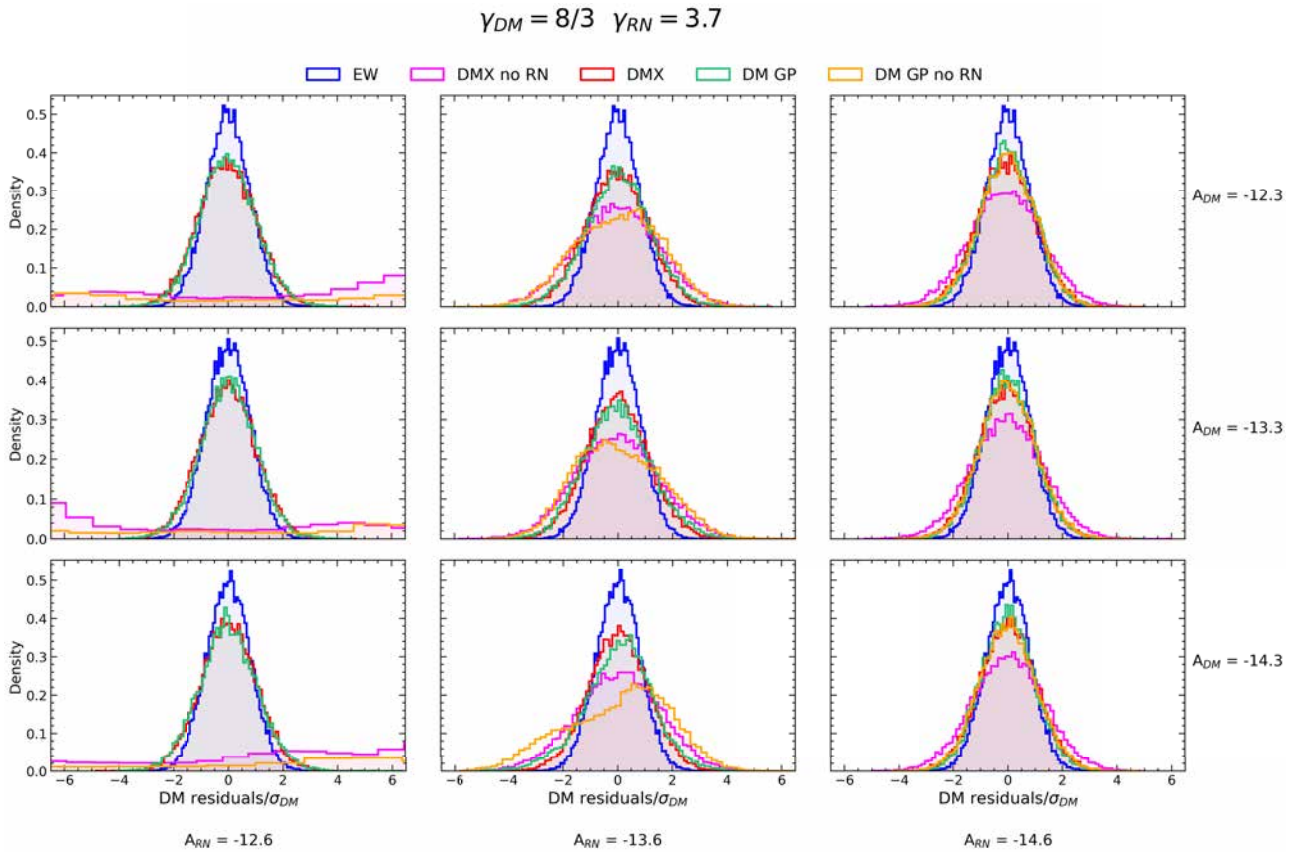


Fig. 4. Histograms of DM residuals normalized by their uncertainty for EW (blue), DMX no RN (magenta), DM GP no RN (orange), DMX (red), DM GP (green). A_{RN} decreases going left to right (as indicated at the bottom of the columns); A_{DM} increases going upwards (as indicated on the right of each row). Spectral indices are $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$.

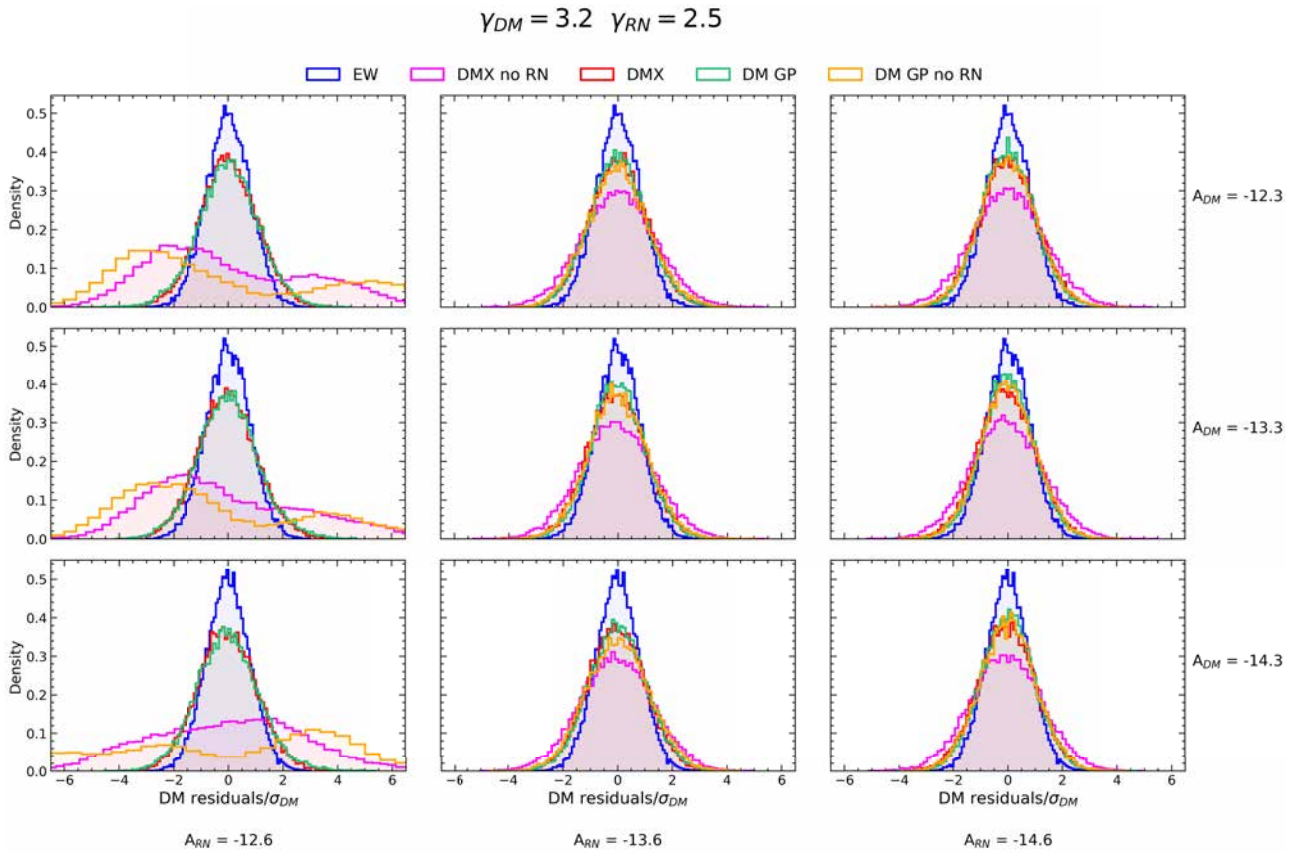


Fig. 5. As in Figure 4. Spectral indices are $\gamma_{RN} = 2.5$, $\gamma_{DM} = 3.2$.

and the true value³. The accuracy test is conducted by studying the histograms in Figures 4 and 5. These are obtained by normalizing the DM residuals by their uncertainties, hence offering a measure that serves as a proxy for the deviation of each bin in the histogram from the true value in terms of sigmas.

