An Overview of the Mock LISA Data Challenges

The *Mock LISA Data Challenge Task Force*: Keith A. Arnaud[*], Stanislav Babak[†], John G. Baker[*], Matthew J. Benacquista[**], Neil J. Cornish[‡], Curt Cutler[§], Shane L. Larson[¶,‖], B. S. Sathyaprakash[††], Michele Vallisneri[§], Alberto Vecchio[‡‡] and Jean–Yves Vinet[§§]

[*]*Gravitational Astrophysics Laboratory, NASA Goddard Space Flight Center, 8800 Greenbelt Rd., Greenbelt, MD 20771, USA*
[†]*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Am Mühlenberg 1, D-14476 Golm bei Potsdam, Germany*
[**]*Center for Gravitational Wave Astronomy, University of Texas at Brownsville, Brownsville, TX 78520, USA*
[‡]*Department of Physics, Montana State University, Bozeman, MT 59717, USA*
[§]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA*
[¶]*Center for Gravitational Wave Physics, 104 Davey Laboratory, University Park, PA 16802, USA*
[‖]*Department of Physics, Weber State University, 2508 University Circle, Ogden, UT 84408, USA*
[††]*School of Physics and Astronomy, Cardiff University, Cardiff, CF243YB, UK*
[‡‡]*School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B152TT, UK*
[§§]*Department ARTEMIS, Observatoire de la Côte d'Azur, BP 429, 06304 Nice, France*

Abstract. The LISA International Science Team Working Group on Data Analysis (LIST-WG1B) is sponsoring several rounds of mock data challenges, with the purpose of fostering the development of LISA data-analysis capabilities, and of demonstrating technical readiness for the maximum science exploitation of the LISA data. The first round of challenge data sets were released at this Symposium. We describe the objectives, structure, and timeline of this program.

Keywords: gravitational waves, LISA, data analysis, galactic binaries, black holes
PACS: 04.80.Nn, 07.60.Ly, 95.55.Ym

1. OVERVIEW

The Laser Interferometer Space Antenna (LISA) is a spaceborne gravitational-wave (GW) observatory designed for detailed studies of a wide variety of GW sources throughout the Universe in the frequency range 0.1 mHz–0.1 Hz [1]. LISA is an all-sky monitor with the capability of measuring source parameters such as masses, spins, and distances with unprecedented precision. The LISA data set is expected to contain a large number ($\sim 10^4$) of resolvable overlapping sources, ranging from galactic subsolar-mass binary systems, to high-redshift massive black-hole binaries; in addition, LISA has the potential of discovering radically new classes of GW sources, such as primordial stochastic backgrounds, cosmic strings, and exotic compact objects [2].

Most sources detectable by LISA are long lived compared to the mission lifetime (> two years), and the LISA data will contain strong GW foregrounds generated by abundant populations of galactic and extra-galactic white-dwarf binary systems, and possibly of solar-mass compact objects captured by massive black holes in galactic nuclei. There

CP873, *Laser Interferometer Space Antenna—6th International LISA Symposium,* edited by S. M. Merkowitz and J. C. Livas

is no established expertise for this kind of data, although much relevant experience has already been gained in the analysis of GW data collected by ground-based detectors (see Saulson's contribution in this volume). In ground-based observations GWs are rare and weak, whereas in the low-frequency band they are numerous and (a fair fraction of them) yield high signal-to-noise ratios (SNR). It is therefore vital, in preparation for the mission, to tackle these new analysis problems early in order to develop the tools and methods necessary for the maximum science exploitation of such a revolutionary data set.

At the LISA International Science Team (LIST) meeting of December 2005, the Working Group on Data Analysis (LIST-WG1B) decided to embark in the organization of several rounds of mock data challenges (MLDC), with the dual purpose of (i) fostering the development of LISA data-analysis tools and capabilities, and (ii) demonstrating the technical readiness already achieved by the GW community in distilling a rich science payoff from the LISA data output. The LISA Mock Data Challenges were discussed at meetings organized by the US and European LISA Project, which were attended by a broad cross section of the international GW community. These challenges are meant to be blind tests, but not really contests; the greatest scientific benefit stemming from them will come from the quantitative comparison of results, analysis methods, and implementations.

