ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202602.00083

Mock Observations for the CSST Mission: HSTDM-Synthetic Data Generation Postprint

Authors: Siyuan Tan, Wenyin Duan, Yilong Zhang, Yiping Ao, Yan Gong, Zhenhui Lin, Xuan Zhang, Yong Shi, Jing Tang, Jing Li, Ruiqing Mao, Sheng-Cai Shi

Date: 2026-01-28T11:17:08+00:00

Abstract

The High Sensitivity Terahertz Detection Module (HSTDM), a key component of the backend system on board the Chinese Space Station Survey Telescope (CSST), will create significant opportunities for discoveries in terahertz astronomy, with implications extending well beyond China to the global astronomical community. To ensure the accuracy and scientific utility of the final archived data products, the raw data collected by HSTDM must undergo rigorous calibration and processing through the HSTDM data processing pipeline (hereafter the HSTDM pipeline). This requires the HSTDM pipeline to correct instrumental artifacts and effects, and to coordinate the data flow arising from scheduled observing sequences across all HSTDM observing modes within the fully automated CSST processing environment.

As understanding of CSST HSTDM data characteristics and reduction strategies evolves during pipeline development, it is essential to evaluate the accuracy, robustness, and performance of the HSTDM pipeline under all HSTDM observing modes, so that individual pipeline components can be rationally added, removed, modified, or extended within a modular framework. In this paper, we develop practical simulation methods to meet this need. The synthetic data generation for HSTDM observations comprises two main parts: (1) simulation of HSTDM instrumental effects based on both real test measurements and theoretical models; and (2) generation of observing data streams based on representative HSTDM observing mode scenarios. These simulation methods have been implemented and demonstrated to be effective for testing and validating the HSTDM pipeline during its development stage.

ChinaRxiv 中X

Full Text

Mock Observations for the CSST Mission: HSTDM-Synthetic Data Generation

Siyuan Tan[12], Wenyin Duan[12], Yilong Zhang[1], Yiping Ao[1], Yan Gong[1], Zhenhui Lin[1], Xuan Zhang[12], Yong Shi[34], Jing Tang[5], Jing Li[1], Ruiqing Mao[1], and Sheng-Cai Shi[1]

[1]Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023, China; scshi@pmo.ac.cn
[2]School of Astronomy and Space Science, University of Science and Technology of China, Hefei 230026, China
[3]School of Astronomy and Space Science, Nanjing University, Nanjing 210093, China
[4]Key Laboratory of Space Astronomy and Technology, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China

Received 2025 March 11; revised 2025 June 12; accepted 2025 August 12; published 2026 January 6

Abstract

The High Sensitivity Terahertz Detection Module (HSTDM), a key component of the backend modules on board the Chinese Space Station Survey Telescope (CSST), will offer great opportunities for discovery in terahertz astronomy, with implications that extend well beyond China to the global astronomical community. It is imperative that the raw data collected by HSTDM undergo meticulous calibration and processing through the HSTDM data processing pipeline (hereafter HSTDM pipeline) to ensure the accuracy and effectiveness of the final science data archived for further research. This process necessitates that the HSTDM pipeline address instrumental artifacts and effects as well as coordinate the data flow of scheduled observing sequences under all observing modes of HSTDM within the CSST automated processing environment. As understanding of CSST HSTDM data processing develops during the pipeline development stage, it becomes essential to assess the accuracy, robustness, and performance of the HSTDM pipeline under all observing modes so that components of the HSTDM pipeline can be rationally added, removed, amended, or extended within the modular framework. In this paper, we develop practical simulation methods to facilitate this need. The contribution of synthetic data generation for HSTDM observation includes two parts: (1) HSTDM instrumental effect simulation based on both real testing profiles and simulated models; (2) Observing data flow generation based on HSTDM observing mode scenarios. The simulation methods have been implemented and shown to be practical for testing the HSTDM pipeline during the development stage.

Key words: telescopes –instrumentation: detectors –methods: data analysis

1. Introduction to HSTDM

Featuring the flagship project of Chinese space astronomy, the Chinese Space Station Survey Telescope (CSST) is expected to be the largest space telescope developed by China in the coming years, with outstanding characteristics such as a large field of view, high image quality, and wide band coverage. Its detection sensitivity and spatial resolution are on par with those of the globally renowned NASA/ESA Hubble Space Telescope (HST), but its field of view and data acquisition capacity will significantly surpass HST. Equipped with an array of precision detection modules, CSST is poised to be highly competitive and is expected to achieve significant breakthroughs in the realms of cosmology, active galactic nuclei, galaxies and stars, astrometry, extrasolar planets, and celestial bodies within our solar system \cite{Cao_{2018}, Gong_{2019}, Zhan_{2021}}.