At first, we check the fraction of residuals that lies within the 3σ threshold for each method, which for a Gaussian distribution is expected to be 99.7%. For **EW** more than 99.9% of the values fall within 3σ , and this happens mostly because the uncertainties on the DM measurements are the highest among the methods. **DMX** is the second most accurate method, with a percentage of data points within 3σ greater than 99.5%, followed by **DM GP** with 98.5%. The remaining methods converge to higher percentages, comparable with the three just mentioned, only when the RN amplitude is at the lowest simulated level (-14.3).

4.3. Whitening of the DM residuals

Other than precision and accuracy, we also investigate the ability of each method to whiten the DM residuals, meaning that we check if they are random independent without correlated structures. This is extremely important for noise modeling, since any structure left in the DM residuals is associated with either the absorption of power in other noise components or with artifacts introduced by the method itself. We have conducted a Gaussianity test on the normalized DM residuals, and the associated reduced chi-squared values χ_{red}^2 are reported in Tables 4 and 5. For **EW**, $0.87 \leq \chi_{red}^2 \leq 1.35$, which implies that the data are well described by a Gaussian distribution and that there are no, or very few, remaining time-correlated structures in the DM residuals. For **DMX**, $0.71 \leq \chi_{red}^2 \leq 1.49$, showing that **EW** does not outperforms appreciably **DMX** in whitening the DM residuals. For **DMX no RN** and **DM GP no RN** the results are as expected: as long as A_{RN} is high, they are not able to properly recover the DM values and hence the χ_{red}^2 associated with the histograms of the normalized residuals is very high. If $A_{RN} = -13.6$, **DMX no RN** can reach χ_{red}^2 values which are closer to 1 with respect to **DM GP no RN**, which does so only when $A_{RN} = -14.6$. Lastly, **DM GP**, even though it is the most precise method, is not the most accurate and also not the best in whitening the DM residuals. The distributions in Figures 4 and 5 are most of the times similar to the **DMX** ones, but the corresponding χ_{red}^2 are not as close to 1. Time correlated structures, both short and long-term, are still present in **DM GP** DM residuals, emphasized by the aforementioned reduced errorbars. The short-term time correlated structures are caused by the high Fourier-frequency components of the injected signal (the highest of which is $f_{30} = 30/T_{span}$), while long-term ones are mainly due to the presence of RN in the data. A more explicit representation of this feature is shown in Figure 6, where the first two rows show DM residuals for a single and 100 realizations, respectively, using **DM GP** (similar results to **DM GP** can be obtained with **DMX**). The third row reports the comparison between the structure left in the DM residuals and the injected RN, showing that the DM modeling is indeed absorbing power from the RN process, as it displays the same signature of the injected RN signal. However, we also need to consider whether these structures have a significant impact on the timing residuals by comparing their power spectrum with the WN level. To do so we computed a discrete Fourier transform and subsequently the power spectrum of the DM residuals at different A_{RN} and compared it with the level of the WN

injected for each of the three methods. The result is shown in Figure 7. On average, the WN dominates over the DM residuals which confirms that signature remaining in the DM residuals cannot significantly be detected, and therefore it will not affect the timing residuals. On the other hand, we can say that **EW** is not affected at all by the strength of the RN, while it significantly influences **DM GP** and **DMX**.

Importantly, Figure 7 shows that the remaining power associated with the DM residuals is the lowest for **DM GP** (and **DMX**). This hints that **DM GP** (and **DMX**) diminishes the dispersive noise level in the ToAs to the minimum among the techniques tested, a result that would be beneficial in studies that need the most optimal noise mitigation possible. Hence, while a full study of the **DM GP** analysis and its interaction with various types of RN (particularly GWs) is beyond the scope of this paper, this finding indicates that **DM GP** (and **DMX**) might work best for whitening timing residuals, for example in a context such as the search for low-frequency GWs.

4.4. Alternative simulation settings

Here we tested the methods on datasets that we simulated by varying specific settings: first, we reduced the uncertainties of the simulated ToAs from our initial value (from $5\mu s$ to $1\mu s$ and $0.1\mu s$) to improve our sensitivity to the RN process and DM variations, then we simulated the ToAs at L-band frequencies and finally we changed the injected DM time series.

4.4.1. Testing different ToA uncertainties

We used $\sigma_{temp} = 5\mu s$ since it represents the typical mean uncertainties on ToAs obtained from LOFAR observations (Donner et al. 2020). However, there might be pulsars with higher ToA precision, and we expect that future facilities such as the low-frequency part of the Square Kilometer Array (Janssen et al. 2015) will improve the ToA sensitivity. For this reason, we run two new simulation sets, with decreased values for the σ_{temp} , firstly at $1\mu s$ and then at $0.1\mu s$. Furthermore, as this reduces the WN floor level, we enhance the methods' capability to identify RN and properly disentangle it from DM variations. The WN parameters injected in the simulations were also changed to: $EFAC = 1$, $EQUAD = 10^{-8} s$. Chromatic and achromatic red processes were set both to 30 Fourier-frequency components, with $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$, $A_{RN} = -13.6$ and $A_{DM} = -13.3$.

In Table 6 we report the DM uncertainty levels for each method and each uncertainty. Given the smaller errorbars of the ToAs, the precision on the DM measurements increased, as expected. As we did previously, we can rank **DM GP** as the most precise method followed by **DMX** and **EW**.

We recall that an important finding from our simulations (see Section 4.3) was about the RN signal left in the DM residuals. In fact, Figure 6 (bottom panel) shows that the signal left in the DM residuals matches exactly the one of the injected RN. The corresponding delay, that is thus left on average in the timing residuals, was about $\pm 2\mu s$. This changes in the two new cases characterized by low ToA uncertainties: when $\sigma_{temp} = 1\mu s$ the mean DM signal that is left due to an erroneous RN absorption contributes to the timing residuals at the level of ± 150 ns; when $\sigma_{temp} = 0.1\mu s$ this becomes ± 20 ns. Since the injected RN signal was the same, we can confirm that higher precision on the ToAs allows to increase the sensitivity to the RN and improve the accuracy of the measurements of DM variations.

