Taiji Data Challenge II Manual

(Taiji Data Challenge Working Group)

(Dated: May 20, 2025)

CONTENTS

I. INTRODUCTION

A. Taiji project for space-based gravitational wave detection

Taiji is a Chinese space mission proposed to detect gravitational waves (GWs) in the 0.1 mHz - 1 Hz frequency band [1, 2], which is expected to be launched in the 2030s. Similar to the European mission Laser Interferometer Space Antenna (LISA), the Taiji detector consists of three spacecrafts (SCs), and each SC follows a heliocentric orbit, forming a giant equilateral triangle with nominal arm lengths of approximately 3 million kilometers. The center of mass of the constellation leads the Earth by about 20 degrees, and is about 1 AU away from the Sun.

The science operation of Taiji will last for at least 5 years, during which it will be observing burst, continuous and stochastic GW signals. Taiji's target GW sources include $\mathcal{O}(10^7)$ Galactic and extra-Galactic binaries (GBs) ($\mathcal{O}(10^4)$ resolvable, others forming a confusion foreground), $\mathcal{O}(10)$ - $\mathcal{O}(10^2)$ massive black hole binaries (MBHBs), $\mathcal{O}(1)$ - $\mathcal{O}(10^3)$ extreme mass-ratio inspirals (EMRIs), $\mathcal{O}(1)$ - $\mathcal{O}(10)$ stellar-mass black hole binaries (sBHBs), as well as the astrophysical and/or cosmological stochastic GW backgrounds (SGWBs), *etc*.

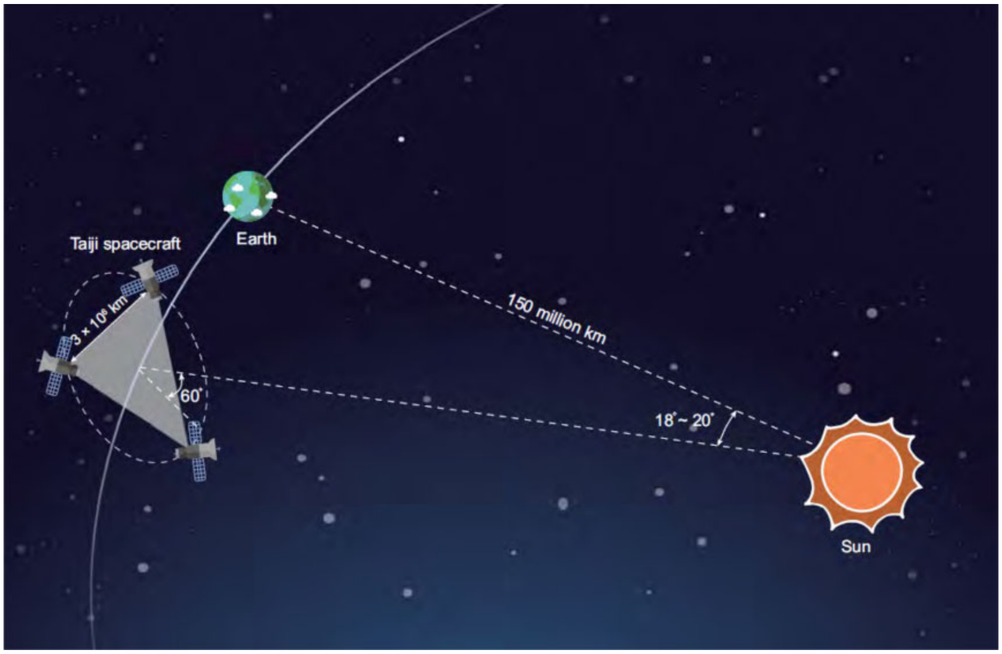

FIG. 1. Taiji mission concept (Credit: Ref. [1]).

B. A brief introduction to Taiji Data Challenge II and `Triangle`

Compared to the current observations of LIGO-Virgo-KAGRA, the data of Taiji is predicted to exhibit several distinct features: long duration, source overlapping in the time and frequency domains (especially a considerable fraction of signals with high signal-to-noise ratios (SNRs) up to $\mathcal{O}(10^2)$ - $\mathcal{O}(10^3)$), as well as inevitable glitches, gaps, and non-stationary noises during the observation period of signals, *etc*. Consequently, "global fit", namely simultaneously fitting all the parameters of GW sources and detector instruments, has been regarded as the primary challenge for space-based GW data analysis, necessitating both high speed and high accuracy of the algorithms. To provide testbeds for these algorithms, both LISA and Taiji have released their simulation datasets, namely the "Data Challenges", such as the mock LISA data challenge (MLDC) [3], LISA data challenge (LDC) [4], and the 1st Taiji data challenge (TDC I) [5].

As more than one group has published their prototype global fit analyses [6–8], LDC has basically fulfilled its purpose as a simulation dataset based on idealized orbit configurations, instrumental noises and GW waveforms, and we believe it is now necessary to take a further step to introduce more realities and complexities to the mock data. The purpose of Taiji data challenge II (TDC II) is to discover and address the "new" challenges [TDC paper].

TDC II includes 5 groups of datasets, each designed to manifest specific challenge topics. All the datasets are stored in HDF5 (`.h5`) files, whose download link can be found in TDC II website [9], Within the HDF5 files, all the data are organized under the attribute "`eta`" (for dataset groups 1 and 2) or attributes "`interspacecraft_interferometer_carrier`", *etc* (for dataset groups 3 and 4), with the corresponding sampling times stored under "`time`".

Accompanying TDC II, an open-source toolkit "`Triangle`" is also released, which offers the instructions for accessing TDC II data, and enables the injection of customized GW signals and noises, therefore users may have the opportunity to uncover other potential challenges and scientific prospects for Taiji. The whole `Triangle` toolkit consists 3 code repositories:

- `Triangle-Simulator`: time-domain prototype simulator for the data of space-based GW detectors, which is the code utilized for the creation of TDC II. `Triangle-Simulator` encapsulates the simulation of GW responses, noises, instrumental effects (*e.g.* clock deviations), TDI and other pre-processing steps in a unified pipeline. Tutorials are included to introduce the related concepts, models, as well as how to access TDC II data.

- `Triangle-GB`: frequency-domain 2nd-generation TDI response calculator for GBs, adapted from the `GBGPU` [10–12] (the GPU version of `FastGB` [13]) to support the numerical orbit interface of `Triangle-Simulator` and 2nd-generation TDI. Examples are provided showcasing simple Bayesian analysis on the TDC II verification dataset 0.1.