One of the powerful modules aboard CSST is the High Sensitivity Terahertz Detection Module (HSTDM; \cite{Zhang_{2018}}), whose core components are two superconducting (SIS) mixers (superconductor-insulator-superconductor) operating within a working frequency range of 0.41–0.51 THz (corresponding working wavelength of 590–730 μm), with a spectral resolution less than 100 kHz and a system noise temperature less than 300 K \cite{Yao_{2020}}. HSTDM serves as an excellent exemplar of terahertz technology \cite{Li_{2025}} and is designed for spectral line observation, offering both high spectral resolution and exceptionally high sensitivity. The technology behind spectral line observations is known as heterodyne spectroscopy. In heterodyne spectroscopy, the incident sky signal $\nu_{\rm sky}$ is mixed with a local oscillator (LO) at a tunable frequency $\nu_{\rm LO}$ close to $\nu_{\rm sky}$, and passed onto the nonlinear mixer. The mixer is designed to be very sensitive to the beat frequencies $\nu_{\rm IF} = |\nu_{\rm sky} - \nu_{\rm LO}|$, which are called intermediate frequencies (IFs) that cover a frequency band with significantly lower frequencies than $\nu_{\rm sky}$ but retain the same information as in $\nu_{\rm sky}$. The IF band signals are then amplified and passed onto the spectrometer to obtain the final raw spectrum data for further processing.

The signals emanating from the cosmos in the 0.41–0.51 THz frequency band contain spectral signatures that reveal a range of interstellar atomic, molecular, and atmospheric tracers, including CI, H_2O, O_2, NH_3, CO, CS, SO, and others. There have been several space missions worldwide that carried detection modules targeting a similar frequency band, such as the Odin satellite (486–504 GHz, 541–580 GHz) \cite{Hjalmarson_{2004}}, the Submillimeter Wave Astronomy Satellite (SWAS) (lower frequency band: 487–493 GHz) \cite{Melnick_{2002}}, and Herschel HIFI (Band 1: 480–640 GHz) \cite{Roelfsema_{2012}}. The HSTDM will complement Herschel HIFI in the frequency band of 410–480 GHz, with scientific objectives in two aspects: (1) The evolution of cosmic carbon, mainly to observe the 0.492 THz neutral carbon (CI) line emission from gas clouds in the Milky Way and neighboring galaxies, and to understand the distribution, dynamic characteristics, relationship with the environment, and the process of atom-to-molecule transformation; (2) Molecular spectral line surveys

to obtain chemical composition of different kinds of celestial bodies in star-forming regions of the Milky Way (such as Orion-KL, Sgr B2, IRAS 16293-2422, etc.), late-type stars (such as IRC+10216, etc.), planets, and comets \cite{Ao_{2023}}.

The conversion of HSTDM raw telemetry data to scientifically usable products is performed by a series of standard processing steps that constitute the HSTDM pipeline. This involves the recombination of telemetry segment data to form complete spectrum data, the coordination of the raw data flow of scheduled observing sequences under all observing modes of HSTDM, the efficient correction and removal of instrumental artifacts in the observation data, along with the generation of standard data products of different levels. The HSTDM pipeline and all associated data products were designed and developed by the HSTDM data processing team within the CSST science data processing group. The HSTDM pipeline is currently nearing the end of its development phase and requires comprehensive simulation and testing to assess the accuracy, robustness, and performance of the HSTDM pipeline under all observing modes, which serves as the purpose of this paper. This article is organized as follows: Section 2 reviews the observation modes of HSTDM during nominal observation, which forms the basis for generating synthetic observation data flow of the scheduled observing sequences. Section 3 briefly introduces the concepts and structures of the HSTDM pipeline. Section 4 discusses the simulation method for the HSTDM data processing pipeline at length, with simulation experiments and discussion presented in Section 5, and our conclusions drawn in Section 6.

[Figure 1: see original paper] The illustration of position switch and chop load operation during HSTDM observation. The position switch operation involves telescope slewing to change the actual pointing from the source position (ON) to the reference position (OFF). The chop load operation involves the internal mirror' s adjustment to change the sky path toward the internal load, which is usually a carefully engineered blackbody radiator with high emissivity and known physical temperature. This figure is an adaptation from the HIFI observing mode illustration on pages 33 and 36 in Herschel Space Observatory (2017).

2. Observing Modes

According to the technical report of HSTDM \cite{Ao_{2023}}, HSTDM has two types of observing modes: target mode, in which the telescope observes a fixed point in the sky, and On-The-Fly (OTF) mode, which enables continuous sky scanning over a larger region. To better present these two observing modes in the following subsections, it is important to describe the specific operations that must be performed in both modes. For both observing modes during nominal observation and on-orbit calibration (OOC), there exist elementary operations that HSTDM must conduct: position switch and chop load. As illustrated in Figure 1, position switch refers to telescope slewing between source and reference positions. ON integration can commence when either beam of HSTDM is

directed to the source position. Once the beam is redirected to the OFF position, OFF integration can initiate accordingly. The OFF position should be selected in close proximity to the ON position, yet at an angular distance of no less than representing the full width at half maximum (FWHM) of the Gaussian beam of HSTDM at frequency range of 0.41–0.51 THz \cite{Ao_{2023}}. The chosen OFF position should be either free of emissions within this range or exhibit an already known emission profile, as also outlined in \cite{Ossenkopf_{2005}}. Chop load refers to the internal mirror' s adjustment, which alters the sky path to direct the observation toward the internal load (cold load). This operation is essential for both HSTDM instrumental sensitivity measurement and intensity calibration of the source. Compared with telescope slewing that requires the main mirror to move, chop load is done through internal motor control which takes relatively less time, thereby minimizing observational overhead as much as possible.