A MLDC Task Force was constituted at the beginning of 2006 and has been working since then to formulate challenge problems of maximum efficacy, to establish criteria for the evaluation of the analyses, to develop standard models of the LISA mission and GW sources, to provide computing tools – LISA response simulators, source-waveform generators, and a MLDC file format – and more generally to provide any technical support necessary to the challengers. The challenges involve the distribution of several data sets, encoded in a simple standard format, and containing combinations of realistic simulated LISA noise with the signals from one or more GW sources of parameters unknown to the challenge participants, who are asked to return the maximum amount of correct information about the sources, and to produce technical notes detailing their work. In this short contribution we summarize the objectives, structure, and timeline of the MLDCs. In the companion contribution by the MLDC Task Force in this volume [3] we provide more technically oriented details about the MLDCs. Details can be found on the official MLDC website [4], in the living *Omnibus* document for Challenge 1 [5], and on the MLDC Task Force wiki [6].

2. STRUCTURE AND TIMELINE

The MLDCs consist in extracting the maximum amount of information about the source(s) that generate GW signal(s) contained in the (mock) data sets that are distributed. In order to ensure the incremental development of data-analysis approaches, software, pipelines, and infrastructure, the Task Force has decided to issue several rounds of challenges at intervals of approximately six months, each containing progressively more complex GW signals and noise realizations. Each round of MLDCs consists of multiple data sets with signals of different nature and strength embedded in synthetic noise. Two classes of data sets are distributed at each release: the proper challenge data

sets, where the source parameters are unknown, and training data sets, with GW signals of similar nature to those included in the blind tests, but whose parameters are made public. All the software and the necessary documentation to generate the data sets are public [4], so that interested parties can produce other mock data streams to facilitate the development of analysis tools, and the testing and validation of the algorithms. The challenge data sets are generated by two members of the Task Force who do not take part in the challenges, and who are the only repositories of the "keys" needed to recover the source parameters. Teleconferences open to all challenge participants will be organized to monitor progress. MLDC results, including analyses of how different methods performed, will be disseminated through technical documents and articles in peer-reviewed journals. Prospective challenge participants are asked to subscribe to the `lisatools-challenge` mailing list (see [4]), which offers a natural communication link between the Task Force and the broader community.

The goal of the first round of challenges (Challenge 1) is to foster the development and validation of building blocks and basic tools for LISA data analysis, and to tackle the analysis of data sets containing a single signal or a small number of nonoverlapping signals (with one important exception), embedded in Gaussian and stationary noise, with no contribution from galactic and extragalactic foregrounds. Challenge-1 data sets were released in June 2006 (at the end of this Symposium), and results are due on or before December 1, 2006. They will be discussed and presented to the broader GW community in a dedicated session at the 11th Gravitational-Wave Data Analysis Workshop (in December 2006 at the Albert Einstein Institute in Golm, Germany).

The next set of challenges (Challenge 2), expected to be released in December 2006 with results due in June 2007, is regarded by the Task Force as a key milestone in the MLDC effort: the goal is to tackle a *global* analysis problem – the distinctive feature of LISA from a GW data-analysis point of view – in the presence of foregrounds, but within a restricted parameter space, in order to limit computational requirements at this early stage. We expect the results of Challenge 2 to provide a proof of concept for the LISA data-analysis approach that can serve as a basis for more detailed studies and further development. Although a final decision on the exact GW content of this data set has yet to be made – progress on Challenge 1 will surely affect decisions on the format of Challenge 2 – data streams will contain a stochastic foreground (produced by a galactic-population–synthesis model), LISA verification binaries, a small frequency band with strongly overlapping yet resolvable stellar-mass binary systems, a few coalescences of massive black-hole binary systems, and ~ 1 extreme–mass-ratio inspiral (EMRI). Moreover the total number of sources in the data set will be unknown. Other data sets containing just one source – such as EMRIs, stochastic backgrounds, and short broadband bursts – will also be produced to address source specific analysis problems.