- `Triangle-BBH`: frequency-domain 2nd-generation TDI response calculator for MBHBs. The responses are modeled based on the same numerical orbit interface as `Triangle-Simulator`. Current supported waveform approximants are IMRPhenomD/HM implemented in `WF4PY` [14–16] (CPU) and `BBHx` [17–19] (GPU). Examples based on verification dataset 0.2 is offered.

With the main features, targets and data components (signals, noises, artifacts) of TDC II data introduced in Ref. [TDC paper], in the following we provide detailed descriptions on the models and mathematical formalism for these components, as well as their code implementations in `Triangle-Simulator`. More ready-to-use instructions can be found in the tutorials of the code.

II. SOURCES & WAVEFORMS

A. Massive black hole binaries

The parametrization of MBHB is shown in TABLE I. Note that this is the parametrization of IMRPhenomD and IMRPhenomT, while for SEOBNRv5EHM, φ_c should be replaced by the reference phase $\varphi_{\rm ref}$ [rad], and there are 2 additional parameters: reference frequency $f_{\rm ref}$ [Hz] and eccentricity e [1]. For time-domain simulation, we employed the IMRPhenomD and IMRPhenomT approximants implemented in `PyCBC` [20] and SEOBNRv5EHM in `pySEOBNR` [21].

TABLE I. The parametrization of MBHB.

parameter	description	unit
$\mathcal{M}_{c,z}$	chirp mass (redshifted)	$M_\odot$
q	mass ratio	1
χ_{z1}	spin of black hole 1	1
χ_{z2}	spin of black hole 2	1
t_c	time of coalescence	day
φ_c	phase at coalescence	rad
D_L	luminosity distance	Mpc
ι	inclination angle	rad
λ	Ecliptic longitude	rad
β	Ecliptic latitude	rad
ψ	polarization angle	rad

B. Extreme mass-ratio inspirals

The parametrization of EMRI (AK model) is shown in TABLE II. The sky location parameters are related to spherical coordinates $\{\theta_S, \phi_S\}$ as $\lambda = \phi_S$ and $\beta = \pi/2 - \theta_S$. The parametrizations of other EMRI waveforms differ from those of AK and can all be found in the data files.

TABLE II. The parametrization of EMRI (AK model).

parameter	description	unit
μ	mass of compact object	$M_\odot$
M	mass of MBH	$M_\odot$
Λ	lambda angle	rad
S	spin of the MBH	1
e_0	initial eccentricity	1
ν_0	initial azimuthal orbital frequency	Hz
β	Ecliptic latitude of source	rad
λ	Ecliptic longitude of source	rad
θ_K	polar angle of spin	rad
ϕ_K	azimuthal angle of spin	rad
Φ_0	initial azimuthal orbital phase	rad
$\tilde{\gamma}_0$	initial tidal gamma	rad
α_0	initial alpha angle	rad
D	distance	Gpc

C. Galactic binaries

The parametrization of GB is shown in TABLE III.

In TDC II, the amplitude of GB is defined as

$$A = \frac{2(G\mathcal{M}_c)^{5/3}(\pi f_0)^{2/3}}{c^4 D}, \quad (1)$$

TABLE III. The parametrization of GB.

parameter	description	unit
A	amplitude	1
f_0	initial GW frequency	Hz
$\dot{f}_0$	initial derivative of GW frequency	Hz/s
φ_0	initial GW phase	rad
ι	inclination angle	rad
λ	Ecliptic longitude	rad
β	Ecliptic latitude	rad
ψ	polarization angle	rad

and our convention for the the source-frame (denoted by superscript "S") polarizations are

$$h_+^S(t) = A(1+\cos^2\iota)\cos\left[\varphi\left(t\right)\right], \quad h_\times^S(t) = 2A\cos\iota\sin\left[\varphi\left(t\right)\right], \tag{2}$$

where the phase of GW takes a Taylor expansion form:

$$\varphi(t) = 2\pi\left(f_0 t + \frac{1}{2}\dot{f}_0 t^2 + \frac{1}{6}\ddot{f}_0 t^3\right) + \varphi_0, \tag{3}$$

with

$$\ddot{f}_0 = \frac{11}{3}\frac{\dot{f}_0^2}{f_0}. \tag{4}$$

D. Stochastic gravitational wave background

The parametrization of SGWB varies for different models. For astrophysical SGWB, we adopt a power law model:

$$\Omega_{\mathrm{GW}}(f) = A_{\mathrm{astro}}\left(\frac{f}{f_{\mathrm{astro}}}\right)^{\gamma_{\mathrm{astro}}}, \tag{5}$$

where we fix f_{astro} at 1 mHz, so that the parameter space is 2-D: $\{A_{\mathrm{astro}}, \gamma_{\mathrm{astro}}\}$. While for cosmological SGWB, a double broken power law model is adopted:

$$\Omega_{\mathrm{GW}}(f) = A_{\mathrm{pt}} s^9 \left(\frac{1+r_b^4}{r_b^4+s^4}\right)^{(9-b)/4}\left(\frac{b+4}{b+4-m+ms^2}\right)^{(b+4)/2}, \tag{6}$$

where

$$m = \frac{9r_b^4+b}{r_b^4+1}, \quad s = \frac{f}{f_{\mathrm{pt}}}. \tag{7}$$

Therefore it has a 4-D parameter space: $\{A_{\mathrm{pt}}, f_{\mathrm{pt}}, r_b, b\}$.

III. MOCK DATA GENERATION

Taiji's data flow incorporates sophisticated in-orbit measurements and on-ground processing steps. In order to simulate the performances of instruments and the characteristics of data within a reasonable timescale, we currently focus on the key information propagated through the system, rather than implementing a full physical simulation. Given that Taiji's scientific data mainly originate from laser interferometric measurements, a brief workflow for the simulation of TDC II data is presented in FIG. 2.