2.1. Target Mode

For point source observations, HSTDM offers a dedicated target mode observation. This mode essentially comprises two fundamental procedural sequences: (1) HSTDM performs integration at the source position for a predefined duration, concurrently transferring data to the compressed storage unit of CSST. After an interval determined by the system stability time, the source integration is interrupted, and the system transitions to the reference integration phase via telescope slewing to the designated OFF position. (2) HSTDM assesses the instrumental sensitivity by leveraging the quantifiable disparity of the load measurements between the hot and cold internal loads. The hot and cold loads are usually carefully engineered blackbody radiators with high emissivity and known physical temperature. With respect to the HSTDM load measurements, we refer to those carried out with the blackbody immersed in liquid nitrogen as the cold load. Conversely, the hot load corresponds to measurements taken when the blackbody is placed at room temperature. The method to measure the instrumental sensitivity is the well-known Y-Factor method \cite{Tiemeijer_{2005}}, which can be formulated briefly as follows:

$$Y = \frac{P_h}{P_c}, \quad T_{\text{sys}} = \frac{T_h - YT_c}{Y - 1}$$

where P_h and P_c are the channel outputs of HSTDM at hot load and cold load, respectively. T_h and T_c denote the thermodynamic temperature of the blackbody at hot load and cold load, respectively. Y represents the Y-factor, which is determined by the measurements under both hot and cold load conditions. Additionally, T_{sys} is the system noise temperature, which also characterizes the instrumental sensitivity of the whole system.

This load calibration should be strategically scheduled during the telescope' s slew to the OFF position, along with other operations by leveraging an obser-

vation scheduling optimization method like that in \cite{Tan_{2024}}. These two foundational sequences are organized in a time series manner. A typical timeline of this mode is illustrated in Figure 2 [Figure 2: see original paper]. As shown in this figure, the timeline consists of telescope slews from the initial position to the science target, integrations at this position, a subsequent slew to a user-designated OFF position, and integrations at the OFF position. The durations of integrations at both positions are chosen to be the same and the sequence of pointing follows an ON–OFF patterned cycle. Load calibration measurements are interspersed during the telescope slews between the source and OFF positions, although it is not mandatory to execute load calibration after each slew. In instances where the instrumental configuration differs from the previous setup, an instrument reset is required during the initial slew to the target source position, followed by a load calibration procedure. There are five parameters that characterize this timeline sequence, as listed in Table 1 .

2.2. On-The-Fly Mode

For observations of extended sources, HSTDM utilizes the OTF observation technique. OTF is an observing technique in which the telescope is driven smoothly and rapidly across a region of sky, or "field," while data and antenna position information are recorded continuously \cite{Mangum_{2007}}. The schematic view of typical OTF observation and its timeline sketch are presented in Figures 3 and 4 respectively.

As depicted in Figure 3 [Figure 3: see original paper], integration of the source occurs and data dumps are captured during the telescope' s scan of a specific row within the map grid. Following each row, the telescope reverses direction to initiate the scan of the subsequent row. Integration and data dumping are paused during these directional changes. After a duration set by the system' s stability requirements, the mapping process is temporarily halted for reference measurements at the OFF position. Load calibration is commonly executed during the telescope's slew to the OFF position as illustrated in Figure 4 [Figure 4: see original paper].

There are six parameters that characterize this timeline sequence, as listed in Table 2 .

Typically, the mapped area is observed through multiple coverages, accumulating the total integration time per source to meet the necessary requirements. The rms noise of switched measurements is given by:

$$\sigma = \frac{T_{\mathrm{sys}}}{\sqrt{\Delta\nu \cdot t_{\mathrm{on}} \cdot \eta_{\mathrm{spec}}}} \sqrt{1 + \frac{t_{\mathrm{on}}}{t_{\mathrm{off}}}}$$

where T_{sys} denotes the system noise temperature, $\Delta\nu$ is the spectral resolution of the measurement, t_{on} denotes the source measurement period, t_{off} denotes the OFF measurement period, and η_{spec} is the spectrometer efficiency.

The scanning velocity of OTF mode, due to its high scanning velocity, significantly exceeds that of traditional target mode observations. The optimum duration of an OFF measurement for OTF is given by \cite{Mangum_{2007}}:

$$t_{\text{off, opt}} = \frac{t_{\text{on}}}{\sqrt{N_{\text{on}}}}$$

where N_{on} denotes the number of ON measurements made per OFF measurement. It is very clear from Equation (4) that the efficiency of OTF can be improved with a larger N_{on}. However, the maximum allowable time between two OFF integrations is constrained by the system stability time, denoted as $t_{\text{stability}}$. Meanwhile, the minimum integration time, t_{on}, for each source point is limited by the data storage rate.

The scanning velocity must be calibrated such that the telescope's motion during a single integration between two data readouts covers less than half (Nyquist frequency) of, or even less than one third of the beamwidth. This adjustment is to minimize dynamic blurring in the direction of crucial telescope motion while facilitating rapid map scanning. Furthermore, when accounting for the dead time associated with the telescope' s slew from the source position to the OFF position, as well as the change in scan direction, a complex optimization process is required to determine these parameters effectively. In Figure 4, the scanning motion within the map is symbolized by a series of small, step-like structures in purple. Concurrently, at the designated OFF position, multiple data readouts, indicated by green rectangles, are conducted for identical locations.

