Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

DATA PAPER

LUNDI$_{sim}$: model meshes for flow simulation and compression benchmarks

Laurent Duval[1] | **Frédéric Payan**[2] | **Christophe Preux**[1] | **Lauriane Bouard**[1,2]

[1]IFP Energies nouvelles, France

[2]Université Côte d'Azur, CNRS, I3S, France

Correspondence
Corresponding author Laurent Duval
Email: laurent.duval@ifpen.fr

Abstract

The volume of scientific data produced for and by numerical simulation workflows, particularly in geoscience, is increasing at an incredible rate. Among remedies, lossless and lossy compression techniques are becoming popular. Their assessment require open datasets shared under FAIR principles (Findable, Accessible, Interoperable, Reusable), with MRE (Minimal Reproducible Example) ancillary data for reuse. We share LUNDI$_{sim}$, an exemplary faulted geological mesh. Enhanced by porosity/permeability datasets, the model proposes four distinct subsurface environments. They are primarily designed for flow simulation in porous media. Several consistent resolutions are proposed for each model. We also provide a set of reservoir features for reproducing typical two-phase flow simulations on all LUNDI$_{sim}$ models in a reservoir engineering context. This dataset is chiefly meant for benchmarking data reduction (upscaling) or genuine mesh compression algorithms. It is also suitable for advanced mesh processing workflows in geology, from visualization to machine learning.

KEYWORDS
open data, scientific data compression, geology, volumic model mesh, flow simulation

1 | INTRODUCTION

Science has entered the so-called "fourth paradigm" of data-intensive computing for discovery [1]. Increasingly accurate models yield unprecedented access to more precise simulations, resorting to high-performance computing (HPC) facilities. The exploitation of massive datasets is however hampered by many size-related issues, such as storage, workflow management, visualization [2,3], etc.

As a result, data compression is making a comeback from an influential 1990's multimedia era‡ to the many worlds of modeling and simulation. At stake are legal long-term storage issues for instance in climate modeling [4], checkpoint restart or snapshotting for fault-tolerance in HPC [5], approximate computing [6], faster selection of parameters with smaller simulation models, progressive result retrieval [7], *in situ*/in storage processing [8], objective and subjective performance evaluation, etc.

For a comprehensive evaluation of compression performance and challenges in modeling [9], several issues deserve attention, sometimes in contrast to what was held true for multimedia data coding. We thereafter detail five most prominent issues: efficiency, discrepancy, diversity, interpretability, availability, before we inflect them to geological models in Section 2, thereby motivating the proposed LUNDI$_{sim}$ mesh in the remaining of the paper.

Issue 1 (efficiency) perfect or lossless compression like "zip" notoriously yields very limited reduction ratios. More than two- to three-fold reduction in size is rare (except for highly-structured data). Therefore, approximate, near-lossless, progressive or lossy compression algorithms are required to ensure a significant byte-size reduction compatible with the pressing needs occurring from gigantic simulation volumes. They however entail careful assessments of the data loss impact on performance [10,11,12,13], especially for bounding errors [14,15].

Abbreviations: FAIR, Findable, Accessible, Interoperable, Reusable ; MRE, Minimal Reproducible Example ; HPC, High Performance Computing ; HS, HexaShrink ; CPG, Corner Point Grid.

‡ Famous data compression standards: jpeg, gif, png, mp3.

 © 2024 Copyright Holder Name

Issue 2 (discrepancy) simulation dataset are typically very heterogeneous. Coming from different sources, at various workflow steps, they include structured, semi-structured, and unstructured data, are of various dimensionality (1D, 2D, 3D and 4D) and stored in various formats (booleans, discrete labels, integers, floats, etc.) or containers (json, HDF, netCDF, FITS, Mexus, ASDF, XLM). They cannot be addressed with generic compression tools easily, nor with optimal performance. They require dedicated algorithms, taking into account complicated data morphology [16].

Issue 3 (diversity) mixed and high-dynamics data: each type of data may exhibit a huge diversity of types, ranges, or statistical distributions, from a handful of finite nominal categories (Likert scale, data labels, attributes) to high-precision values (covering several range scales) with unbalanced histograms [17]. At stake here are quantities whose variations have highly non-linear behavior or non-proportional effects. For instance, small values that would be discarded with traditional lossy compression may need to be faithfully preserved.

Issue 4 (interpretability) direct interpretation of the different — and often visually combined — types of scientific data [18] is less straightforward than with standard audio, image or video [19]. First, it is heavily coupled with physical modeling. Second, models potentially undergo long-lasting simulations whose outputs are subject to a host of objective and subjective assessments. Simulation evaluations gather teams with diverse skills. Their expertise is deployed iteratively, at different stages of the workflow. Owing to simulation complexity and compression recency, overall quality assessment is restricted to a small number of individuals from distinct backgrounds, with little universally-accepted metrics and huge policy options. Acceptable objective losses with no influence on simulation may become unacceptable to an expert subjective interpretation, focusing on specific modalities. In contrast, knowledge of the human sensory systems and the world-wide dissemination of multimedia devices allowed the persistence of widely-accepted compression of audio and visual contents.