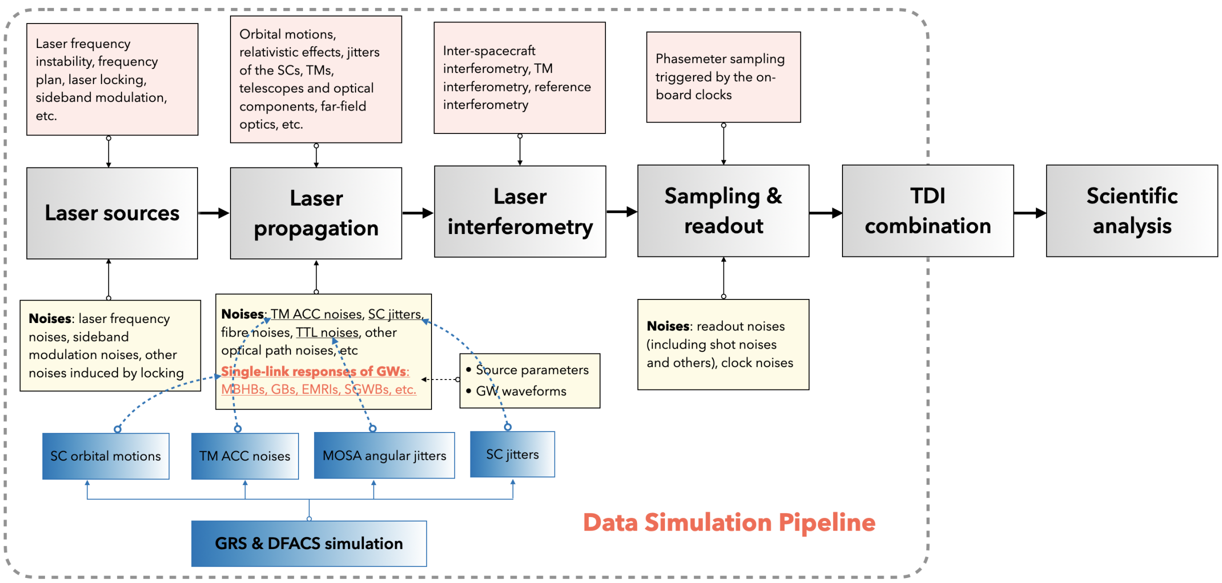

FIG. 2. A brief workflow for the simulation of TDC II data.

A. Laser interferometric measurements

The models of interferometric measurements basically follows Ref. [22], and we make adaptions according to the conventions of TDC II. To begin with, we introduce the notations for two operators, namely the "delay" operator and the "Doppler delay" operator:

$$\mathbf{D}_{ij} f(t) \equiv f\left[t - d_{ij}(t)\right], \quad \dot{\mathbf{D}}_{ij} f(t) \equiv \left[1 - \dot{d}_{ij}(t)\right] \times f\left[t - d_{ij}(t)\right], \tag{8}$$

where $d_{ij}(t)$ stands for the light travel time (LTT) from SC$_j$ to SC$_i$ at the reception time t. The delay operator $\mathbf{D}_{ij}$ is commonly present in the expressions of interferometric measurements and TDI in the early literature. While it was found by Ref. [23] that when

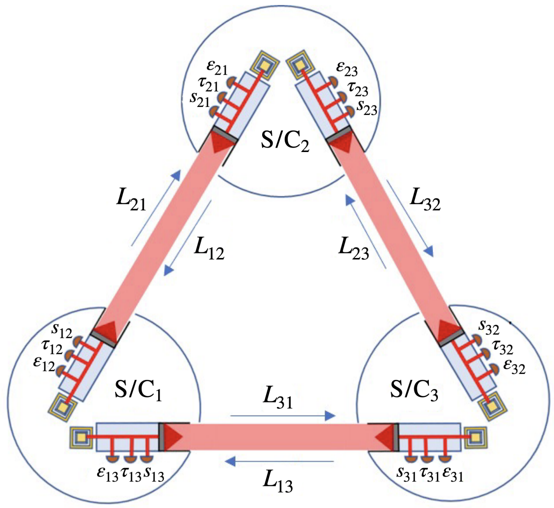

FIG. 3. A schematic of Taiji's detector constellation. Credit: Ref. [26].

interferometric measurements are expressed in units of frequency, the delay operators $\mathbf{D}_{ij}$ should be replaced by their "Doppler" version $\dot{\mathbf{D}}_{ij}$. Note that the difference between $\mathbf{D}_{ij}$ and $\dot{\mathbf{D}}_{ij}$ is only considerable when acted on laser frequency noises, and in the modeling of GW signals and secondary noises, one may safely replace $\dot{\mathbf{D}}_{ij}$ by the original $\mathbf{D}_{ij}$. For brevity we also define the multiple delay operator $\mathbf{D}_{i_1 i_2 \ldots i_{n-1} i_n} f(t)$ as $f(t)$ acted upon by $\mathbf{D}_{i_{n-1} i_n}$, $\mathbf{D}_{i_{n-2} i_{n-1}}$ ... successively.

In our simulation the LTTs are calculated to the 1st post-Newtonian (PN) order, which takes into consideration the motions of SCs as well as the relativistic corrections due to the gravity of the Sun [24]:

$$d_{ij}(t) = d_{ij}^{0\mathrm{PN}}(t) + d_{ij}^{0.5\mathrm{PN}}(t) + d_{ij}^{1\mathrm{PN}}(t). \tag{9}$$

TDI processing typically requires the LTTs to be measured to the nanosecond accuracy [25]. In above equation, $d_{ij}^{1\mathrm{PN}}(t)$ is usually several hundreds of nanoseconds, and the PN coefficient of Sun-detector system is at the 10^{-8} order, so the next half PN correction is less than 10^{-9} s and can be safely neglected.

As is schematically shown in FIG. 3, there are two movable optical sub-assemblies (MOSAs) symmetrically installed on each SC. MOSA is a structure composed of an optical bench (OB), a laser source, a telescope, and a gravitaional reference sensor hosting the test-mass (TM). Each MOSA is labeled by ij ($ij \in \{12, 23, 31, 21, 32, 13\}$), with i being the index of SC carrying this MOSA, and j the index of distant SC that transmits lasers with this MOSA. All the interferometric measurements used for GW detection are taken on the OBs. Specifically, the ISI mixes the local beam with the distant beam (coming from the distant MOSA); the TMI mixes the local and adjacent beams, after it has bounced on the local TM; and the RFI mixes the local and adjacent beams without interaction with the TM. Moreover, for the purposes of clock noise reduction and inter-spacecraft ranging, *etc*, "sidebands" are created by modulating clock signals to the lasers, resulting in an additional sideband interferometric measurement for each interferometer. We label the "original" interferometry with subscript "c" ("c" for carrier), and the sideband interferometry with "sb" ("sb" for sideband). Note that this might only be a simplified model, and the actual designs for payloads are still under consideration.