3. HSTDM Data Processing Pipeline Concepts

The CSST HSTDM does not produce high-level science data in orbit; it only generates raw telemetry data which is transmitted to the ground station where Level 0 data are produced and archived according to the Interface Control Document (ICD) of HSTDM and the HSTDM Level 0 data definition file \cite{Tang_{2024}a}. The HSTDM pipeline starts with Level 0 data and is responsible for converting the Level 0 data to Level 1 and Level 2 science data through two levels of calibration procedures that will be briefly discussed in the following subsections. A schematic view of the HSTDM pipeline is displayed in Figure 5 [Figure 5: see original paper].

3.1. Data Format and Calibration

As depicted in Figure 5, calibrating HSTDM data involves converting a series of raw instrument outputs into scientifically usable data, which is typically in the form of antenna temperature versus frequency. Raw telemetry data from a single readout of HSTDM are in the form of spectral count versus corresponding channel numbers. An integration consists of several readouts. Given the

instrument drifts of HSTDM and the changing pointing during different readouts, these readouts cannot be immediately co-added. They have to be first converted into Level 0 data, which contain not only the raw readout that comes from space but also incorporate important positional and state information of the CSST platform at the beginning and end of the specific readout, as well as the observation mode and sequence information at the scheduling level. In fact, the Level 0 data are designed to be self-contained, in that necessary information for subsequent processing is provided.

For HSTDM, it is convenient to take observations as a function of time steps that are prescribed by the concrete observation command sequences based on the two aforementioned observation modes. At any moment, HSTDM is executing the flow of the observation command sequence, be it slewing the telescope, integrating on-source or off-source position, or doing internal load. At any time, HSTDM observes intensity as a function of frequency and outputs spectral data with precise timestamp codes. A given observation sequence could have hundreds and thousands of integrations, and their initial outputs are collected, recombined, restructured, and reformatted into a group of Level 0 data files by addressing the specific protocol and scheduling information meticulously at the ground station. The generated Level 0 data files are then archived in the Data File System (DFS) and can be accessed through the dedicated and unified interface provided by the CSST data processing system. The Level 0 data files and the relevant metadata files from both the scheduling system and the OOC system are all that are needed for the HSTDM pipeline. The HSTDM pipeline is designed in a way that it can handle data calibration, regardless of the specific observation mode, and outputs Level 1 and Level 2 data products that are in the Flexible Image Transport System (FITS) format with detailed definitions in \cite{Lin_{2024}} and \cite{Tang_{2024}b} respectively.

3.2. Processing Levels

Level 0 is the rawest form of HSTDM data available for the HSTDM pipeline. To begin with, Level 0 data must undergo a quality check, which consists of several sanity checks of FITS header items and data dimensions before being handed over to the HSTDM L1 pipeline. The main functions of the Level 1 pipeline are: (1) Flux-calibrate the HSTDM data using internal measurements and reference measurements; (2) Calibrate the pointing of HSTDM by applying the pointing model and calibration reference system parameters retrieved from the OOC system. Additionally, the channel numbers of Level 0 data are converted to observed frequencies with additional necessary calibration to remove the Doppler effect during the observations.

The Level 2 pipeline is mainly responsible for combining and merging integration data taken at different times of observations, as well as applying advanced methods to mitigate HSTDM' s instrument effects including baseline distortion, standing waves, and sideband imbalances, to name a few. For target mode observation, the Level 2 pipeline applies radiometric weighting to each Level 1

data file in the same group, which are then co-added to obtain long integration spectral data. For OTF observation, the Level 2 pipeline provides a standard regridding method to obtain the resampled data of the targeted area based on a series of sampling point data.

4. Synthetic Data Generation of HSTDM

Observational uncertainty stems from a lack of comprehensive data regarding the impact of the primary optical system of CSST on HSTDM' s performance. The Level 1 pipeline has completed two rounds of unit and integrated testing. Given this progress, it is now essential to conduct necessary simulation tests on the HSTDM pipeline before commencing comprehensive ground testing of the CSST data processing pipeline. This step will follow the installation and successful hardware testing of all backend modules within the primary optical instrument. These preliminary simulation tests are critical for identifying potential issues and optimizing HSTDM pipeline performance, thereby ensuring the accuracy, reliability, and efficiency of the HSTDM pipeline in subsequent ground testing and in-flight phases.

To facilitate this urgent need, we propose a Synthetic Data Generation method for HSTDM observations, focusing on simulating HSTDM instrumental effects, space environment conditions, and observation data flow based on observing mode scenarios. We also consider the IF characteristics to obtain the basic spectrum profile and the CSST orbit parameters to derive position and velocity information at the timestamps mentioned above. All these aspects are necessary to generate HSTDM simulation data that resemble real-world scenarios as closely as possible. The schematic view of the proposed method is displayed in Figure 6 [Figure 6: see original paper]. We will discuss related techniques at length in the following subsections 4.1, 4.2, and 4.3.