³ ISO 5725-1:2023 in International Organization for Standardization (<https://www.iso.org/home.html>)

Table 4. Reduced χ^2 of a Gaussian fit to the histograms reported in Figure 4. $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$.

	χ_{red}^2					
A_{RN}	EW	DMX	DM GP	DMX no RN	DM GP no RN	A_{DM}
-12.6	0.95	1.37	1.53	1108.49	2691.50	-12.3
-13.6	0.91	1.04	1.73	1.34	18.75	-12.3
-14.6	0.91	1.10	1.21	0.80	1.25	-12.3
-12.6	0.87	0.78	1.67	6154.89	6538.24	-13.3
-13.6	0.88	1.01	2.45	1.33	26.07	-13.3
-14.6	0.90	0.84	1.35	0.76	1.00	-13.3
-12.6	0.89	0.86	1.63	1104.90	1579.96	-14.3
-13.6	0.89	0.71	5.25	1.53	71.90	-14.3
-14.6	0.87	0.83	1.26	0.73	0.93	-14.3

Table 5. Reduced χ^2 of obtained from a Gaussian fit on the histograms χ reported in Figure 5. $\gamma_{RN} = 2.5$, $\gamma_{DM} = 3.2$.

	χ_{red}^2					
A_{RN}	EW	DMX	DM GP	DMX no RN	DM GP no RN	A_{DM}
-12.6	1.35	0.98	1.96	365.57	218.71	-12.3
-13.6	1.35	0.82	1.45	0.90	1.99	-12.3
-14.6	1.35	1.03	1.09	0.94	0.98	-12.3
-12.6	1.17	1.49	3.35	168.97	194.42	-13.3
-13.6	1.16	1.32	1.58	1.02	1.20	-13.3
-14.6	1.17	0.76	1.13	0.97	1.12	-13.3
-12.6	0.98	1.29	2.20	33.12	141.30	-14.3
-13.6	0.98	0.96	0.95	0.74	1.79	-14.3
-14.6	0.98	0.95	0.77	0.77	1.45	-14.3

Table 6. Mean DM uncertainty measured with the **EW**, **DMX** and **DM GP** methods over simulated ToAs with uncertainties fixed at $1\mu\text{s}$ and $0.1\mu\text{s}$

σ_{temp}	$\langle\sigma_{DM}\rangle \cdot 10^{-6}$ (pc cm $^{-3}$)		
	EW	DMX	DM GP
$1\mu\text{s}$	6.3	2.0	1.6
$0.1\mu\text{s}$	0.6	0.5	0.2

4.4.2. Testing ToAs at L-band

In this part we report the results obtained by applying the DM-computation methods to L-band ToAs. To simulate high-frequency ToAs, we refer to the data collected by the Nançay decimeter radiotelescope (NRT), which has a relatively large bandwidth (512 MHz) centered at 1.4 GHz divided in 4 sub-bands (EPTA Collaboration et al. 2023b). The input parameter are: EFAC = 1, EQUAD = 10^{-8} s, $A_{DM} = -13.3$ and $\gamma_{DM} = 8/3$. We use a conservative $1\mu\text{s}$ ToA template-fitting error.

Concerning the precision of the methods, the results are similar as for the LOFAR frequency case: **DM GP** is the most precise, followed by **DMX** and **EW**. However, the size of these errors is now larger by approximately two orders of magnitude (see Table 7). This is due to two main reasons: the first is the increased central frequency of the analysed ToAs (which is 1.4 GHz against the 153 MHz of the previous LOFAR-like case), because DM is inversely proportional to the squared of the observing frequency. The second is the fractional bandwidth B (see Equation 5): the higher B , the better the precision in the DM measurements (Verbiest & Shaifullah 2018). For the NRT-like case $B \approx 0.34$, while for the LOFAR-like case $B \approx 0.41$. Concerning accuracy, we also report similar conclusions as be-

Table 7. As in Table 6 but for L-band ToAs with $\sigma_{temp} = 1\mu\text{s}$ and injected DM noise parameters: $A_{DM} = -13.6$, $\gamma_{DM} = 3.7$.

$\langle\sigma_{DM}\rangle \cdot 10^{-4}$ (pc cm $^{-3}$)		
EW	DMX	DM GP
7.8	2.1	0.7

Table 8. Reduced χ^2 of the normalized DM residuals.

χ_{red}^2		
EW	DMX	DM GP
1.07	1.10	1.54

Notes. It has been obtained with a Gaussian fit on histograms of DM residuals normalized by their uncertainty for the three main tested methods (**EW**, **DMX** and **DM GP**) over ToAs simulated at L-band with $\sigma_{temp} = 1\mu\text{s}$ and injected DM noise with A_{DM} , γ_{DM} of, respectively, -13.3 and $8/3$.

fore: **EW** has more than 99.76% of the values within 3σ and it is the most accurate method, followed by **DMX** with more than 99.44% and finally **DM GP** with 99.2%. As well, the Gaussianity test provides similar results to the LOFAR-like case. In Table 8 are listed the reduced χ^2 values which confirm that on average **EW** and **DMX** are working marginally better in whitening the residuals due to their larger errorbars (and thus lower precision), on single DM recoveries. In the top panel of Figure 8 are presented the DM residuals for each of the three methods, while the corresponding normalized histograms are reported in the bottom plot. Figure 9 shows the power spectra of the DM residuals compared with the injected signal. Despite the fact that