The results of laser interferometry are read out by the phasemeters, in terms of the instantaneous frequencies of interfered lasers. Therefore we simulate all the raw measurements in the frequency (Hz) unit, and each data stream is a time series uniformly spaced in time. For the sake of understanding and numerical simulation, the laser interferometric data are usually regarded as the sum of two parts [27]. One is a $\sim$ MHz order slow-varying "offset" part, and the other is a $\sim$ Hz order jittering "fluctuation" part. The former includes the effects of laser locking, the frequency plans, the Doppler effects due to orbital motions, *etc*, and the latter is the combination of various noises and GW signals. Note that this is only an artificial division and in realistic detection we only have access to the sum of them. The separation of these two parts is only possible after a "detrending" process. We model laser interferometry in this "two-variable decomposition" manner, denoting the "offset" parts as uppercase letters, and the "fluctuation" parts as lowercase letters. The carrier measurement of ISI reads:

$$\mathrm{ISI}_{c,ij} = S_{c,ij} + s_{c,ij}, \tag{10}$$

where the "offset" part is

$$S_{c,ij} = \dot{\mathbf{D}}_{ij} O^p_{ji} - O^p_{ij} - \nu_0 \dot{d}_{ij}, \tag{11}$$

and the "fluctuation" part is

$$\begin{aligned} s_{c,ij} = \dot{\mathbf{D}}_{ij} p_{ji} - p_{ij} + N^{\mathrm{ro}}_{s_{c,ij}} + \left(\nu_0 + \mathbf{D}_{ij} O^p_{ji}\right) \left(\dot{\mathbf{D}}_{ij} \Delta_{ji} + \Delta_{ij}\right) \\ - \left(\nu_0 + \mathbf{D}_{ij} O^p_{ji}\right) N^{\mathrm{op}}_{s_{ij} \leftarrow ji} + \left(\nu_0 + O^p_{ij}\right) N^{\mathrm{op}}_{s_{ij} \leftarrow ij} + \left(\nu_0 + \mathbf{D}_{ij} O^p_{ji}\right) y_{ij}. \end{aligned} \tag{12}$$

The sideband measurement of ISI reads

$$\mathrm{ISI}_{sb,ij} = S_{sb,ij} + s_{sb,ij}, \tag{13}$$

where

$$S_{sb,ij} = S_{c,ij} + \dot{\mathbf{D}}_{ij}\left[\nu^{\mathrm{m}}_{ji}\left(1+\dot{O}^{q}_{j}\right)\right] - \nu^{\mathrm{m}}_{ij}\left(1+\dot{O}^{q}_{i}\right), \tag{14}$$

$$\begin{aligned} s_{sb,ij} = {} & \dot{\mathbf{D}}_{ij}p_{ji} - p_{ij} + N^{\mathrm{ro}}_{s_{sb,ij}} + \left(\nu_0 + \mathbf{D}_{ij}O^{p,sb}_{ji}\right)\left(\dot{\mathbf{D}}_{ij}\Delta_{ji} + \Delta_{ij}\right) \\ & - \left(\nu_0 + \mathbf{D}_{ij}O^{p,sb}_{ji}\right)N^{\mathrm{op}}_{s_{ij}\leftarrow ji} + \left(\nu_0 + O^{p,sb}_{ij}\right)N^{\mathrm{op}}_{s_{ij}\leftarrow ij} + \left(\nu_0 + \mathbf{D}_{ij}O^{p,sb}_{ji}\right)y_{ij} \\ & + \nu^{\mathrm{m}}_{ji}\dot{\mathbf{D}}_{ij}q_j - \nu^{\mathrm{m}}_{ij}q_i + \nu^{\mathrm{m}}_{ji}\dot{\mathbf{D}}_{ij}N^{\mathrm{m}}_{ji} - \nu^{\mathrm{m}}_{ij}N^{\mathrm{m}}_{ij}. \end{aligned} \tag{15}$$

Similarly, for the carrier of RFI:

$$\mathrm{RFI}_{c,ij} = T_{c,ij} + \tau_{c,ij}, \tag{16}$$

where

$$T_{c,ij} = O^{p}_{ik} - O^{p}_{ij}, \tag{17}$$

$$\tau_{c,ij} = p_{ik} - p_{ij} + N^{\mathrm{ro}}_{\tau_{c,ij}} - \left(\nu_0 + O^{p}_{ik}\right)\mu_{ij} - \left(\nu_0 + O^{p}_{ik}\right)N^{\mathrm{op}}_{\tau_{ij}\leftarrow ik} + \left(\nu_0 + O^{p}_{ij}\right)N^{\mathrm{op}}_{\tau_{ij}\leftarrow ij}. \tag{18}$$

For the sideband of RFI:

$$\mathrm{RFI}_{sb,ij} = T_{sb,ij} + \tau_{sb,ij}, \tag{19}$$

where

$$T_{sb,ij} = T_{c,ij} + \left(\nu^{\mathrm{m}}_{ik} - \nu^{\mathrm{m}}_{ij}\right)\left(1+\dot{O}^{q}_{i}\right), \tag{20}$$

$$\begin{aligned} \tau_{sb,ij} = {} & p_{ik} - p_{ij} + N^{\mathrm{ro}}_{\tau_{sb,ij}} - \left(\nu_0 + O^{p,sb}_{ik}\right)\mu_{ij} - \left(\nu_0 + O^{p,sb}_{ik}\right)N^{\mathrm{op}}_{\tau_{ij}\leftarrow ik} + \left(\nu_0 + O^{p,sb}_{ij}\right)N^{\mathrm{op}}_{\tau_{ij}\leftarrow ij} \\ & + \left(\nu^{\mathrm{m}}_{ik} - \nu^{\mathrm{m}}_{ij}\right)q_i + \nu^{\mathrm{m}}_{ik}N^{\mathrm{m}}_{ik} - \nu^{\mathrm{m}}_{ij}N^{\mathrm{m}}_{ij}. \end{aligned} \tag{21}$$