Issue 5 (availability) open availability of representative models, in FAIR principles [20], is not granted, for proprietary uses or ad-hoc data manipulation that cannot be reproduced. One of the co-authors of this paper has for instance encountered apparently huge private meshes which, candidates for potential challenges, showed up to have been artificially inflated by linear interpolation on mid-scale data. This may jeopardize fair compression evaluation, as data becomes highly predictable. Therefore, openly shared geological models, that may be modified to adapt to different simulation contexts, are convenient.

2 | GEOLOGICAL MODEL ISSUES

Aside gigantic climate-related models, geoscience somehow lacks open, manageable, heterogeneous data models that can be embedded in a processing or simulation workflow. In a similar way to recent initiatives [21,22], we share with this paper a handful of 3D geological models and their multiscale representations. Developed at IFPEN during Lauriane Bouard's PhD thesis [23], this dataset is collectively denoted as LUNDI$_{\text{sim}}$. LUNDI$_{\text{sim}}$ is dedicated to performance evaluation and benchmarks around the compression of 3D geological models [24] targeted to simulation workflows, illustrating the five previously raised issues. We named our dataset LUNDI$_{\text{sim}}$, after the Icelandic name of the (peaceful) Atlantic puffin. This name is a friendly nod to two protagonistic simulation software suites, Petrel™ and SKUA™, named after two (highly competing) seabirds. LUNDI$_{\text{sim}}$ was initially created for testing HexaShrink (HS) [25], a scalable storage and multiresolution (also called hierarchical [26,27,28,29]) visualization framework for hexahedral meshes with mixed attributes and discontinuities. HS was then integrated into a comprehensive compression workflow, enabling progressive and refinable data representation of composite hexahedral meshes. "Composite" here means that the 3D geometric structure (or grid) may itself be encoded by complementary spatial locations. In computational geology, this geometry is traditionally structured by a Corner Point Grid (CPG): a 1D coordinate system along the vertical direction ("Pillar") supports a more horizontal 2D layering ("Zcorn"). This grid may be complemented with numerical properties (porosity, expressed in unit proportion; permeability given in (milli)darcy or (m)d in the following) and discrete categories (cell activity, rock type) designed from rock physics, for flow simulation in reservoir modeling engineering. There, a geological model may be filled by different stochastic distributions. They account for phenomena representing variations in the underground. Once filled with properties, a reservoir model is simulated under varying operating conditions. Such simulations are used to gain insight on how to manage a storage or production facility on a day-to-day basis.

In [25], we observed that different data in composite meshes distinctly affect compression algorithms. For the sake of completeness, we provide here an illustrative example, based on one of our LUNDI$_{\text{sim}}$ models, described thereafter. The dark blue bars in Figure 1 represent the "raw" number of bits per symbol for various data types in a model cell (i.e., Zcorn, Pillar, Activity, Porosity and Permeability). The orange bar depicts the direct application of the generic yet highly lossless LZMA (Lempel-Zip-Markov chain Algorithm) [30] Section 6.26 coder on all components, with mild average compression (see Issue 1 from Section 1). More

specifically, we observe that heterogeneous data types have distinct compression ratios (Issue 2). Boolean Activity property is easily compressed, while Permeability is more challenging. As in[25] we sought at the same time both lossless compression and the possibility to address mesh multiresolution, the gray, yellow and light blue bars of Figure 1 indicate the LZMA performance after respectively one, two or three levels of multiscale (HS×1, HS×2, HS×3 respectively) decompositions of all properties.

As seen per the average number of bits per symbol, integer geometry (Zcorn and Pillar) is increasingly compressed (though mildly) with resolutions, while the effect on continuous scalar properties (Porosity, Permeability) is sightly degraded due to the high-dynamics of the data (Issue 3).

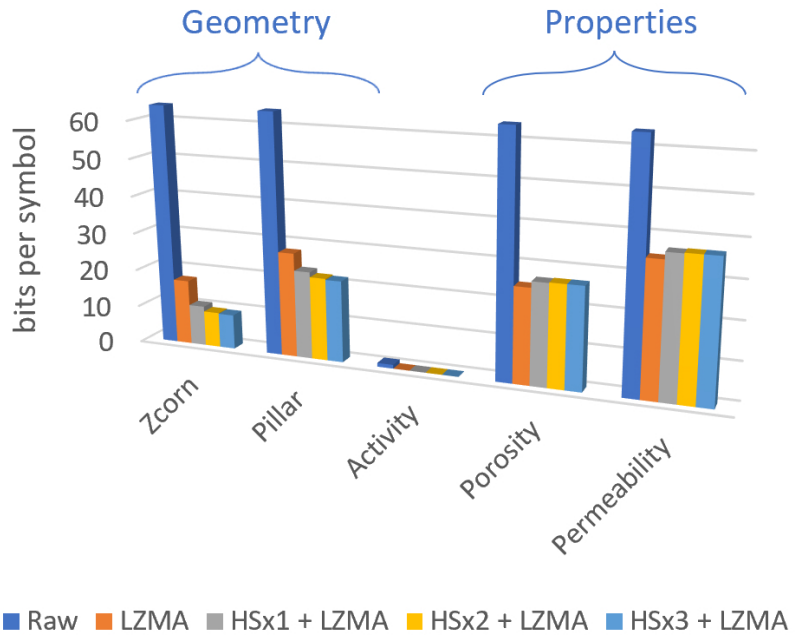

FIGURE 1 (Average) number of bits per symbol for each data type of the nearshore$_0$ LUNDI$_{sim}$ model (Zcorn, Pillar, Activity, Porosity, Permeability) in function of the encoding setting: uncompressed raw data (dark blue bars), raw data compressed with LZMA (orange bars), data decomposed at one (gray bars), two (yellow bars) or three (light blue bars) different resolutions before being compressed with LZMA. See[25] Fig. 17–18 for more visual details.