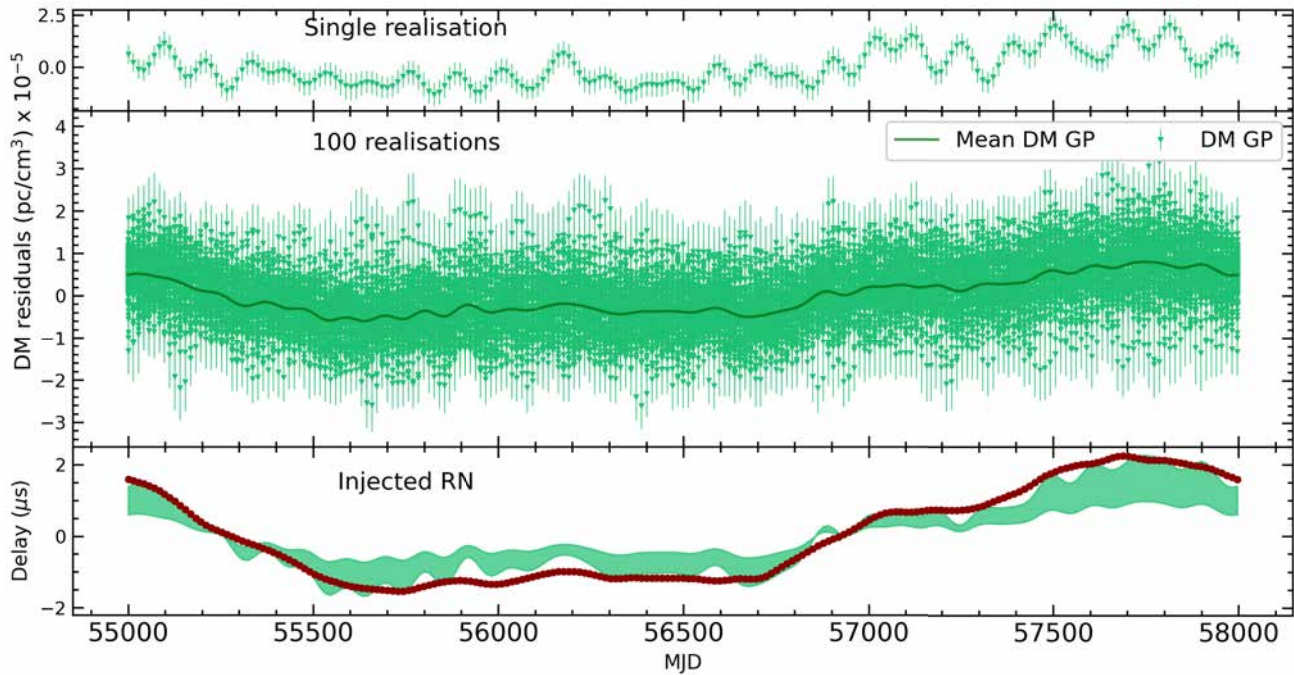


Fig. 6. The top panel shows the DM residuals obtained by applying **DM GP** (with RN modeling) to an individual simulation over the 100 realized for values of the noise parameters $A_{RN} = -13.6$, $A_{DM} = -13.3$, $\gamma_{RN} = 3.7$, $\gamma_{DM} = 8/3$, EFAC and EQUAD as Table 1 and $\sigma_{\text{ToA}} = 5\mu\text{s}$. The second panel reports an average of the DM residuals left by **DM GP** over the 100 simulations performed with the same parameters. The bottom panel shows the injected RN signal (red dots) compared to the time delay associated with the mean **DM GP** (shaded green). Note that we have a shaded region because a single DM value causes a frequency-dependent delay. Hence, the edges of the green region correspond to the maximum and minimum frequencies of the band.

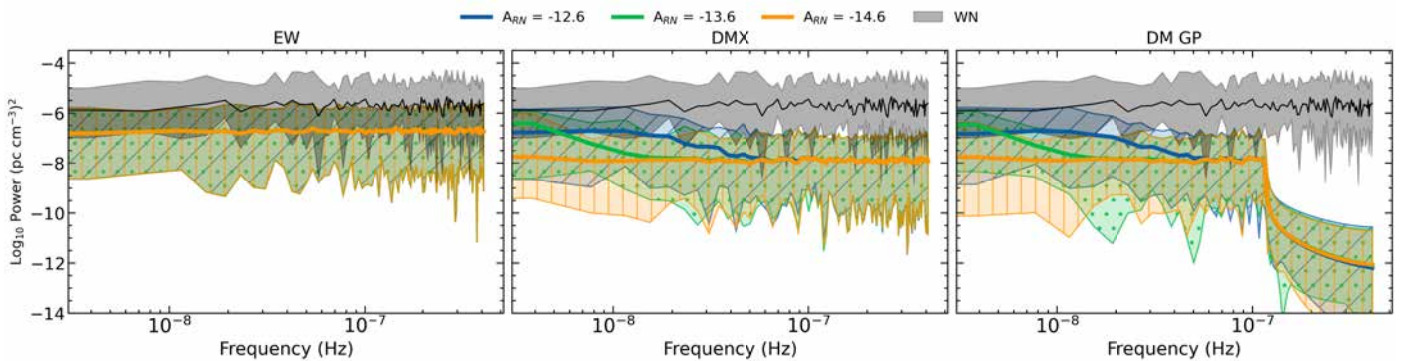


Fig. 7. Power spectra of the DM residuals left by the **EW** (left panel), **DMX** (central panel), and **DM GP** (right panel) methods. We also show the WN level (black) and the power spectra for different RN amplitudes: $A_{RN} = -14.6$ orange with vertical dashes; $A_{RN} = -13.6$ green with dots; $A_{RN} = -12.6$ blue with oblique dashes. With **DM GP** we model the noise up to a Fourier-frequency of $30/T_{\text{span}} \sim 1.15 \times 10^{-7}$ Hz, the same as the injected signal. That is why in the right-hand panel there is a drop at that specific frequency of the power spectrum of the **DM GP** residuals.

the WN is dominating in most of the frequency bins, all of the three methods can model the long-term structures present in the data. At the highest frequencies in the spectra, **EW** and **DMX** cannot match the injected signal anymore due to the limits imposed by the WN. Instead, **DM GP** is the only method that produces a power spectrum which does not cross the injection line, meaning that it is not adding any more power at those frequencies. Note that **DM GP** uses a power-law model to describe the noise process, which is identical to the functional form of the injected signal. Therefore, it is possible that the reported results show a biased result for **DM GP**, since in real-life datasets we may not have pure power-law noise processes.

To evaluate a more realistic case, in the following section we test the methods on an injected DM time series which does not come from a pure power-law.

4.4.3. Custom DM injection

Here we test the performance of the three considered techniques over an individual, injected DM time-series with an empirical power-spectrum shape, derived in Donner (2022) from the LO-FAR dataset of PSR J0139+5814 (see Figure 10). The ToAs are still evenly spaced, with an uncertainty set to $5\mu\text{s}$ and the injected WN is defined by EFAC = 1.2 and EQUAD = $2\mu\text{s}$. Given the short timescale fluctuations in the injected signal, we use 107 Fourier components to define the power-spectrum used in **DM GP**, that corresponds to the limit imposed by the Nyquist theorem given the 215 data points. The result of the analysis are reported in Figure 11 where we show the DM residuals in the top panel and their power spectra in the bottom panel. As previously noted, **EW** and **DMX** provide independent DM measurements

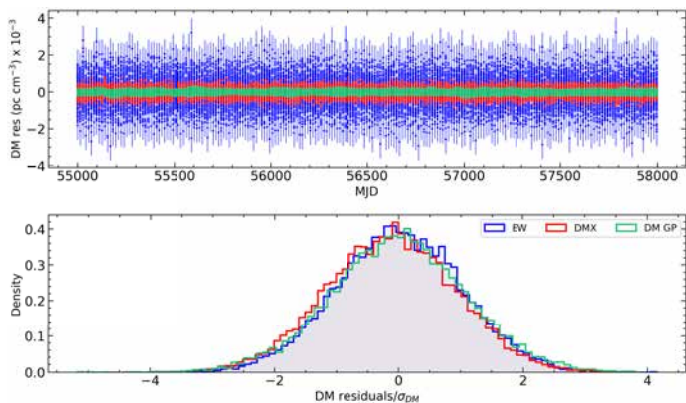


Fig. 8. Top panel: DM residuals left by the **EW** (blue), **DMX** (red), **DM GP** (green) methods when run over the 100 realizations of L-band ToAs. Bottom panel: histograms of the DM residuals normalized by their uncertainty. The input noise parameters are: $A_{DM} = -13.3$, $\gamma_{DM} = 8/3$

(although **DMX** might introduce some degree of correlations for close-by points, see Lam et al. 2017), while **DM GP** assumes an underlying model for the power spectrum to describe the DM variations. The **DM GP** errorbars are now larger than the **DMX** ones and hence the recoveries are extremely accurate. The bottom panel of Figure 11 shows that even with the injected signal resembling a broken power law, **DM GP** was able to absorb the signal and hence successfully recover the injected DM time series.