For TMIs:

$$\mathrm{TMI}_{c,ij} = E_{c,ij} + \varepsilon_{c,ij}, \tag{22}$$

where

$$E_{c,ij} = O^{p}_{ik} - O^{p}_{ij}, \tag{23}$$

$$\varepsilon_{c,ij} = p_{ik} - p_{ij} + N^{\mathrm{ro}}_{\varepsilon_{c,ij}} - 2\left(\nu_0 + O^p_{ij}\right)\delta_{ij} + 2\left(\nu_0 + O^p_{ij}\right)\Delta_{ij} - \left(\nu_0 + O^p_{ik}\right)\mu_{ij}$$
$$- \left(\nu_0 + O^p_{ik}\right) N^{\mathrm{op}}_{\varepsilon_{ij}\leftarrow ik} + \left(\nu_0 + O^p_{ij}\right) N^{\mathrm{op}}_{\varepsilon_{ij}\leftarrow ij}. \tag{24}$$

We do not simulate the sideband of TMI since it's information is duplicated with that of RFI (in the sense of sideband - carrier) In above equations $\nu_0 = 281.6$ THz is the central frequency of lasers, and $\nu^{\mathrm{m}}_{ij} = 2.4/2.401$ GHz for left ($ij \in \{12, 23, 31\}$) /right ($ij \in \{21, 32, 13\}$) MOSAs is the modulation frequency of sidebands. Most importantly, y_{ij} stands for the fractional frequency difference caused by incident GWs. The meanings of other terms can be found in TABLE IV and TABLE V. Notice that above equations are implicitly expressed in a global time frame (Barycenter Coordinate Time of the Solar system, or TCB in short), and the conversion among different time frames (due to clock deviations and relativistic corrections) is not explicitly incorporated. Regarding this part, our treatments are in alignment with Ref. [22].

See Ref. [TDC paper] for the specific noise types injected into each dataset of TDC II. In TDC II, the TM acceleration noises, OB displacement noises (also termed "SC jitters" in Ref. [TDC paper]), and the angular jitters of MOSAs are generated based on a 60-degree-of-freedom numerical simulation of the drag-free & altitude control system (DFACS). The MOSAs' angular jitters are further used to simulate the inter-spacecraft tilt-to-length (TTL) noises via a simple linear coupling model [28]. Other noises are generated as Gaussian stationary noises according to their corresponding designed power spectral density (PSD) models. For these PSDs, both the amplitudes and spectral shapes can be found in the source codes of `Triangle-Simulator` (`Constants.py` and `Noise.py`). Therefore, it seems that we are not so "blind" about these noises. However, we highly recommend the users to stay agnostic about these noise models, since this is the situation we are very likely to encounter in the future. Meanwhile, the knowledge we might possess is the transfer function of each instrumental noise, which could be crucial for the data-driven characterizations of noise spectra and correlations.

Models in this subsection are implemented via the `Triangle-Simulator.Interferometer.Interferometers` class. Notice that for dataset groups 0, 1 and 2 (*i.e.* the "scientific" datasets), we assume that all the data are perfectly detrended and synchronized to TCB. Therefore, in the simulation outputs, the offset terms (the upper case T, S, E) are elimi-

TABLE IV. The "offset" terms.

offset type	symbol	unit
clock drift	O_i^q	time [s]
carrier laser frequency offset	O_{ij}^p	frequency [Hz]
sideband laser frequency offset	$O_{ij}^{p,sb} = O_{ij}^p + \nu_{ij}^{\mathrm{m}} \left(1 + \dot{O}_i^q\right)$	frequency [Hz]

nated, leaving only the fluctuation terms (the lower case s, τ, ε), and the clock drifts along with clock noises should be neglected. In this case the sideband measurements also become unnecessary.

TABLE V. Symbols and alias (in `Triangle`) for the noise terms. The instrumental noises are classified into different types mainly based on the form in which they enter the interferometric measurements, rather than their physical origins.

noise type	symbol	alias	unit
laser noise	p_{ij}	`laser_noise`	frequency [Hz]
readout noise of IFO	$N^{\mathrm{ro}}_{\mathrm{IFO}_{ij}}$	`ro_ifo_noise`	frequency [Hz]
fibre back link noise	μ_{ij}	`bl_noise`	optical path derivative [s/s]
optical path noise	$N^{\mathrm{op}}_{\mathrm{IFO}_{ij\leftarrow kl}}$	`op_ifo_noise`	optical path derivative [s/s]
TM acceleration noise	δ_{ij}	`acc_noise`	optical path derivative [s/s]
OB displacement noise	Δ_{ij}	`ob_noise`	optical path derivative [s/s]
clock noise	q_i	`clock_noise`	time derivative [s/s]
sideband modulation noise	N^{m}_{ij}	`modulation_noise`	time derivative [s/s]
pseudo ranging noise	N^{R}_{ij}	`ranging_noise`	time [s]

B. Gravitational wave response

The single-link response of GW signal is derived under the usual conventions adopted in the literature [29–31], which is also in consistency with the response model of TDC I [5]:

$$\begin{aligned} y_{ij}(t) &\equiv \frac{\nu_{\text{receive}} - \nu_{\text{send}}}{\nu_{\text{send}}} \\ &\approx \frac{1}{2\left(1 - \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{n}}_{ij}(t_i)\right)} \left[H_{ij}\left(t - \frac{d_{ij}(t)}{c} - \frac{\hat{\boldsymbol{k}} \cdot \boldsymbol{R}_j(t)}{c}\right) - H_{ij}\left(t - \frac{\hat{\boldsymbol{k}} \cdot \boldsymbol{R}_i(t)}{c}\right)\right], \end{aligned} \tag{25}$$

where the projection of GW on arm ij is defined as

$$H_{ij} \equiv \boldsymbol{h} : \hat{\boldsymbol{n}}_{ij} \otimes \hat{\boldsymbol{n}}_{ij}, \tag{26}$$

with $\hat{\boldsymbol{k}}$, $\boldsymbol{R}_i$, $\hat{\boldsymbol{n}}_{ij}$ being the wave vector, the position of SC_i in the Solar system barycenter (SSB) frame, and the arm vector. In terms of Ecliptic longitude and latitude, the coordinate of $\hat{\boldsymbol{k}}$ is

$$\hat{\boldsymbol{k}} = [-\cos\beta\cos\lambda, -\cos\beta\sin\lambda, -\sin\beta]\,. \tag{27}$$