Above observations were made on losslessly compressed data. In other terms, decompressed data is faithful to the raw model, hence does not hamper workflow precision, notably in a context of simulation. However, simulation practice often resorts to data at coarser resolutions, for speedups and multi-scenario evaluations. Plus, it is well-known that different data resolutions (scales) or precision (byte-per-symbol) may subjectively impact a simulation workflow (Issue 4). In a typical compress-once/decompress-many context, one may need for instance to address objective mesh size and decompression speed metrics at the beginning of the workflow, and more subjective replays of flow propagation for post-processing. Therefore, our LUNDI$_{sim}$ dataset contains models at four different levels of resolution to address Issue 5.

The remaining of the paper is organized as follows. We provide contextual information on reservoir modeling for simulation in Section 3, inspired from the well-known reservoir engineering challenge SPE10. We craft the two main components of LUNDI$_{sim}$ in Section 4: the common model mesh (Subsection 4.1) and its SPE10-inherited physical properties (Subsection 4.2). Ancillary data for simulation are provided in Section 5: global reservoir characteristics (Subsection 5.1) to allow simulation workflow reproduction (MRE, up to software suite characteristics); the application to fluid production (Subsection 5.2) with traditional simulation observables. Section 6 details data availability and associated software. LUNDI$_{sim}$ potential reuse and limits are given in Section 7, before conclusions (Section 8).

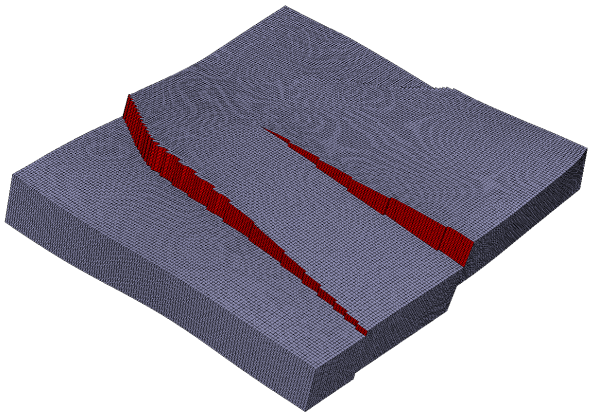

FIGURE 2 LUNDI$_{sim}$ model mesh (128 × 128 × 32 cells, three faults).

3 | RESERVOIR MODELING FOR SIMULATION

We base our work on a previously published challenge known as SPE10, i.e. the Tenth SPE Comparative Solution Project[31] for reservoir simulation. We consider its second problem called Model 2, part of the *Brent* sequence (quoting), "*a waterflood of a large geostatistical model chosen so that it was hard (though not impossible) to compute the true fine-grid solution*"[31] p. 308. In this challenge, eight companies competed to obtain the best possible outcome in the evaluation of this model, using a combination of simulation software and upscaling techniques. Counter-intuitively (since data science is more acquainted with downsampling or downscaling), in the context of reservoir simulation, upscaling and upgridding denote the process with which a fine-scale geological model (a grid assorted with rock properties such as porosity and permeability data) is converted into coarser models that are more computationally tractable, while providing outcomes as close as possible as those expected from the finer grid. Cells of the coarser grid (upgridding[32]) are filled with equivalent properties (upscaling) obtained from finer-resolution cells, using a variety of homogenization or averaging techniques. We refer to[31,33,34] for details.

Upscaling thus reduces the original grid size as well as cell-borne quantities. This results in a global reduction of the size of data with heterogeneous properties, a process similar to what is targeted in genuine data compression, where the modification of data resolution is combined with variations in data precision and additional entropy coding schemes that yield a final compressed file. We refer to[30] for advanced notions in data coding. Meanwhile, one may ask whether suitable data compression, adapted to geological data, is compatible and even maybe beneficial to flow simulation of large heterogeneous models, as partly exposed in[23,35] (whose outcomes are not required here for further understanding). We now focus on LUNDI$_{sim}$ benchmark models.

4 | LUNDI$_{sim}$ MODEL DESCRIPTION

4.1 | LUNDI$_{sim}$ model mesh

Figure 2 provides an overview of the model hexahedral mesh underlying all LUNDI$_{sim}$ models. This mesh bears a geological morphology similar to SPE10 dataset 2 (quoting): a "*simple geometry, with no top structure or faults*". It mainly differs in its lengths in each dimension (chosen as powers of two) and the addition of faults, which are challenging for upscaling/upgridding, multiscale decomposition, mesh compression (as vertices are not conform) and flow simulation (as faults affect fluid displacement).