5. Conclusion

In this paper, we carried out a series of simulations in order to test the performance of the three main methods available in the PTA community to calculate DM variations: **EW**, **DM GP** and **DMX**. The final aim is to provide information on the calculation of the DM time series for, e.g., studies of the IISM and the Solar wind.

Our main conclusion is that, while they all perform well, the most accurate method to calculate DM variations is **EW**, but **DM GP** and **DMX** seem to lower the level of the DM noise in the timing residuals the most.

In more detail, we conclude that all of the methods perform satisfactorily (as long as the RN and DM models are selected correctly for **DMX** and **DM GP**). Figure 3 demonstrates that when the RN power is significant, it is unavoidable to include it in the noise model, otherwise the achromatic, time-correlated structures will be absorbed in the DM modeling in the cases of **DMX** and **DM GP**.

With respect to the precision, the mean uncertainties $\langle\sigma_{DM}\rangle$ on the DM recoveries for **EW** are independent of the amplitudes of the red noise processes. This means that the method is only affected by the ToA precision and the frequency coverage of the data. For **DMX** and **DM GP**, instead, $\langle\sigma_{DM}\rangle$ depends on the amplitude of the injected RN process (see Tables 2 and 3). Overall, the method which gives the most precise DM measurements is **DM GP**.

We also studied the accuracy, defined as the level of agreement between a test result and the true value while taking into account the uncertainties, by analyzing the distributions of the normalized DM residuals (Figures 4 and 5). Because of the larger errorbars, **EW** is the most accurate method having in most of the cases $>99.9\%$ of the data within 3σ . The second most accu-

rate method, according to this criterion, is **DMX** and in the end **DM GP**, which presents more data points in the tail of the distribution because of its high DM precision. This is true not only at LOFAR frequencies, but also in the L-band trials, where we simulated a NRT-like frequency coverage.

The accuracy test, though, as described above, does not imply that the DM residuals are white. Hence, we have conducted a Gaussianity test on each of the normalized DM residuals distributions, which informs on the ability of each method in removing time-correlated structures from the ToAs. The χ^2_{red} of the Gaussian fits are reported in Tables 4 and 5, and they show that **EW** is the method with the narrowest range of χ^2_{red} around 1. **DMX** shows similar results, but only when the RN is modeled. However, the χ^2_{red} range shown by **DMX** fluctuates more than for **EW**. **DM GP**, which uses a linear combination of Fourier components, produces a time series where close data points are not completely independent. This affects the overall residual distribution, resulting in the widest χ^2_{red} range spanned among the three methods.

To better understand the reasons behind the **DMX** and **DM GP** distributions, we studied the effect of the RN on the DM residuals. In Figure 6 we show that even when the RN is modeled, a signature can still be found in the DM residuals yielded by **DM GP** and **DMX**. This happens when the WN level exceeds the RN, especially at the highest Fourier frequencies. Nonetheless, Figure 7 shows that, at LOFAR frequencies, the unmodeled RN signal remains below the WN level. In order to properly disentangle the RN and DM noise signal it is necessary to have high-precision ToAs. In fact, we investigated this scenario by simulating datasets with reduced uncertainties, $\sigma_{ToAs} = 1 \mu s$, $0.1 \mu s$, and, as expected, **DMX** and **DM GP** do not show any significant structures related to the injected RN signal in the DM residuals anymore. The aforementioned precision-accuracy hierarchy holds in this simulation scenario as well. Similar conclusions are reached in (albeit limited) simulations carried out at L-band and with an injected DM time series with empirical spectral shape.

However, we stress that these results are based on simulations that do not take into account uneven observational cadence, variable ToA uncertainty, poor, if not absent, frequency coverage of the data and also combination of datasets, i.e., the most common characteristics of real-life data. When these additional complexities are present, it is possible that **DM GP** performs better than other methods thanks to its flexibility and the fact that its characteristics allow a more optimal blending of data with diverse properties. Moreover, the performance of **EW** and **DMX** might be affected by more realistic datasets because of their windowing constraints. A firmer conclusion will be obtained once these scenarios are taken into consideration. Also, the impact of these DM recovery schemes on timing model parameters has not been assessed in our work, but Kramer et al. 2021 already showed that astrometric parameters can be corrupted by some of the schemes tested. Consequently, future work on this topic would require a complete assessment of that aspect, as well.

Last, we reached the important conclusion that **DM GP** and **DMX** seem to have the capacity to reduce the DM noise level to the minimum (see Figure 7). Future works should assess the effectiveness of the analyzed methods in modeling DM variations as a source of noise when a GW background is also present in the data.

Acknowledgements. FI is supported by the University of Cagliari (IT). FI, CT, AP are supported by the Istituto Nazionale di Astrofisica. ACh and GMS acknowledge financial support provided under the European Union's H2020 ERC

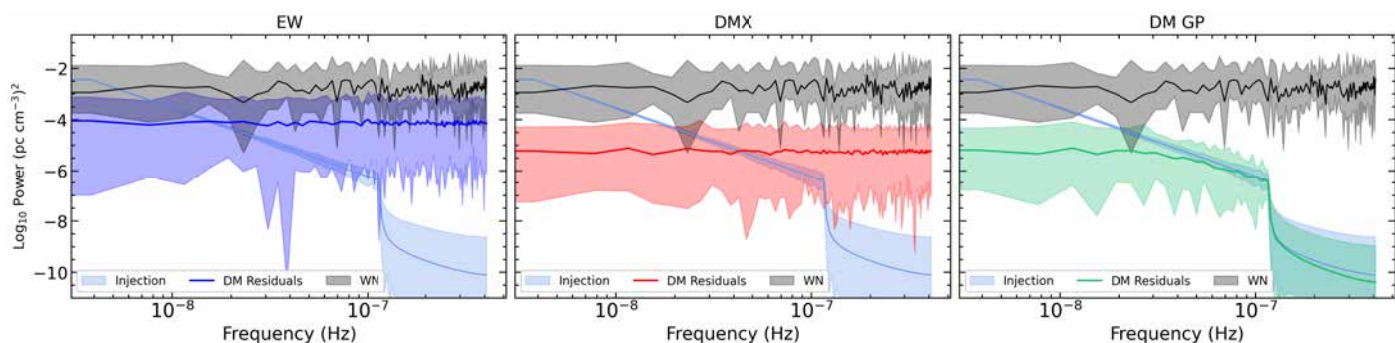


Fig. 9. Power spectra of the DM residuals obtained using **EW** (left), **DMX** (center), **DM GP** (right), compared with the injected signal (light blue) at L-band. Input noise parameters: $A_{DM} = -13.3$, $\gamma_{DM} = 8/3$. The injected signal power (light-blue) drops at a frequency of $30/T_{span} \sim 1.15 \times 10^{-7}$ Hz. We use the same number of frequency components in the modeling with **DM GP**.