The GW tensor $\boldsymbol{h}$ reads

$$\boldsymbol{h} = h_+ \boldsymbol{e}_+ + h_\times \boldsymbol{e}_\times. \tag{28}$$

We transform the GW waveform from the source frame (denoted with superscript "S") to the SSB frame via

$$h_+ = \cos 2\psi \ h_+^{\mathrm{S}} - \sin 2\psi \ h_\times^{\mathrm{S}}, \quad h_\times = \sin 2\psi \ h_+^{\mathrm{S}} + \cos 2\psi \ h_\times^{\mathrm{S}}, \tag{29}$$

where

$$\boldsymbol{e}_+ = \boldsymbol{u} \otimes \boldsymbol{u} - \boldsymbol{v} \otimes \boldsymbol{v}, \quad \boldsymbol{e}_\times = \boldsymbol{u} \otimes \boldsymbol{v} + \boldsymbol{v} \otimes \boldsymbol{u} \tag{30}$$

are the polarization basis tensors, which are associated to vectors $\boldsymbol{u}$, $\boldsymbol{v}$:

$$\boldsymbol{u} = [\sin\lambda, -\cos\lambda, 0], \quad \boldsymbol{v} = [-\sin\beta\cos\lambda, -\sin\beta\sin\lambda, \cos\beta]. \tag{31}$$

`Triangle-Simulator` provides multiple (equivalent) implementations for the GW response function. For dataset groups 0, 1 and 2 (scientific datasets), due to the large number of GW signals to be calculated, we utilize the `Triangle-Simulator.GW. GeneralTDIResponse` class (setting `return_eta=True` and `Pstrings=eta_string` since single-link responses are desired). While for dataset groups 3 and 4 (raw datasets), the `SimulateGW` function integrated to the `Interferometers` class is adopted, since it accounts for the realistic scenarios such as signal sampling error due to clock deviations.

C. Data anomalies

Glitches and gaps are two typical data anomalies that may be present in realistic Taiji data. LISA Pathfinder and Taiji-1 demonstrated that the TMs could be affected by unexpected transient disturbances. Thus it is conservative to expect that these glitches would take place during the science operations of Taiji. According to the observations of LISA Pathfinder and Taiji-1, we model the glitches with a double decaying exponential form in terms of TM acceleration [32, 33]:

$$\delta^{\text{glitch}}_{\text{acceleration}}(t) = \frac{\Delta v}{\tau_1 - \tau_2}\left(e^{-\frac{t-t_0}{\tau_1}} - e^{-\frac{t-t_0}{\tau_2}}\right)\Theta(t-t_0), \tag{32}$$

where t_0 is the time of injection, Δv is the velocity gain caused by the glitch, $\tau_{1,2}$ are the time scales of the exponentials, and $\Theta(t)$ stands for the Heaviside function. In Datasets 4, glitches of this form are injected at random times and random TMs. On the other hand, we model data gaps originating from scheduled maintenances (e.g. re-pointing of the antennas, switching of frequency plans, and in-orbit maneuvers, *etc*) as a 7-hour duration with no data collection (set to `numpy.nan`) starting at the 15th day of Datasets 4.

D. Time-delay interferometry

The complete TDI processing includes 3 steps. Firstly, the intermediate variable ξ_{ij} is constructed, which is free of OB displacement noises:

$$\xi_{ij} = s_{ij} + \frac{\tau_{ij} - \varepsilon_{ij}}{2} + \dot{\mathbf{D}}_{ij}\frac{\tau_{ji} - \varepsilon_{ji}}{2}, \tag{33}$$

where $ij \in \{12, 23, 31, 21, 32, 13\}$.

Secondly, we construct the intermediate variable η_{ij} to reduce half of the laser noises:

$$\eta_{ij} = \xi_{ij} + \dot{\mathbf{D}}_{ij}\frac{\tau_{ji} - \tau_{jk}}{2}, \quad \eta_{ik} = \xi_{ik} + \frac{\tau_{ij} - \tau_{ik}}{2}, \tag{34}$$

with $ijk \in \{123, 231, 312\}$. To convert to the fractional frequency difference unit, we further divide all the η_{ij} by the central frequency ν_0. These are just the data streams that users have access to for Datasets 0, 1 and 2. For a signal-only simulation, one has $\eta_{ij} = y_{ij}$. While for the 2-component noise model usually employed in simplified investigations,

$$\eta_{ij} = y_{ij} + n_{ij}, \quad n_{ij} = N_{ij} + \delta_{ij} + \dot{\mathbf{D}}_{ij}\delta_{ji}, \tag{35}$$

with N_{ij} and δ_{ij} being the optical metrology system (OMS) noise and TM acceleration (ACC) noise, respectively.

The third step is constructing TDI variables to mitigate all the remaining laser noises. This step varies for different TDI schemes. Despite that there are hundreds of TDI schemes in the literature, they can all be abstracted into a unified form:

$$\text{TDI} \ = \sum_{ij} \mathbf{P}_{ij}\eta_{ij}. \tag{36}$$

Taking the second-generation Michelson channel X_2 as an example, the fiducial rule of X_2 channel used by `Triangle` is

$$\begin{aligned} X_2 = & \left(1 - \dot{\mathbf{D}}_{131} - \dot{\mathbf{D}}_{13121} + \dot{\mathbf{D}}_{1213131}\right)\left(\eta_{12} + \dot{\mathbf{D}}_{12}\eta_{21}\right) \\ & - \left(1 - \dot{\mathbf{D}}_{121} - \dot{\mathbf{D}}_{12131} + \dot{\mathbf{D}}_{1312121}\right)\left(\eta_{13} + \dot{\mathbf{D}}_{13}\eta_{31}\right), \end{aligned} \tag{37}$$

which can be expressed in terms of $\mathbf{P}_{ij}$ operators as

$$\begin{aligned} \mathbf{P}_{12} &= 1 - \dot{\mathbf{D}}_{131} - \dot{\mathbf{D}}_{13121} + \dot{\mathbf{D}}_{1213131}, \\ \mathbf{P}_{23} &= 0, \\ \mathbf{P}_{31} &= -\dot{\mathbf{D}}_{13} + \dot{\mathbf{D}}_{1213} + \dot{\mathbf{D}}_{121313} - \dot{\mathbf{D}}_{13121213}, \\ \mathbf{P}_{21} &= \dot{\mathbf{D}}_{12} - \dot{\mathbf{D}}_{1312} - \dot{\mathbf{D}}_{131212} + \dot{\mathbf{D}}_{12131312}, \\ \mathbf{P}_{32} &= 0, \\ \mathbf{P}_{13} &= -1 + \dot{\mathbf{D}}_{121} + \dot{\mathbf{D}}_{12131} - \dot{\mathbf{D}}_{1312121}. \end{aligned} \tag{38}$$

The expressions for Y_2 and Z_2 channels can be obtained using the permutation rule $1 \to 2, 2 \to 3, 3 \to 1$. The conventions for Michelson channels might differ by a minus sign in other works. The `Triangle-Simulator.TDI.TDI` class is capable of combining any TDI variable, given the strings representing the $\mathbf{P}_{ij}$ operators.