The topography of LUNDI$_{sim}$ models stems from a realistic reservoir engineering case. It forms one quarter of an anticline structure (Figure 3), common in hydrocarbon trap reservoir study. The highest point (P_1) corresponds to the top of the anticline (3360 m depth). The opposite corner is situated 50 m below, on the same horizon.

LUNDI$_{sim}$ model mesh contains three continuous vertical stair-step faults. Two are apparent in red within Figure 2, the third one bulging from the top-right side. They are not aligned along grid axes and possess different offsets to emulate mildly complex environments. Its structure is composed of 128 × 128 × 32 cells to allow reasonable simulation times. The average cell size

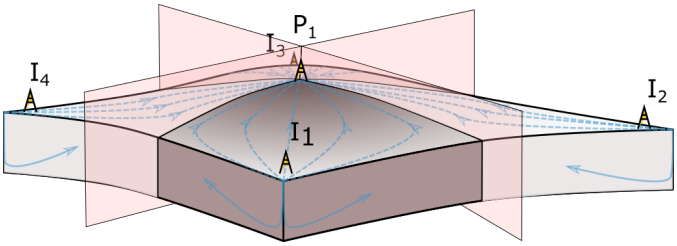

FIGURE 3 Reservoir engineering: an injection/production system. In this quarter five-spot configuration model[36], waterflows (depicted by arrowed blue lines) are produced via injection from four wells ($\{I_i\}$) up to one central producer well P_1, located on the highest point. LUNDI$_{sim}$ represents one quarter of this reservoir model (highlighted in brown and limited by vertical reddish planes), including the wells I_1 and P_1.

represents a volume of size 1.70 m × 1.70 m × 0.95 m, which is common in sedimentary geology for modeling horizontal fine deposits of geologic material over the years. The numbers of cells (128, 128 and 32) in each dimension are powers of two ($2^7; 2^7; 2^5$). This choice allows to implement the most standard dyadic computations, subsampling or decompositions, to better benchmark compression methods. This choice allows to scale the mesh dimensions by five scales, down to LEGO® brick sizes. Note that size reduction by one or two dyadic scales often suffices. Therefore in practice, non-dyadic dimensions may be handled by only padding cells to the next even or quadruple integer, or using activity labels.

4.2 | LUNDI$_{sim}$ model properties

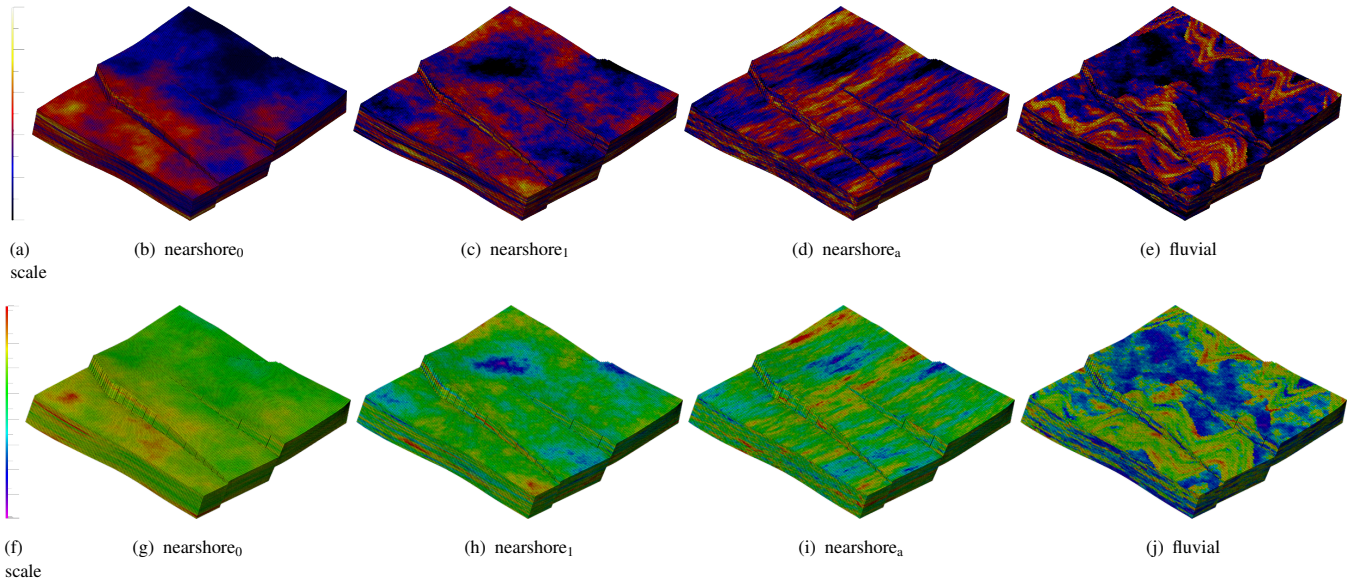

FIGURE 4 Four geological environments supplied in LUNDI$_{sim}$ dataset: nearshore$_0$, nearshore$_1$, nearshore$_a$, and fluvial. Porosity (top) and permeability (bottom) properties range from 0.0–0.5 (as a unit fraction) and 700 × 10^{-6} md–20 × 10^3 md, respectively.