The third round of challenges (Challenge 3), expected to be released in June 2007 with results due in December 2007, will focus on testing an even more realistic data-analysis scenario, including noise non-Gaussianity and nonstationarity, data gaps, more general gravitational waveforms, and a larger number of sources; details will clearly be influenced by the results of Challenges 1 and 2. MLDCs will continue beyond the third round, but the exact schedule and format has not been discussed yet.

3. THE FIRST ROUND OF CHALLENGES

The first round of challenges is aimed at enabling the development of the building blocks necessary for simple LISA data-analysis tasks. It focuses on data sets that contain a single source (generally with moderate-to-high SNR), or a small number of sources that do not overlap in the frequency domain; an important exception is however made for two data sets that contain a few tens of galactic binary systems whose radiation overlaps significantly in a small frequency region. These latter data sets provide a fairly realistic representation of the novel aspects of LISA data-analysis, although restricted to a limited frequency band and to a particular class of signals. It should therefore provide an early assessment of the soundness of the techniques that have been investigated so far to tackle this particular problem. The first round of challenges concentrates on sources currently listed as minimum science requirements [7]: galactic binaries (including verification binaries), and massive black-hole binary systems. Data sets containing EMRIs are also made available now, in order to facilitate the development of data-analysis tools for this especially demanding class of sources. However, the results of this EMRI challenge are due only in June 2007 (the due date for Challenge 2).

All the data sets, with the exception of EMRIs, are one-year long, and contain instrumental noise modeled as Gaussian and stationary, with the assumption that laser frequency noise has been removed perfectly. No foreground radiation is included. More details are provided in the companion contribution in this volume [3] and in Refs. [4, 5].

3.1. Galactic stellar mass binaries

Challenge 1 includes seven distinct data sets containing radiation from stellar-mass galactic binary systems. Waveform are approximated as exactly monochromatic in the source reference frame: the signal is therefore described by seven parameters [3]. The first three data sets contain a single signal from a galactic binary (with optimal SNR for a single TDI output ≈ 20) whose geometrical parameters are drawn randomly over the whole relevant range. The three signals differ primarily in the frequency, which was chosen randomly in the interval 0.9 mHz–1.1 mHz, 2.9 mHz–3.1 mHz, and 9 mHz–11 mHz, respectively.

Another data set contains "verification binaries," defined for the purpose of the MLDCs as stellar-mass binary systems whose orbital period and location in the sky are *exactly* known, whereas all other parameters are completely unknown—this is a fairly realistic approximation for the actual known systems in the Galaxy. We injected six signals from real verification binaries (whose supposedly unknown parameters were randomly drawn from uniform distributions over the relevant range) – RXJ0806, V407Vul, ESCet, AMCVn , HPLib, and EIPsc – and 14 signals from synthetic verification binaries that were generated randomly using a binary-population–synthesis code. In all cases we ensured that the coherent SNR over one year of integration was > 10. In order to incrementally increase the difficulty of the analysis process, Challenge 1 includes also a data set containing 20 unknown galactic binaries whose parameters were all chosen randomly and with frequency in the range 0.1 mHz–10 mHz.

The next two (and final) data sets for galactic binaries included in Challenge 1

are of somewhat different nature, and embody the main novel feature of LISA data analysis: the presence of an unknown number of sources overlapping in the time and frequency domains. We generated two data streams, each containing (in a small band of frequencies) a number of sources comparable to what is predicted by our present understanding of galactic substellar-mass binary populations. Challenge participants are given only a range for the number of binaries present in the data stream (between 40 and 60), and the frequency band in which they are located (between 3 mHz and 3.015 mHz for the mildly overlapping sources in the first set, and between 2.9985 mHz and 3.0015 mHz for the strongly overlapping sources in the second set). All the other parameters are chosen randomly, with the only constraint that the coherent single–TDI-observable SNR must exceed five.