In the presence of clock noises, an additional step should be performed to calculate and subtract the clock noise correction term [34], which doesn't affect the signal content of data.

[1] Z. Luo, Y. Wang, Y. Wu, W. Hu, and G. Jin, The taiji program: A concise overview, Progress of Theoretical and Experimental Physics **2021**, 05A108 (2020), https://academic.oup.com/ptep/article-pdf/2021/5/05A108/37953044/ptaa083.pdf.

[2] Z. Luo, Z. Guo, G. Jin, Y. Wu, and W. Hu, A brief analysis to taiji: Science and technology, Results in Physics **16**, 102918 (2020).

[3] K. A. Arnaud *et al.*, The Mock LISA Data Challenges: An overview, AIP Conf. Proc. **873**, 619 (2006), arXiv:gr-qc/0609105.

[4] Q. Baghi (LDC Working Group), The LISA Data Challenges, in *56th Rencontres de Moriond on Gravitation* (2022) arXiv:2204.12142 [gr-qc].

[5] Z. Ren, T. Zhao, Z. Cao, Z.-K. Guo, W.-B. Han, H.-B. Jin, and Y.-L. Wu, Taiji data challenge for exploring gravitational wave universe, Front. Phys. (Beijing) **18**, 64302 (2023), arXiv:2301.02967 [gr-qc].

[6] T. B. Littenberg and N. J. Cornish, Prototype global analysis of LISA data with multiple source types, Phys. Rev. D **107**, 063004 (2023), arXiv:2301.03673 [gr-qc].

[7] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler, and D. Giardini, Global analysis of LISA data with Galactic binaries and massive black hole binaries, Phys. Rev. D **110**, 024005 (2024), arXiv:2403.15318 [gr-qc].

[8] M. L. Katz, N. Karnesis, N. Korsakova, J. R. Gair, and N. Stergioulas, Efficient GPU-accelerated multisource global fit pipeline for LISA data analysis, Phys. Rev. D (2024), arXiv:2405.04690 [gr-qc].

[9] TDC Working Group, TDC II website, `http://gr.imech.ac.cn/overview/`.

[10] M. L. Katz, mikekatz04/GBGPU: First official public release (2022).

[11] M. L. Katz, C. Danielski, N. Karnesis, V. Korol, N. Tamanini, N. J. Cornish, and T. B. Littenberg, Bayesian characterization of circumbinary sub-stellar objects with LISA, Mon. Not. Roy. Astron. Soc. **517**, 697 (2022), arXiv:2205.03461 [astro-ph.EP].

[12] T. Robson, N. J. Cornish, N. Tamanini, and S. Toonen, Detecting hierarchical stellar systems with LISA, Phys. Rev. D **98**, 064012 (2018), arXiv:1806.00500 [gr-qc].

[13] N. J. Cornish and T. B. Littenberg, Tests of Bayesian Model Selection Techniques for Gravitational Wave Astronomy, Phys. Rev. D **76**, 083006 (2007), arXiv:0704.1808 [gr-qc].

[14] CosmoStatGW, WF4PY, `https://github.com/CosmoStatGW/WF4Py`.

[15] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, Forecasting the Detection Capabilities of Third-generation Gravitational-wave Detectors Using GWFAST, Astrophys. J. **941**, 208 (2022), arXiv:2207.02771 [gr-qc].

[16] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, GWFAST: A Fisher Information

Matrix Python Code for Third-generation Gravitational-wave Detectors, Astrophys. J. Supp. **263**, 2 (2022), arXiv:2207.06910 [astro-ph.IM].

[17] M. L. Katz, mikekatz04/BBHx: First official public release (2021).

[18] M. L. Katz, S. Marsat, A. J. K. Chua, S. Babak, and S. L. Larson, GPU-accelerated massive black hole binary parameter estimation with LISA, Phys. Rev. D **102**, 023033 (2020), arXiv:2005.01827 [gr-qc].

[19] M. L. Katz, Fully automated end-to-end pipeline for massive black hole binary signal extraction from LISA data, Phys. Rev. D **105**, 044055 (2022), arXiv:2111.01064 [gr-qc].

[20] A. Nitz, I. Harry, D. Brown, C. M. Biwer, J. Willis, T. D. Canton, C. Capano, T. Dent, L. Pekowsky, G. S. C. Davies, S. De, M. Cabero, S. Wu, A. R. Williamson, B. Machenschalk, D. Macleod, F. Pannarale, P. Kumar, S. Reyes, dfinstad, S. Kumar, M. Tápai, L. Singer, P. Kumar, veronica villa, maxtrevor, B. U. V. Gadre, S. Khan, S. Fairhurst, and A. Tolley, gwastro/pycbc: v2.3.3 release of pycbc (2024).

[21] pyseobnr, `https://waveforms.docs.ligo.org/software/pyseobnr/index.html`.

[22] O. Hartwig, *Instrumental modelling and noise reduction algorithms for the Laser Interferometer Space Antenna*, Ph.D. thesis (2021).

[23] J.-B. Bayle, O. Hartwig, and M. Staab, Adapting time-delay interferometry for lisa data in frequency, Phys. Rev. D **104**, 023006 (2021).

[24] B. Chauvineau, T. Regimbau, J.-Y. Vinet, and S. Pireaux, Relativistic analysis of the lisa long range optical links, Phys. Rev. D **72**, 122003 (2005).

[25] M. Tinto and S. V. Dhurandhar, TIME DELAY INTERFEROMETRY, Living Rev. Rel. **8**, 4 (2005), arXiv:gr-qc/0409034.