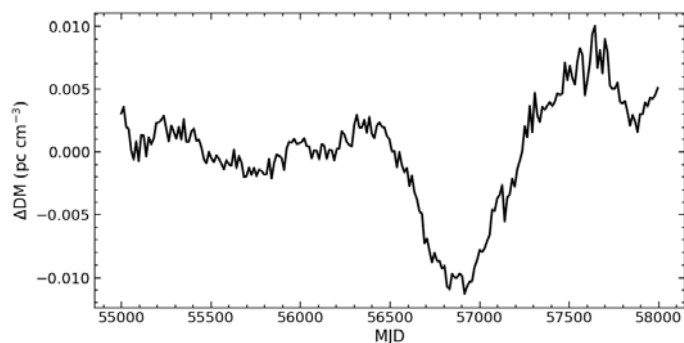


Fig. 10. DM time series injected in the data with an empirical power spectrum. (Donner 2022)

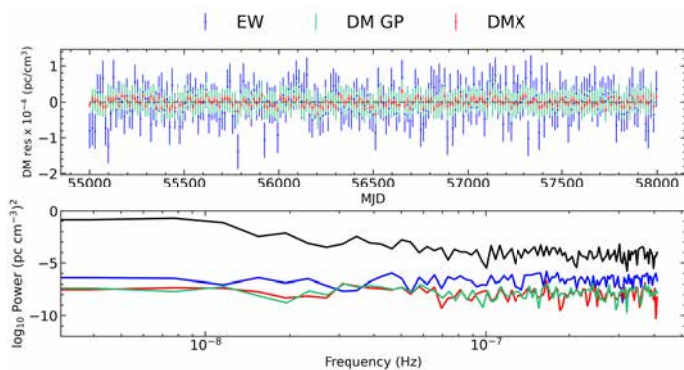


Fig. 11. Top panel: DM residuals for **EW** (blue), **DM GP** (green), **DMX** (red) obtained by taking the difference of the DMs recovered by each method and the injected DM time series of Figure 10 with unknown power spectrum. WN parameters are $EFAC = 1.2$ and $EQUAD 2\mu s$ and ToA uncertainty set at $5\mu s$. Bottom panel: Power spectra of the injected signal (black) and of the DM residuals.

Consolidator Grant “Binary Massive Black Hole Astrophysics” (B Massive, Grant Agreement: 818691). JPWV acknowledges support from NSF AccelNet award No. 2114721. SCS acknowledges the support of a College of Science and Engineering University of Galway Postgraduate Scholarship. MTL graciously acknowledges support received from NSF AAG award number 2009468, and NSF Physics Frontiers Center award number 2020265, which supports the NANOGrav project. Work at NRL is supported by NASA.

References

Agazie, G., Alam, M. F., Anumarlapudi, A., et al. 2023, *ApJ*, 951, L9
 Armstrong, J. W., Rickett, B. J., & Spangler, S. R. 1995, *ApJ*, 443, 209

- Backer, D. C., Kulkarni, S. R., Heiles, C., Davis, M., & Goss, W. 1982, *Nature*, 300, 615
 Bondonneau, L., Griefmeier, J. M., Theureau, G., et al. 2021, *A&A*, 652, A34
 Donner, J. Y. 2022, PhD thesis, Universität Bielefeld, Bielefeld (Germany)
 Donner, J. Y., Verbiest, J. P. W., Tiburzi, C., et al. 2020, *A&A*, 644, A153
 Donner, J. Y., Verbiest, J. P. W., Tiburzi, C., et al. 2019, *A&A*, 624, A22
 Edwards, R. T., Hobbs, G. B., & Manchester, R. N. 2006, *MNRAS*, 372, 1549
 Ellis, J. & Van Haasteren, R. 2017, *jellis18/PTMCMCSampler: Official Release*
 Ellis, J. A., Vallisneri, M., Taylor, S. R., & Baker, P. T. 2020, *ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust Pulsar Inference Suite*
 EPTA Collaboration, InPTA Collaboration, Antoniadis, J., et al. 2023a, *A&A*, 678, A49
 EPTA Collaboration, InPTA Collaboration, Antoniadis, J., et al. 2023b, *A&A*, 678, A48
 Foster, R. S. & Cordes, J. M. 1990, *ApJ*, 364, 123
 Hassall, T. E., Stappers, B. W., Hessels, J. W. T., et al. 2012, *A&A*, 543, A66
 Hazboun, J. S. 2020, *La Forge*
 Hellings, R. W. & Downs, G. S. 1983, *ApJ*, 265, L39
 Hobbs, G., Lorimer, D. R., Lyne, A. G., & Kramer, M. 2005, *MNRAS*, 360, 974
 Hobbs, G., Lyne, A. G., Kramer, M., Martin, C. E., & Jordan, C. 2004, *MNRAS*, 353, 1311
 Janssen, G., Hobbs, G., McLaughlin, M., et al. 2015, in *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, 37
 Jones, M. L., McLaughlin, M. A., Lam, M. T., et al. 2017, *ApJ*, 841, 125
 Keith, M. J., Johnston, S., Karastergiou, A., et al. 2024, *MNRAS*, 530, 1581
 Kramer, M., Stairs, I. H., Manchester, R. N., et al. 2021, *Phys. Rev. X*, 11, 041050
 Krishnakumar, M. A., Manoharan, P. K., Joshi, B. C., et al. 2021, *A&A*, 651, A5
 Lam, M. T., Cordes, J. M., Chatterjee, S., et al. 2017, *ApJ*, 834, 35
 Lam, M. T., Cordes, J. M., Chatterjee, S., & Dolch, T. 2015, *ApJ*, 801, 130
 Lam, M. T., Cordes, J. M., Chatterjee, S., et al. 2016, *ApJ*, 821, 66
 Lentati, L., Alexander, P., Hobson, M. P., et al. 2013, *Phys. Rev. D*, 87, 104021
 Lorimer, D. R. & Kramer, M. 2005, *Handbook of pulsar astronomy*, Vol. 4 (Cambridge university press)
 Luo, J., Ransom, S., Demorest, P., et al. 2021, *ApJ*, 911, 45
 Maggiore, M. 2018, *Gravitational Waves: Volume 2: Astrophysics and Cosmology* (Oxford University Press)
 Phinney, E. S. 2001, *arXiv e-prints*, arXiv:0108028
 Rasmussen, C. E. & Williams, C. K. I. 2006, *Gaussian Processes for Machine Learning*
 Rickett, B. J. 1990, *ARA&A*, 28, 561
 Taylor, S. R. 2021, *arXiv e-prints*, arXiv:2105.13270
 Tiburzi, C. 2018, *PASA*, 35, e013
 Tiburzi, C., Shaifullah, G. M., Bassa, C. G., et al. 2021, *A&A*, 647, A84
 Tiburzi, C., Verbiest, J. P. W., Shaifullah, G. M., et al. 2019, *MNRAS*, 487, 394
 Vallisneri, M. 2020, *libstempo: Python wrapper for Tempo2*, *Astrophysics Source Code Library*, record ascl:2002.017
 van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, *A&A*, 556, A2
 van Haasteren, R. & Vallisneri, M. 2014, *Phys. Rev. D*, 90, 104012
 Verbiest, J. P. W., Lentati, L., Hobbs, G., et al. 2016, *MNRAS*, 458, 1267
 Verbiest, J. P. W., Osłowski, S., & Burke-Spolaor, S. 2021, *Pulsar Timing Array Experiments* (Springer Singapore), 1–42
 Verbiest, J. P. W. & Shaifullah, G. M. 2018, *Classical and Quantum Gravity*, 35, 133001
 Zarka, P., Girard, J. N., Tagger, M., & Denis, L. 2012, in *SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics*, ed. S. Boissier, P. de Laverny, N. Nardetto, R. Samadi, D. Valls-Gabaud, & H. Wozniak, 687–694