The current development of the HSTDM pipeline is nearing completion, except for certain high-level calibration methods within the Level 2 pipeline that remain undetermined.

4.1. Simulation of HSTDM Instrumental Effects

HSTDM is meticulously engineered to detect incoming signals that possess exceptionally low power levels with ample output. This necessitates a substantial receiver gain. In reality, however, the receiver gain is not perfectly stable. Consequently, even minor fluctuations in gain can contribute predominantly to the thermal noise of the receiver. Hence, the instability of HSTDM should be studied and considered in our simulated data.

To characterize the instability of HSTDM, we apply Allan variance analysis \cite{Riley_{2008}} to all channels' output data of HSTDM collected during the integration testing phase of HSTDM qualification components. The main configuration of the testing environment is summarized in Table 3 .

We use the open source project AllanTools \cite{Wallin_{2024}} to calculate the overlapping Allan deviation (OAD) of HSTDM output $x_{n,i}$, where i denotes the channel number, and $x_{n,i}$ is the time-series of the ith channel output, spaced by the measurement interval τ_0, with length N. The OAD of $x_{n,i}$ can be formulated as:

$$\sigma_{\mathrm{OAD}}(x_{n,i}) = \sqrt{\frac{1}{2(N-2m)} \sum_{n=1}^{N-2m} [x_{n+2m,i} - 2x_{n+m,i} + x_{n,i}]^2}$$

where $m = \tau/\tau_0$ and $\sigma_{\mathrm{OAD}}(x_{n,i})$ is the OAD of $x_{n,i}$.

The OAD of six randomly selected channel outputs of HSTDM under the testing environment is shown in Figure 7 [Figure 7: see original paper]. As the figure illustrates, the OAD of each channel output is very different from each other, although all seem to have a similar trend of OAD value changes. The turning points marked in red circles indicate τ values where the corresponding OAD reaches its minimal value, which we call the Allan time t_A, or system stability time as mentioned in subsection 2.2. We observed that t_A of the six displayed channel outputs ranges from 10 to 90 s, indicating differences in instability for each channel output. In total, there are 16,384 channels, which correspond to 0–1.2 GHz in the IF range of each of the two HSTDM detectors, with the ranges 0–0.16 GHz and 1.16–1.20 GHz omitted due to performance deterioration.

Therefore, the channel numbers of interest are from 2185 to 15,838. To further uncover the differences in t_A between all channel outputs of interest, we use a histogram to display the distribution of all calculated t_A values. The results are shown in Figure 8 [Figure 8: see original paper]. As the figure presents, we observe that the t_A values for outputs from channel 2185 to channel 15,838 range from 10 to 255 s, with a median value of 59 s. Thus, from a statistical perspective, we choose $t_A = 60$ s in our simulation. t_A, which also refers to $t_{\mathrm{stability}}$, is an important parameter both to characterize the stability of HSTDM and to determine the concrete values of parameters concerning the observation mode profile that will be discussed in subsection 4.3.

To simulate the instability effects in the simulated output data, we model the measurement $X_{i,t}$ in the output of channel i \cite{Caceres_{1997}} using the following stochastic differential equation (SDE), an Ornstein–Uhlenbeck (OU) process:

$$dX_{i,t} = \theta(\mu - X_{i,t})dt + \sigma dW_t$$

where μ is the long-term mean or equilibrium level to which the process reverts, $\theta > 0$ is the rate at which the process reverts to the mean μ, and $\sigma > 0$ is the volatility or the standard deviation of the random fluctuations. W_t is a standard Wiener process (i.e., a process for which dW_t/dt is a white noise process), which

represents Brownian motion. The analytical solution to $X_{i,t}$ can be formulated as:

$$X_{i,t} = X_{i,0}e^{-\theta t} + \mu(1 - e^{-\theta t}) + \sigma \int_0^t e^{-\theta(t-s)} dW_s$$

Equation (7) resembles the auto-regressive (AR) process form with the constant part $A = (1 - e^{-\theta\tau})\mu$, $B = e^{-\theta\tau}$, and the random part $CN_{0,1}$, where $N_{0,1}$ represents standard normal distribution.

In the simulation of instrumental effects, we propose a dual-pronged approach. The initial strategy involves utilizing measured output with the LO activated yet in the absence of any imposed incoming signals to establish a baseline representative of background noise. Subsequently, a predefined signal is superimposed upon this baseline, thereby emulating the response output during on-source observations. The alternative strategy employs the above-mentioned OU process model, tailored by assigning discrete parameter values $X_{i,0}$, μ_i, σ_i, and θ_i. This customization can effectively simulate the variegated channel output of HSTDM.

4.2. Space Environment Simulation

In simulating the space environment, we primarily consider the following three aspects: (1) The coupling coefficient of HSTDM' s antenna with the same target source differs between sunlit and shaded areas; (2) Cosmic ray interference in the frequency range of 410–510 GHz; (3) Cosmic microwave background (CMB) noise.