3.2. Massive black hole binary systems

Challenge 1 includes two data sets containing a single MBHB inspiral. The waveforms are approximated at the restricted second post-Newtonian order (see [3, 5, 4] for more details), with no spin-orbit– nor spin-spin–induced modulations. The waveform is described by nine parameters; the mass of the primary object was chosen from a uniform distribution in the range 1–5 $\times 10^6 M_\odot$, and the mass ratio in the range 1–4. The main differences between the two data sets are the time at which coalescence occurs, and the optimal SNR: the first data set was chosen to contain about half a year of effective inspiral (with coalescence time unknown within a range of 40 days), and to have SNR ≈ 500; the second data set to cover an earlier stage in the life of a binary (with coalescence occurring about a month ± 20 days from the end of the data set), and to have a reduced SNR (in the range $\approx$ 20–100).

3.3. Extreme–mass-ratio inspirals

The final sets of Challenge 1 contain radiation from an EMRI. At present, this source class is not included in the minimum science requirements; however, the scientific payoff from such observations is so high, and their analysis so demanding, that the Task Force felt it was important to release EMRI data sets early on, to help in focusing the development of data-analysis techniques. Challenge 1 includes five two-year long data sets containing EMRIs with SNR in the range 20–120. The computation of the EMRI waveforms represents in itself a major theoretical challenge that still needs work; for Challenge 1, we adopted the "analytical kludge waveforms" [8], which are fastest to compute, while retaining most of the qualitative features of actual signals (for more details see [3, 5, 4]). The waveforms depend on 14 parameters (we assume that the captured object has negligible spin), chosen in an astrophysically realistic range: the total and reduced mass of the system were drawn from uniforms distribution in the ranges 9.5×10^5–$1.05 \times 10^6 M_\odot$ and 9.5–$10.5 M_\odot$, respectively; the spin of the central massive object was chosen in the range 0.5–$0.7 M^2$, and the orbital eccentricity at plunge in the range 0.1–0.25. The plunge was set to take place between 1.5 years and 2 years + 2 weeks from the beginning of the observation.

ACKNOWLEDGMENTS

M.V.'s work was supported by the LISA Mission Science Office and by the Human Resources Development Fund at the Jet Propulsion Laboratory, California Institute of Technology, where it was performed under contract with the National Aeronautics and Space Administration.

REFERENCES

1. LISA Study Team, "LISA: Laser Interferometer Space Antenna for the detection and observation of gravitational waves," Pre-Phase A Report, 2nd ed. (Max Planck Institut für Quantenoptik, Garching, Germany, 1998).
2. C. Cutler, and K. S. Thorne, in *General Relativity and Gravitation*, N. T. Bishop and D. M. Sunil, eds. (World Scientific, Singapore, 2002), p. 72.
3. MLDC Task Force, "A How-To for the Mock LISA Data Challenges," in this volume.
4. Mock LISA Data Challenge Homepage, `astrogravs.nasa.gov/docs/mldc`.
5. Mock LISA Data Challenge Task Force, "Document for Challenge 1," `svn.sourceforge.net/viewvc/lisatools/Docs/challenge1.pdf`.
6. Mock LISA Data Challenge Task Force wiki, `www.tapir.caltech.edu/dokuwiki/listwg1b:home`.
7. T. A. Prince and K. Danzmann, "LISA Science Requirement Document," v. 3.0, 12 May 2005, at `www.srl.caltech.edu/lisa/documents`.
8. L. Barack and C. Cutler, *Phys. Rev. D* **69**, 082005 (2004).