The mesh is enhanced by two continuous petrophysical properties, porosity and permeability, required for the simulation benchmark, partly presented in[35]. The spatial distributions of those properties are inspired by two geological formations in[31]:

Ness[§] and *Tarbert*[¶] Note that we do not consider here rock types: though they are important in overall compression schemes[25], they were not required for our flow simulation purpose. As there is no obvious mapping from one geological object to another, we draw four different stochastic realizations to emulate four distinct environments, from homogeneous to anisotropic, which are displayed in color scales in Figure 4.

The three first correspond to prograding nearshore environments (*Tarbert* formation) with smooth property variations: nearshore_0 and nearshore_1 have been generated by an isotropic distribution with different ranges of dependence, while nearshore_a exhibits more anisotropy. The fourth "fluvial" model (*Upper Ness* formation) exhibits sharper contrasts, with distinctive heterogeneous geological objects. This discrepancy between environments emulates a wide range of petrophysical system behaviors.

The conception of the initial common grid, the inclusion of faults and property filling have been performed with Paradigm™ 3D geological modeling software GOCAD (formerly known as GeOCAD, Geological Objects Computer-Aided Design; now SKUA)[#] and the MATLAB Reservoir Simulation Toolbox (MRST)[36].

5 | SIMULATION SETTINGS

The following provides ancillary data, adapted from SPE10, so that $\text{LUNDI}_{\text{sim}}$ can be simulated through a Minimal Reproducible Example (MRE) workflow. Precise outcomes may of course depend on alternative software choices and expert decisions.

5.1 | Global reservoir characteristics

As for the reservoir model, the rock compressibility is set to $1 \times 10^{-6}\,\text{bar}^{-1}$ and the reservoir pressure is set to 200 bar at the water-oil contact, fixed at a depth of 3410 m. Finally, the reservoir temperature is set to 60 °C.

The simulation workflow is backed on the test case of a so-called black-oil model[27]. It consists of two liquid phases: water and dead oil (with no gas dissolved). We introduce the Formation Volume Factor (FVF) quantity: ratio of volumes occupied by a fluid at reservoir conditions versus surface conditions. Quantities below are again borrowed from SPE10, and recalled for completeness. For water, viscosity pressure, density and FVF are computed by correlation from the reservoir simulator Pumaflow®[‖]. For oil, some quantities are given by tabulations. The viscosity pressure (in centipoise, cP) is computed from Table 1, the density is set to $1\,\text{kg/m}^3$, and the oil FVF (B_o) is tabulated in Table 2.

TABLE 1 Oil viscosity in centipoise (cP), tabulated as a function of pressure.

Pressure (bar)	Viscosity (cP)
50	2.85
200	2.99

TABLE 2 Oil Formation Volume Factor (FVF) B_o, tabulated as a function of pressure.

Pressure (bar)	B_o
50	1.05
200	1.02
500	1.01

[§] https://data.bgs.ac.uk/id/Lexicon/NamedRockUnit/NESS.
[¶] https://data.bgs.ac.uk/id/Lexicon/NamedRockUnit/TARB.
[#] https://www.aspentech.com/en/products/sse/aspen-skua
[‖] https://www.beicip.com/pumaflow

TABLE 3 Relative permeability curves tabulation for water (Kr_w) and oil (Kr_o) as a function of water saturation (S_w).

S_w	Kr_w	Kr_o
0.200†	0.0000	1.0000
0.250	0.0069	0.8403
0.300	0.0278	0.6944
0.350	0.0625	0.5625
0.400	0.1111	0.4444
0.450	0.1736	0.3403
0.500	0.2500	0.2500
0.550	0.3403	0.1736
0.600	0.4444	0.1111
0.650	0.5625	0.0625
0.700	0.6944	0.0278
0.750	0.8403	0.0069
0.800‡	1.0000	0.0000

†Irreducible water saturation value (S_{wi}).
‡Residual oil saturation value (S_{or}).

We now turn to water/oil mixture characteristics with relative permeabilities for water (S_w) and oil (S_o), respectively. Given that the latter is obtained from the former by $S_o = 1 - S_w$, we tabulate relatives permeability curves in Table 3, for water (Kr_w) and oil (Kr_o). Here, the irreducible water saturation is $S_{wi} = 0.2$ (Table 3, top of first column) and the residual oil saturation is $S_{or} = 0.2$ (complement to the S_w given in Table 3, bottom of first column).

5.2 | Application to fluid production

We finally present a typical two-phase flow simulated on LUNDI$_{sim}$ (full resolution, nearshore$_0$ environment). Initially, two phases in the reservoir are horizontally stratified, with oil above water. The two wells are drilled in the whole depth of the reservoir. At $t = 0$, water is injected by I_1 in the lower part of the reservoir (Figure 3). The water pressure pushes the oil through the reservoir up to the producer P_1 (distant from 300 m). Injector pressure and producer rate remain constant, respectively set at 300 bar and 300 m^3 per day.