Appendix A: Evaluating time-domain realizations of stochastic time-correlated signals with LaForge

This part describes the method used in LaForge software to obtain time-domain realization of stochastic signals measured in the frequency domain. The stochastic and time-correlated signals in PTA data such as DM variations are mainly modeled as Gaussian processes (GPs). GPs are defined as a collection of (infinite) random variables representing the function values (e.g., time delay of the modeled signal in the timing residuals) for all input locations (i.e., observing epochs). Any finite set of these random variables has a joint Gaussian distribution. Thus, GPs allows us to define the mean value and the variance of the signal at any input location.

GPs can be fully specified with one of each following approaches (Rasmussen & Williams 2006):

- Weight-Space View, where the process is defined as a linear sum of deterministic basis functions $T_\mu(t)$ multiplied by weights a_μ as

$$f(t) \sim \sum_{\mu}^m a_{\mu} T_{\mu}(t), \quad (\text{A.1})$$

with $\mu = 1, \dots, m$ and where the weights are Gaussian random variables as $a_{\mu} \sim \mathcal{N}(a_{\mu}^0, \phi_{\mu\nu})$, with a_{μ}^0 assumed to be a zero vector in our case, and $\phi_{\mu\nu}$ corresponding to the weights covariance matrix.

The PTA likelihood including a GP described in the weight-space view can be written as

$$p(\delta t|a_{\mu}, GP) = \frac{\exp\left[-\frac{1}{2} \sum_{i,j} (\delta t_i - \sum_{\mu} a_{\mu} T_{\mu}(t_i)) N_{ij}^{-1} (\delta t_j - \sum_{\mu} a_{\mu} T_{\mu}(t_j))\right]}{\sqrt{(2\pi)^n \det(N)}} \times \frac{\exp\left[-\frac{1}{2} \sum_{\mu,\nu} a_{\mu} \phi_{\mu\nu}^{-1} a_{\nu}\right]}{\sqrt{(2\pi)^m \det(\phi)}}, \quad (\text{A.2})$$

with δt and N respectively corresponding to the timing residuals and the covariance matrix, $i, j = 1, \dots, n$ and $\mu, \nu = 1, \dots, m$.

- Function-Space View, where the GP directly describes the targeted signal as

$$f(t) \sim GP(g(t), k(t, t')), \quad (\text{A.3})$$

including the mean function $g(t)$ and the covariance function $k(t, t')$, also referred to as kernel.

The PTA likelihood in that form is marginalized over the basis weights mentioned above and can be expressed as

$$p(\delta t|GP) = \frac{\exp\left[-\frac{1}{2} \sum_{ij} \delta t_i (N_{ij} + K_{ij})^{-1} \delta t_j\right]}{\sqrt{(2\pi)^n \det(N + K)}}, \quad (\text{A.4})$$

where $K = k(t, t')$.

In PTA analysis, we usually employ Equation A.4 and express $k(t, t')$ from the correspondence between both views as (Rasmussen & Williams 2006)

$$k(t, t') = \sum_{\mu,\nu} T_{\mu}(t) \phi_{\mu\nu} T_{\nu}(t'), \quad (\text{A.5})$$

In practice, we set the basis functions T as a finite set of sine/cosine functions and

$$\phi_{\mu\nu} = P(f_{\mu}) \delta_{\mu\nu} / t_{\text{span}}, \quad (\text{A.6})$$

with r.h.s. terms respectively corresponding to the power spectral density (PSD), the Kronecker delta and the time span of the data set. This approach corresponds to the Fourier-sum approach described in (van Haasteren & Vallisneri 2014).

To summarize, we use a time-domain likelihood function with a parameterized covariance matrix that contains a PSD that is most often defined as a simple power-law described with two parameters: the amplitude A , usually set at the reference frequency of 1 yr^{-1} and the spectral slope γ .

The Gaussian process signal in the time-domain $f(t)$ can be evaluated from Eq. A.1 after estimating the weights a with Eq. A.2. Let us now describe a method to estimate realizations of a , where we perform a Maximum Likelihood Estimation on Eq. A.2 (Lentati et al. 2013). Let us rewrite the logarithm of this equation in a matrix notation,

$$\log [p(\delta t|a, GP)] = -\frac{1}{2} (\delta t - aT)^T N^{-1} (\delta t - aT) - \frac{1}{2} a^T \phi^{-1} a - \frac{1}{2} [(n+m) \log(2\pi) + \log(\det(N) \det(\phi))], \quad (\text{A.7})$$

where the third term is just a constant number "cst", thus

$$\begin{aligned} \log [p(\delta t|a, GP)] &= -\frac{1}{2} [\delta t^T N^{-1} \delta t + (aT)^T N^{-1} (aT) + a^T \phi^{-1} a] \\ &\quad + (T^T N^{-1} \delta t)^T a + \text{cst} \\ &= -\frac{1}{2} [\delta t^T N^{-1} \delta t + a^T (T^T N^{-1} T + \phi^{-1}) a] \\ &\quad + (T^T N^{-1} \delta t)^T a + \text{cst}. \end{aligned} \quad (\text{A.8})$$

Let us now write its derivative over the weights,

$$\frac{\partial \log [p(\delta t|a, GP)]}{\partial a} = (T^T N^{-1} T + \phi^{-1}) a - (T^T N^{-1} \delta t)^T, \quad (\text{A.9})$$

and finally obtain the maximum likelihood vector of weights \hat{a} ,

$$\hat{a} = (T^T N^{-1} T + \phi^{-1})^{-1} (T^T N^{-1} \delta t)^T, \quad (\text{A.10})$$

where we evaluate ϕ using Eq. A.6, for which $P(f)$ is computed by using posterior distributions of our model parameters (e.g., A and γ if defined as a power-law) evaluated from a Bayesian analysis. Furthermore, N is computed after applying white noise parameters (e.g., EFAC, EQUAD) with values taken from the same posterior distributions. This way, we evaluate time-domain realizations that are marginalized over other processes included during the Bayesian analysis.