To simulate the variable coupling coefficients of HSTDM's antenna in both sunlit and shaded regions, we employ a binary analytical approach. This methodology commences with the determination of HSTDM's relative position with respect to the Sun and Earth. Based on this positional assessment, the coupling coefficient is assigned a value of η_1 for periods when the antenna is exposed to sunlight and η_2 during periods of shadow. Consequently, this process generates a temporal sequence of coefficient values that correspond with the orbital data, as delineated in the parameters setting block of the simulation.

For the cosmic ray interference simulation, we assume that cosmic ray interferences are transient in time and narrow-banded. We therefore generate a narrowband spectral signal in the frequency range of 410–510 GHz with amplitude level and occurring timestamp adjustable through parameter settings specified in the parameter setting block of the simulation.

For the CMB spectrum, we generate a white noise spectrum with an amplitude value of 2.7 K.

The schematic diagram of the simulation of space environment effects on HSTDM' s output data is illustrated in Figure 9 [Figure 9: see original paper].

4.3. Observation Data Flow Simulation Based on Observing Mode Scenarios

The observing mode scenarios are essentially structured sequences of temporal events that constitute the framework of standard observational cycles. These events can be delineated into discrete operational segments, each described by a distinct set of parameters as elaborated in Tables 1, 2, and Equation (4).

To accurately simulate data flow generation for typical target mode or OTF observations, it is imperative to initially configure all pertinent parameters along the timeline that define each operational step. These parameters encompass:

1. **Observation Parameters**: These specify the time slots allocated for all procedures conducted during a standard target mode or OTF observation. For comprehensive understanding, refer to Tables 1 and 2 for the definitions and meanings of these parameters.
2. **Time Sequencing Parameters**: These establish the start and end timestamps for the simulation, as well as the minimum duration of the discretized time slot, known as the time slot granularity.
3. **CSST Orbital Data**: Comprising millions of rows of positional and velocity data spanning a decade, with each row representing a 60 s interval. Interpolation is essential to derive positions and velocities at a time granularity finer than 60 s.
4. **Space Environmental Parameters**: As previously discussed in subsection 4.2, these parameters play an important role in the simulation.
5. **Source Spectrum and Instrumental Effects Parameters**: These delineate the standard spectrum of the target source for target mode observation or spectral map of the target region for OTF observation, as well as the instrumental effects specific to HSTDM. For target mode observations, standard spectrum data accessible from existing archives may require interpolation or extrapolation to align with HSTDM' s spectral data due to potential differences in spectral channel numbers. In the context of OTF observation, it is imperative that the spectral data simulated at each timestamp undergo initial interpolation from the standard spectral map data according to the most recently updated positional data. This preliminary step is essential prior to performing any subsequent interpolation or extrapolation to align with HSTDM' s output spectral data. The instrumental effects parameters are utilized as described in subsection 4.1.
6. **Cold Load Data**: The chop load data as previously discussed in Section 2.

Subsequently, employing the time sequencer and observation parameters, we generate a timeline of observation events at each time slot. We then iterate through the observation event list in each time slot, incrementally updating from fundamental information—such as CSST's position, velocity, attitude, and observation

state—to more sophisticated data, including space environmental impacts and observed frequency shifts in the spectral data. This process results in updated spectral data that incorporate all time-variant orbital and instrumental effects. The spectral data, to which the aforementioned effects have been applied, in conjunction with the updated header information for the Level 0 data and the updated observation state, collectively contribute to the generation of standard Level 0 data. This iterative process continues until the end of the observation event list.

We set the time granularity for the observation event list to be 0.25 s, which is the minimum readout time of HSTDM during observations. The length of the observation event list depends on the time span of the simulated scenario. It is important to note that the readout of HSTDM is in the form of spectral count while scientifically usable data (source spectrum data) is in antenna temperature. Therefore, conversion from antenna temperature to spectral count is needed during spectrum data simulation. We choose HSTDM' s output data under hot and cold load operations, collected in the hardware' s integrated test, as our benchmark for converting antenna temperature to spectral count. This conversion can be given by:

$$p_{\mathrm{sim}} = p_c + \frac{(T_{\mathrm{sim}} - T_c)(p_h - p_c)}{T_h - T_c}$$

where p_{sim} is the spectral data to be generated in spectral count, T_{sim} is the simulated antenna temperature of the spectrum, T_c and T_h are the blackbody temperatures of cold and hot loads, respectively, and p_c and p_h are the spectral counts measured during cold and hot load tests of HSTDM.

The overview of the proposed data flow simulation is illustrated in Figure 10 [Figure 10: see original paper].

5. Simulation Experiment and Discussion

We implement the above-mentioned simulation method in Python and construct observation scenarios to generate observation simulation data accordingly. The simulation data are a series of Level 0 data files that try to resemble HSTDM observational data as realistically as possible. We feed these data into our developing HSTDM pipeline and obtain subsequent Level 1 and Level 2 data products. By comparing the spectrum part in the data product with that calculated through another verified approach, we can draw conclusions regarding the accuracy of the HSTDM pipeline and the practicality of the proposed simulation method.

We first build a target mode observation scenario based on HSTDM output data collected during brightness temperature tests in the integration testing phase of HSTDM qualification components. The collected data can be classified into three types: (1) Cold Data: data collected during cold load (77 K) operation;

(2) Room Temperature Data: data collected during room temperature load (293.5 K) operation; (3) Signal Imposed Data: data collected when a narrowband radio frequency (RF) signal is imposed on the antenna.