One valuable indicator in oil production to determine field exploitation is the estimated water cut, i.e., the ratio between water and total liquid volumes, at the producer well. It is regularly recorded over a period of time expressed in days (Figure 5). The inflection point (red point on the curve) is the water breakthrough, which denotes the water arrival at P_1. From that instant, the extracted liquid contains more and more water. To avoid expensive post-processing and optimize the exploitation configuration, reservoir engineers aim to delay this instant. Simulation is a powerful tool to estimate such predicted water cut curves. To determine the best exploitation configuration, many simulations with varying parameters can thus be run until satisfaction, involving very long computation times. We expect to evaluate the positive impact of resolution and precision variations (on computational time) on the latter.

6 | DATA FORMAT AND ACCESS

The four LUNDI$_{sim}$ models presented in this article (one per environment) are provided as “Grid Eclipse" (GRDECL) data, a *de facto* standard for grids with hexahedral cells, developed by Schumberger for the ECLIPSE™ Reservoir Simulator**. They are available at Zenodo†† and from author website‡‡. Lower resolutions of LUNDI$_{sim}$, produced by HexaShrink, are also available.

LUNDI$_{sim}$ illustrations from Figures 4 and 2 were made with *ResInsight*§§ (v.2023.06¶¶), an open source cross-platform 3D visualization and post-processing tool for reservoir models and simulations (developed in Python, available for Windows and

** https://www.software.slb.com/products/eclipse
†† [TO DO. GIT HUB, PANGAEA?]
‡‡ http://www.laurent-duval.eu/opus-lundisim.html
§§ https://resinsight.org
¶¶ https://github.com/OPM/ResInsight

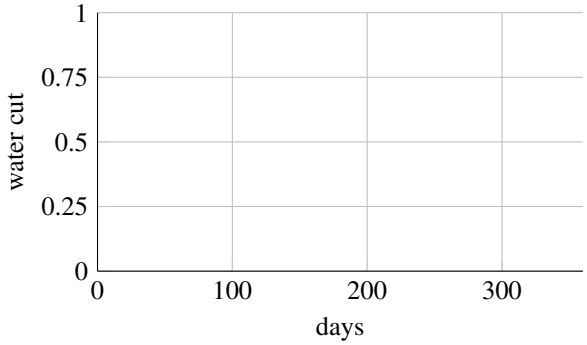

FIGURE 5 Two-phase flow water cut simulated on LUNDI$_{sim}$ (full resolution, nearshore$_0$ environment) and measured at P_1 according to the configuration of Figure 3 (quarter five-spot model). Two-phase flow simulated water cut is measured at P_1. The red dot indicates the time (in days) of the water breakthrough, i.e. when water starts being produced at well P_1.

Linux). Other Python libraries support GRDECL format, for instance *PyGRDECL*[##] or *XTGeo*[‖‖], and can be also used for visualization or other processings.

7 | POTENTIAL DATASET USE/REUSE

Inspired by the SPE10 simulation challenge[31], LUNDI$_{sim}$ with its different environments are primarily meant for evaluating the performance of lossy or lossless compression algorithms with respect to reservoir modeling and simulation. Openly-shared models are scarce in reservoir geoscience and engineering. LUNDI$_{sim}$ serves other purposes as well.

It can be used to test more geologically-oriented upscaling methods and their reliability regarding information loss, through quality indicators[33]. While initially developed for hydrocarbons, our approach may conceptually be used for more sustainable projects, for instance geothermy, hydrogen (H_2) or carbon dioxide (CO_2[21]) storage projects. Note that the Society of Petroleum Engineers has just released a call on the 11th SPE challenge for safe and efficient implementation of geological carbon storage.

Being complex volume meshes, LUNDI$_{sim}$ models can be used to benchmark scientific data compression algorithms. They are also adapted to investigate the impact of reduced data precision[37] or resolution change on pure objective metrics (for instance in a context of mesh visualization, storage or checkpoint restart), but also on faithfulness of any simulation.

As for precision, current practice favors the IEEE 754 floating-point format — in double, quadruple or even octuple precision[38] — to ensure both accuracy and simplicity of data management. As a result, some data fields are represented, stored and transferred with an excessive number of bits[12]. Plus, it is being recognized that for a given simulation workflow, quantities from an homogeneous data field may possess widely different statistical distributions, in which distinct scales of magnitude are associated to different spread/precision/impact. For instance, a permeability value of zero or below 50 md means "no to meaningless" water flows (rocks working as "seals"), while values greater by orders of magnitude (over 10 000 md) may yield "full permeation". As a consequence, a fine precision for small permeabilities is meaningful, when higher permeability values would not affect results when changed by $\pm 20\,\%$. As HPC sparks interest on so-called next-generation arithmetic (such as UNUM or POSIT formats[39,40]), with so few simulation tools already adapted to such hybrid data formats, it is important to be able to emulate them on shared and representative dataset with minimal, reproducible examples of workflows.