Appendix B: Achromatic noise recoveries with DMX and DM GP

In this section we report on how **DMX** and **DM GP** methods are able to recover achromatic noise parameters from our simulations.

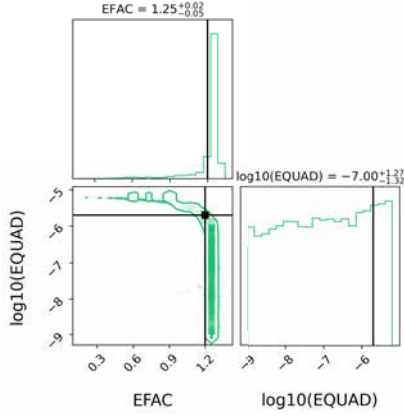


Fig. B.1. Corner plot of WN parameters using **DM GP** for an individual realization (see Section 2) with injected noise parameters: $A_{DM} = -13.3$, $\gamma_{DM} = -2.7$, $A_{RN} = -13.6$ and $\gamma_{RN} = -3.7$. Black lines correspond to the injected values.

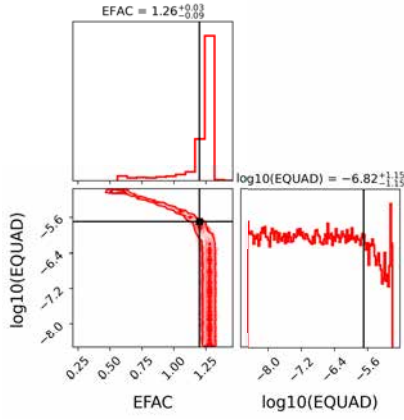


Fig. B.2. Corner plot of WN parameters using **DMX** for an individual realization (see Section 2) with injected noise parameters: $A_{DM} = -13.3$, $\gamma_{DM} = -2.7$, $A_{RN} = -13.6$ and $\gamma_{RN} = -3.7$. Black lines correspond to the injected values.

Appendix B.1: White Noise

WN parameters are inferred during the **DMX** and **DM GP** analysis through Equation (6) in the *enterprise* software suite. Since in our simulations the ToA uncertainties are all the same, the EFAC and EQUAD parameters are strongly correlated as shown in Figures B.1 and B.2. Overall, the EFAC parameters are always well constrained and consistent with the injected values for both **DM GP** and **DMX**. On the other hand, the EQUADs show a flat posterior (meaning that they are totally unconstrained) but still consistent with the injection. Moreover, **DM GP** tends to present a small peak at higher EQUADs, while **DMX** does not. In order to disentangle EFAC and EQUAD parameters, simulations with variable ToA uncertainties are required.

Appendix B.2: Achromatic red noise

In Figures B.3 and B.4 we report the 100 posterior distributions of RN parameters obtained from the simulations described in section 2. We show the only fiducial case with $A_{DM} = -13.3$ and varying A_{RN} . *Enterprise* can well constrain the RN parameters alongside **DM GP** and **DMX** as long as the power spectrum of the injected RN is above the WN floor (i.e., when

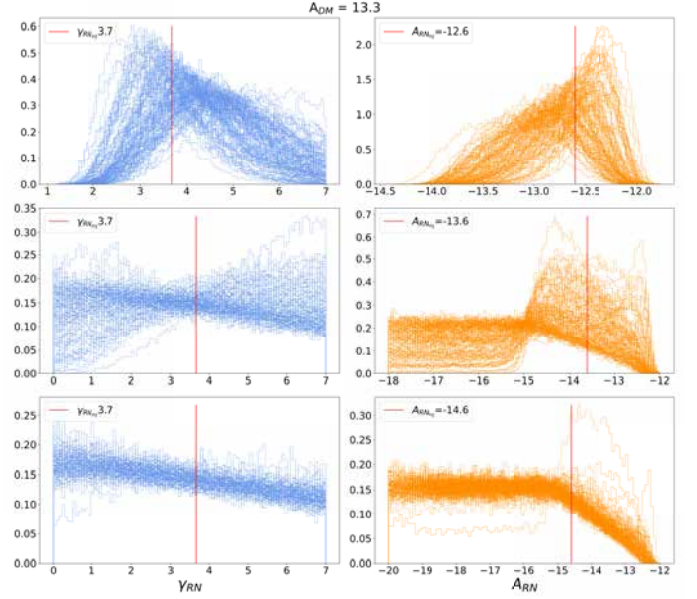


Fig. B.3. Posterior distributions of RN parameters using **DM GP** for 100 realizations (see Section 2) with injected DM noise parameters: $A_{DM} = -13.3$, $\gamma_{DM} = -2.7$. On the left-hand column there is the spectral index γ_{RN} ; on the right-hand column there is the RN amplitude A_{RN} . Each row correspond to a different injected A_{RN} . Red vertical lines report the injected values.

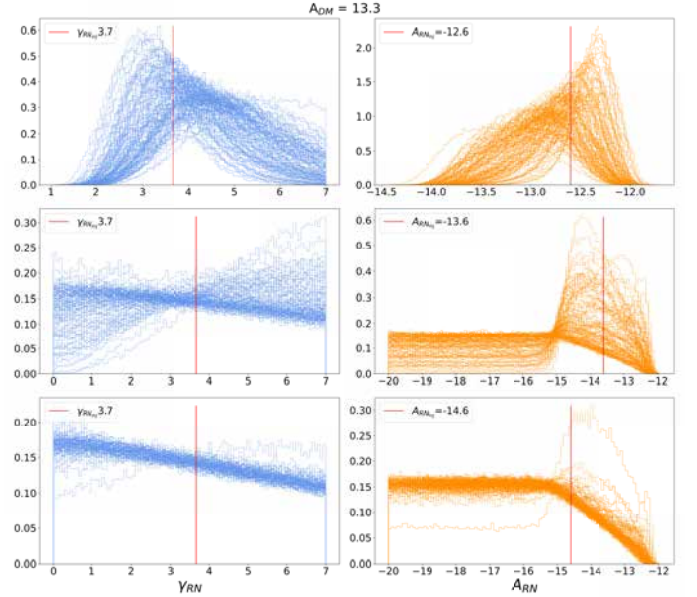


Fig. B.4. Posterior distributions of RN parameters using **DMX** for 100 realizations (see Section 2) with injected DM noise parameters: $A_{DM} = -13.3$, $\gamma_{DM} = -2.7$. On the left-hand column there is the spectral index γ_{RN} ; on the right-hand column there is the RN amplitude A_{RN} . Each row correspond to a different injected A_{RN} . Red vertical lines report the injected values.

$A_{RN} = -12.6$ and barely for $A_{RN} = -13.6$. See also Figure 1). When this does not happen, the posterior distributions of A_{RN} are flat and unconstrained. In all of the cases the recoveries are consistent with the injected values. This result is also confirmed when we reduce the ToA uncertainties to 1 or 0.1 μ s, hence decreasing the WN power level. In these cases the RN pa-

parameters are well constrained and consistent with the injection even for lower A_{RN} .