[26] P. Wu, M. Du, and P. Xu, Suppressing data anomalies of gravitational reference sensors with time delay interferometry combinations, Opt. Express **32**, 43249 (2024), arXiv:2206.10664 [gr-qc].

[27] J.-B. Bayle and O. Hartwig, Unified model for the LISA measurements and instrument simulations, Phys. Rev. D **107**, 083019 (2023), arXiv:2212.05351 [gr-qc].

[28] S. Paczkowski, R. Giusteri, M. Hewitson, N. Karnesis, E. D. Fitzsimons, G. Wanner, and G. Heinzel, Postprocessing subtraction of tilt-to-length noise in lisa, Phys. Rev. D **106**, 042005 (2022).

[29] LDC Working Group, LISA Data Challenge Manual, `https://lisa-ldc.lal.in2p3.fr/`

static/data/pdf/LDC-manual-002.pdf.

[30] M. L. Katz, J.-B. Bayle, A. J. K. Chua, and M. Vallisneri, Assessing the data-analysis impact of lisa orbit approximations using a gpu-accelerated response model, Phys. Rev. D **106**, 103001 (2022).

[31] Q. Baghi, N. Karnesis, J.-B. Bayle, M. Besançon, and H. Inchauspé, Uncovering gravitational-wave backgrounds from noises of unknown shape with LISA, JCAP **04**, 066, arXiv:2302.12573 [gr-qc].

[32] M. Armano, H. Audley, J. Baird, P. Binetruy, M. Born, D. Bortoluzzi, E. Castelli, A. Cavalleri, A. Cesarini, V. Chiavegato, A. M. Cruise, D. Dal Bosco, K. Danzmann, M. De Deus Silva, I. Diepholz, G. Dixon, R. Dolesi, L. Ferraioli, V. Ferroni, E. D. Fitzsimons, M. Freschi, L. Gesa, D. Giardini, F. Gibert, R. Giusteri, C. Grimani, J. Grzymisch, I. Harrison, M. S. Hartig, G. Heinzel, M. Hewitson, D. Hollington, D. Hoyland, M. Hueller, H. Inchauspé, O. Jennrich, P. Jetzer, B. Johlander, N. Karnesis, B. Kaune, N. Korsakova, C. J. Killow, J. A. Lobo, J. P. López-Zaragoza, R. Maarschalkerweerd, D. Mance, V. Martín, L. Martin-Polo, F. Martin-Porqueras, J. Martino, P. W. McNamara, J. Mendes, L. Mendes, N. Meshksar, M. Nofrarias, S. Paczkowski, M. Perreur-Lloyd, A. Petiteau, E. Plagnol, J. Ramos-Castro, J. Reiche, F. Rivas, D. I. Robertson, G. Russano, L. Sala, P. Sarra, J. Slutsky, C. F. Sopuerta, T. Sumner, D. Texier, J. I. Thorpe, D. Vetrugno, S. Vitale, G. Wanner, H. Ward, P. Wass, W. J. Weber, L. Wissel, A. Wittchen, C. Zanoni, and P. Zweifel (LISA Pathfinder Collaboration), Transient acceleration events in lisa pathfinder data: Properties and possible physical origin, Phys. Rev. D **106**, 062001 (2022).

[33] The Taiji Scientific Collaboration, Taiji program in space for gravitational universe with the first run key technologies test in Taiji-1, International Journal of Modern Physics A **36**, 2102002 (2021).

[34] O. Hartwig and J.-B. Bayle, Clock-jitter reduction in lisa time-delay interferometry combinations, Phys. Rev. D **103**, 123027 (2021).

[35] N. J. Cornish and J. Crowder, LISA data analysis using MCMC methods, Phys. Rev. D **72**, 043005 (2005), arXiv:gr-qc/0506059.

[36] J. W. Armstrong, F. B. Estabrook, and M. Tinto, Time-delay interferometry for space-based gravitational wave searches, The Astrophysical Journal **527**, 814 (1999).

[37] M. Muratore, O. Hartwig, D. Vetrugno, S. Vitale, and W. J. Weber, Effectiveness of null

time-delay interferometry channels as instrument noise monitors in LISA, Phys. Rev. D **107**, 082004 (2023), arXiv:2207.02138 [gr-qc].

[38] G. Wang, Time delay interferometry with minimal null frequencies, Phys. Rev. D **110**, 042005 (2024).

[39] V. Korol, N. Hallakoun, S. Toonen, and N. Karnesis, Observationally driven Galactic double white dwarf population for LISA, Mon. Not. Roy. Astron. Soc. **511**, 5936 (2022), arXiv:2109.10972 [astro-ph.HE].

[40] N. Karnesis, S. Babak, M. Pieroni, N. Cornish, and T. Littenberg, Characterization of the stochastic signal originating from compact binary populations as measured by lisa, Phys. Rev. D **104**, 043019 (2021).

[41] M. Tinto, M. Vallisneri, and J. W. Armstrong, Time-delay interferometric ranging for space-borne gravitational-wave detectors, Phys. Rev. D **71**, 041101 (2005).

[42] M. Colpi *et al.*, LISA Definition Study Report, arXiv e-print (2024), arXiv:2402.07571 [astro-ph.CO].

[43] O. Hartwig, J.-B. Bayle, M. Staab, A. Hees, M. Lilley, and P. Wolf, Time-delay interferometry without clock synchronization, Phys. Rev. D **105**, 122008 (2022).

[44] J. Luo, L.-S. Chen, H.-Z. Duan, Y.-G. Gong, S. Hu, J. Ji, Q. Liu, J. Mei, V. Milyukov, M. Sazhin, C.-G. Shao, V. T. Toth, H.-B. Tu, Y. Wang, Y. Wang, H.-C. Yeh, M.-S. Zhan, Y. Zhang, V. Zharov, and Z.-B. Zhou, Tianqin: a space-borne gravitational wave detector, Classical and Quantum Gravity **33**, 035010 (2016).

[45] A. Gamboa, A. Buonanno, R. Enficiaud, M. Khalil, A. Ramos-Buades, L. Pompili, H. Estellés, M. Boyle, L. E. Kidder, H. P. Pfeiffer, H. R. Rüter, and M. A. Scheel, Accurate waveforms for eccentric, aligned-spin binary black holes: The multipolar effective-one-body model seobnrv5ehm, arXiv e-print (2024), arXiv:2412.12823 [gr-qc].