As for resolution, with edge computing, or the necessity sometimes to assess crude estimations in real time[41] on low-power devices using cloud resources, it becomes increasing important to provide users data with adapted granularity. One straightforward scheme consists in sharing the original data source as well as several lower-resolution versions, either with pyramid schemes[29] or with embedded multiresolution mechanisms, for instance with wavelets[42,43] as in[25]. For this reason, we provide LUNDI$_{sim}$ models with their associated lower resolution representations. The latter may also be probed with varying precision, as mentioned previously. Evaluating the combined impact of resolution and precision is briefly evoked in[35], and the

[##] https://github.com/BinWang0213/PyGRDECL
[‖‖] https://pypi.org/project/xtgeo

topic for a forthcoming companion paper. Additional reuse cases reside in combining simulation and compression with machine learning or artificial intelligence tools, which are being used more intensively in simulation[44].

Future research may be interested in larger-size models than those we share here. By providing here the main ingredients and philosophy used to build LUNDI$_{sim}$ models, we hope they will help in creating novel meshes along our open methodological guidelines.

8 | CONCLUSION

A couple of years ago, due to the lack of openly shared heterogeneous and realistic geoscience data to study influence of compression on simulation workflows, we designed our own models, inspired by the SPE10 challenge. For other researchers on this field to overcome this pitfall, we now share our models named LUNDI$_{sim}$ to the scientific community in the FAIR spirit. Based on a typical geoscientific mesh containing several faults, and two formations proposed in SPE10, we generated four models with distinct environments, including porosity and permeability information. Thanks to the multiresolution Hexa-Shrink framework, our dataset also includes lower-resolution versions of each model (mesh and attributes), with consistent fault preservation whatever the level of decomposition. We hope that this dataset will be useful to other geoscience researchers in taking their projects forward.

ACKNOWLEDGMENTS

The research presented was mainly performed during the PhD thesis of Lauriane Bouard, following the post-doctoral position of Jean-Luc Peyrot and the internship of Lenaïc Chizat. The authors are grateful to IFP Energies nouvelles for the permission to share LUNDI$_{sim}$ models. They acknowledge the support of AIR (Action IUT Recherche) of IUT Côte d'Azur. They thank Christophe Latry, Carole Thiébaut (CNES), Laurent Astart, Nadine Couëdel, Frédéric Douarche, Thomas Guignon (IFP Energies nouvelles), Corinne Maihles (TéSA) and Marc Antonini (UniCA, CNRS) for their support.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

None.

References

1. Hey T, Tansley S, Tolle K., eds.*The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

2. Ahrens J. Technology Trends and Challenges for Large-Scale Scientific Visualization. *IEEE Comput. Graph. Appl.*. 2022;42(4):114–119. doi: 10.1109/mcg.2022.3176325

3. Sarton J, Zellmann S, Demirci S, et al. State-of-the-art in Large-Scale Volume Visualization Beyond Structured Data. *Comput. Graph. Forum.* 2023;42(3):491–515. doi: 10.1111/cgf.14857

4. Overpeck JT, Meehl GA, Bony S, Easterling DR. Climate Data Challenges in the 21st Century. *Science.* 2011;331:700–702. doi: 10.1126/science.1197869

5. Yakushin I, Mehta K, Chen J, et al. Feature-preserving Lossy Compression for In Situ Data Analysis. In: Proc. Int. Conf. Parallel Processing. 2020

6. Mittal S. A Survey of Techniques for Approximate Computing. *ACM Comput. Surv.*. 2016;48(4):1–33. doi: 10.1145/2893356

7. Magri VAP, Lindstrom P. A General Framework for Progressive Data Compression and Retrieval. *IEEE Trans. Visual Comput. Graph.*. 2024;30:1358–1368. doi: 10.1109/tvcg.2023.3327186

8. Childs H, Ahern SD, Ahrens J, et al. A terminology for *in situ* visualization and analysis systems. *Int. J. High Perform. Comput. Appl.*. 2020;34(6):676—691. doi: 10.1177/1094342020935991

9. Schweiger G, Nilsson H, Schoeggl J, Birk W, Posch A. Modeling and simulation of large-scale systems: A systematic comparison of modeling paradigms. *Appl. Math. Comput.*. 2020;365:124713. doi: 10.1016/j.amc.2019.124713

10. Murillo R, Del Barrio AA, Botella G. The effects of numerical precision in scientific applications. In: Proc. Annual Modeling and Simulation Conference. 2002.

11. Taurone F, Lucani DE, Fehér M, Zhang Q. Lossless Preprocessing of Floating Point Data to Enhance Compression. In: Distributed Computing and Artificial Intelligence, Special Sessions I, 20th International Conference, Lecture Notes in Networks and Systems, Springer, 2023:457–466

12. Walters MS, Wong DC. The impact of altering emission data precision on compression efficiency and accuracy of simulations of the community multiscale air quality model. *Geosci. Model Dev.*. 2023;16(4):1179–1190. doi: 10.5194/gmd-16-1179-2023

13. Wang D, Pulido J, Grosset P, et al. TAC+: Optimizing Error-Bounded Lossy Compression for 3D AMR Simulations. *IEEE Trans. Parallel Distrib. Syst.*. 2024;35(3):421–438. doi: 10.1109/tpds.2023.3339474