By taking the cold data as the chop load data during nominal observation, the room temperature data as the output during OFF position integration, and the signal imposed data as the output during source position integration, we build a simple target mode observation scenario with observation data flow. We then pass these data into the HSTDM pipeline as illustrated in Figure 5. To better compare the spectrum data yielded by the HSTDM pipeline with those obtained through hot-cold calibration used in the brightness temperature tests \cite{Jin_{2024}}, we turn both the frequency calibration (Doppler frequency) module and pointing calibration module off in the HSTDM Level 1 pipeline.

Figure 11 [Figure 11: see original paper] illustrates the spectral plots as follows: the top subfigure depicts the channel output of HSTDM when a faint RF signal is applied; the middle subfigure presents the channel output of HSTDM under both hot and cold load conditions; and the bottom subfigure displays the spectrum data derived from the HSTDM Level 1 pipeline. Given that the imposed RF signal is a faint single-frequency RF with a known frequency of 492.3 GHz, it is barely discernible over a broad channel range. We opt for a zoomed plot around its corresponding channel number, as shown in the inset in the top subfigure. Additionally, the recovered RF signal and its corresponding channel output, along with its measured antenna temperature, are displayed in the inset in the bottom subfigure.

It is obvious from the figure that we can obtain the antenna temperature T_A (K) of the imposed RF signal after the routine process of the HSTDM pipeline, though we turn the pointing calibration module and frequency calibration module off. The spectral spike, as clearly displayed through all the subplots in Figure 11, is perhaps caused by IF chain noises. The intensity of the spectral spike is much larger than that of the imposed RF signal, as observed in the top and bottom subfigures.

We measure the antenna temperature T_A of the RF signal to be 2.253104×10^2 K. For comparison, T_A calculated using an alternative method for the brightness temperature test yields 2.253039×10^2 K. This alternative approach, as elaborated in \cite{Jin_{2024}}, involves leveraging the approximate linear relationship between antenna temperature of the imposed signal source and the channel output. This linear relationship is mathematically articulated in Equation (1). This alternative approach, leveraging measurement data obtained from both cold and hot load conditions to conduct a fitting procedure, is fundamentally analogous to the core module within the HSTDM Level 1 pipeline, albeit distinct in its approach to processing the measurement data. Both values are in good agreement, albeit not exactly the same. The minor differences in the calculation result stem from the specific details in how the two methods process the measurement data. Therefore, we have roughly verified the accuracy of the HSTDM Level 1 pipeline.

We also generate a series of simulation data files based on an archived OTF observation scenario by focusing on three critical aspects. Initially, we obtain the APEX spectral map data of the CO(4–3) emission line for SERP-MM18 as part of an effort to further investigate the outflows from Serpens South, complementing previous studies on its chemistry \cite{Gong_{2023}}. This dataset provides both the off-position spectral data and on-source spectral data after subtracting the off-position spectral data. Subsequently, we convert the header information and data dimension of each data file into the HSTDM Level 1 data format. Ultimately, we utilize these data files as inputs for the Regridding module within the HSTDM Level 2 pipeline, which produces the integrated intensity map of the CO(4–3) emission line. We then compare this intensity map with one generated by a verified pipeline using the GILDAS software \cite{Badia_{2025}}.

It is important to note that the generation of the aforementioned OTF simulation data encompasses solely the conversion of FITS header information and data dimension. We opt for this approach primarily because our objective is to validate the Regridding module within the HSTDM Level 2 pipeline.

Figure 12 [Figure 12: see original paper] depicts the distribution of observation points within the simulated OTF scenario. The source map is populated with source position markers, distinctly indicated by red crosses in the zoomed section of the plot. In stark contrast, the OFF position is marked by a blue circle. Collectively, there are 26,478 source position points accounted for in the simulation.

Figure 13 [Figure 13: see original paper] displays two distinct subplots: the left subplot illustrates the intensity map of CO(4–3) derived through the Regridding module within the HSTDM Level 2 pipeline, while the right subplot represents the intensity map generated using verified GILDAS software. Both methods use the same sampled data files, differing only in header format and data dimension of each data file. Upon comparing the intensity maps generated by these two methods, it is observed that the intensity values are largely consistent across corresponding positions within the scanned rectangular region. Additionally, discrepancies in intensity are noted along the border regions of the scanned area. Upon examining the reasons behind these discrepancies, we surmise that variances in the regridding algorithms, encompassing the respective parameters of both methods, could potentially account for the observed differences.

While we have thus far implemented only the essential part of the proposed simulation method to generate Level 0 data for the HSTDM Level 1 pipeline and Level 1 data for the Regridding module within the HSTDM Level 2 pipeline, we have already carried out preliminary validations to ascertain the accuracy of the HSTDM Level 1 pipeline as well as the Regridding module within HSTDM' s Level 2 pipeline. This initial validation underscores the practical utility of our proposed simulation approach. Moving forward, we are committed to developing the remaining aspects of the proposed simulation method and will persistently enhance them through integration with the most recent ground-based measure-

ment data and the evolving testing requirements of the HSTDM pipeline.

6. Conclusions

In this paper, we investigate synthetic data generation methods for HSTDM observations. First, we parameterize the basic operations in the observational cycles for both target mode observation and OTF observation. Second, we propose a simulation framework that focuses on emulating instrumental effects and space environment, as well as generating data streams based on observational modes. We provide detailed introductions to these simulation methods, which altogether strive to mimic the actual output of HSTDM as closely as possible. Third, we implement the simulation method and conduct simulation experiments, thereby validating the practical applicability of the proposed simulation method.

Acknowledgments

The authors wish to express their gratitude to Dr. Youhua Xu from the National Astronomical Observatories, Chinese Academy of Sciences, for generously supplying the orbital data of the CSST platform, which has been pivotal to our research endeavors. Sincere appreciation is extended to the HSTDM hardware development team for their invaluable contribution of test data during the system integration testing phase. Their dataset has undeniably served as the foundational bedrock upon which the research presented in this article is built. We also extend our heartfelt thanks to Visiting Scholar Gong Yan from Purple Mountain Observatory, Chinese Academy of Sciences, for providing the CO(4-3) spectral data, which were instrumental in conducting the OTF simulation analysis presented in this article. Additionally, we acknowledge the guidance and leadership of the CSST science data processing group, as well as the insightful discussions with many distinguished members of this group, which have greatly enhanced the depth and quality of our work. This article is made possible under the support of the National Natural Science Foundation of China (NSFC, grant Nos. 12427901 and 12403095) and the support of the program of the Ministry of Science and Technology of the People' s Republic of China under grant 2023YFA1608200. It is important to mention that some results in this paper have been derived using the Astropy, NumPy, and Pandas packages under the Python programming environment. We thank the developers of the Python programming language and the maintainers of these packages for making their code available on a free and open-source basis.

References

Ao, Y., Du, F., Li, D., et al. 2023, Report on the Assessment of Scientific Returns of High-sensitivity Terahertz Detection Module, Purple Mountain Observatory, CAS

Badia, F., Broguière, D., Buisson, G., et al. 2025, GILDAS, https://www.iram.fr/IRAMFR/GILDAS/

Cáceres, M. O., & Budini, A. A. 1997, JPhA, 30, 8427
Cao, Y., Gong, Y., Meng, X.-M., et al. 2018, MNRAS, 480, 2178
Gong, Y., Du, F. J., Henkel, C., et al. 2023, A&A, 679, A39
Gong, Y., Liu, X., Cao, Y., et al. 2019, ApJ, 883, 203
Herschel Space Observatory 2017, Herschel Explanatory Supplement, Volume II: The Heterodyne Instrument for the Far Infrared (HIFI) Handbook, HERSCHEL-HSC-DOC-2097
Hjalmarson, A. 2004, AdSpR, 34, 504
Jin, J., Liu, D., Lou, Z., & Lin, Z. 2024, Report on Brightness-temperature Tests of the HSTDM, Purple Mountain Observatory, CAS
Li, J., Deng, X., Li, Y., et al. 2025, Resea, 8, 0586
Lin, M., Shen, S., Liu, C., et al. 2024, Input Data Requirements and Level 1 Data Structure Design Specification for the CSST Scientific Data Processing System, KSC-00-JK-0002-03.01, National Astronomical Observatories, CAS
Mangum, J., Emerson, D., & Greisen, E. 2007, A&A, 474, 679
Melnick, G. J. 2002, AdSpR, 30, 2051
Ossenkopf, V., & Morris, P. 2005, HIFI Observing Mode Descriptions, ICC/2003-008, SRON
Riley, W. J., & Howe, D. A. 2008, Handbook of Frequency Stability Analysis, NIST 136
Roelfsema, P., Helmich, F., Teyssier, D., et al. 2012, A&A, 537, A17
Tan, S., Yao, Q., Li, J., & Shi, S.-C. 2024, JATIS, 10, 037002
Tang, J., Hu, Y., Nie, J., et al. 2024a, Input Data Requirements and Level 2 Data Structure Design Specification for the CSST Scientific Data Processing System, KSC-00-JK-0003-01.01, National Astronomical Observatories, CAS
Tang, J., Zhang, T., Chen, J., et al. 2024b, Input Data Requirements and Level 0 Data Structure Design Specification for the CSST Scientific Data Processing System, KSC-00-JK-0001-03.01, National Astronomical Observatories, CAS
Tiemeijer, L. F., Havens, R. J., de Kort, R., & Scholten, A. J. 2005, ITMTT, 53, 2917
Wallin, A. E., Price, D., Cantwell, G. C., et al. 2024, AllanTools, https://github.com/aewallin/AllanTools/releases
Yao, M., Liu, D., Liu, B.-L., et al. 2020, in IEEE 2020 33rd General Assembly and Scientific Symposium of the International Union of Radio Science
Zhan, H. 2021, ChSBu, 66, 1290
Zhang, K., Shi, S., Yao, Q., et al. 2018, in Int. Conf. on Microwave and Millimeter Wave Technology (ICMMT)

Data Availability: The data underlying this paper will be shared on reasonable request to the corresponding author.

ORCID iDs: Siyuan Tan https://orcid.org/0009-0001-6331-8708

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.