14. Liang X, Whitney B, Chen J, et al. MGARD+: Optimizing Multilevel Methods for Error-Bounded Scientific Data Reduction. *IEEE Trans. Comput.*. 2022;71(7):1522–1536. doi: 10.1109/tc.2021.3092201

15. Liu J, Tian J, Wu S, et al. cuSZ-I: High-Fidelity Error-Bounded Lossy Compression for Scientific Data on GPUs. *PREPRINT.* 2024. doi: 10.48550/ARXIV.2312.05492

16. Klöwer M, Razinger M, Dominguez JJ, Düben PD, Palmer TN. Compressing atmospheric data into its real information content. *Nat. Comput. Sci.*. 2021;1(11):713–724. doi: 10.1038/s43588-021-00156-2

17. Underwood R, Bessac J, Krasowska D, Calhoun JC, Di S, Cappello F. Black-box statistical prediction of lossy compression ratios for scientific data. *Int. J. High Perform. Comput. Appl.*. 2023;37(3–4):412–433. doi: 10.1177/10943420231179417

18. Baker AH, Hammerling DM, Mickelson SA, et al. Evaluating lossy data compression on climate simulation data within a large ensemble. *Geosci. Model Dev.*. 2016;9(12):4381–4403. doi: 10.5194/gmd-9-4381-2016

19. Poppick A, Nardi J, Feldman N, Baker AH, Pinard A, Hammerling DM. A statistical analysis of lossily compressed climate model data. *Comput. Geosci.*. 2020;145:104599. doi: 10.1016/j.cageo.2020.104599

20. Peters K, Höck H, Thiemann H. FAIR long term preservation of climate and Earth System Science data with a focus on reusability at the World Data Center for Climate (WDCC). *Authorea.* 2020. doi: 10.1002/essoar.10501879.1

21. Alumbaugh D, Gasperikova E, Crandall D, et al. The Kimberlina synthetic multiphysics dataset for CO2 monitoring investigations. *Geosci. Data J.*. 2023. doi: 10.1002/gdj3.191

22. Haehnel P, Freund H, Greskowiak J, Massmann G. Development of a three-dimensional hydrogeological model for the island of Norderney (Germany) using GemPy. *Geosci. Data J.*. 2023. doi: 10.1002/gdj3.208

23. Bouard L. *Refinable resolution and precision for volume mesh compression and simulation in geosciences*. PhD thesis. Université Côte d'Azur, France; 2021.

24. Wellmann F, Caumon G. 3-D Structural geological models: Concepts, methods, and uncertainties. In: Advances in Geophysics. 59. Elsevier, 2018:1–121

25. Peyrot JL, Duval L, Payan F, et al. HexaShrink, an exact scalable framework for hexahedral meshes with attributes and discontinuities: multiresolution rendering and storage of geoscience models. *Computat. Geosci.*. 2019;23:723–743. doi: 10.1007/s10596-019-9816-2

26. Suter E, Kårstad T, Escalona A, Friis HA, Vefring EH. Principles for a Hierarchical Earth Model Representation Aiming Towards Fit-For-Purpose Grid Resolution. In: Proc. EAGE Conf. Tech. Exhib. EAGE 2019

27. Abraham F, Celes W. Multiresolution visualization of massive black oil reservoir models. *Vis. Comput.*. 2019;35:837–848. doi: 10.1007/s00371-019-01674-x

28. Devarajan H, Kougkas A, Logan L, Sun XH. HCompress: Hierarchical Data Compression for Multi-Tiered Storage Environments. In: IEEE International Parallel and Distributed Processing Symposium. 2020.

29. Ceballos L, Conche B, Dupuy G, Patel D. Visualization of Large Scale Reservoir Models. In: Patel D., ed. *Interactive Data Processing and 3D Visualization of the Solid Earth*, , Springer, 2021:209–232

30. Salomon D, Motta G. *Handbook of Data Compression*. Springer, 2009.

31. Christie MA, Blunt MJ. Tenth SPE Comparative Solution Project: A Comparison of Upscaling Techniques. *SPE Reserv. Eval. Eng.*. 2001;4:308–317. doi: 10.2118/66599-ms

32. King MJ. Recent Advances in Upgridding. *Oil Gas Sci. Tech.*. 2007;62(2):195–205. doi: 10.2516/ogst:2007017

33. Preux C. About the Use of Quality Indicators to Reduce Information Loss When Performing Upscaling. *Oil Gas Sci. Tech.*. 2014;71(1). doi: 10.2516/ogst/2014023

34. Misaghian N, Assareh M, Sadeghi M. An upscaling approach using adaptive multi-resolution upgridding and automated relative permeability adjustment. *Computat. Geosci.*. 2018;22:261–282. doi: 10.1007/s10596-017-9688-2

35. Bouard L, Duval L, Payan F, Preux C, Antonini M. Étude comparative de l'impact d'un codage à précision variable sur des données de simulation en géosciences. In: Proc. colloque COmpression et REprésentation des Signaux Audiovisuels (CORESA). 2021.

36. Lie KA. *An introduction to reservoir simulation using MATLAB/GNU Octave*. Cambridge, United Kingdom: Cambridge University Press, 2019. Gesehen am 14.05.2020.

37. Moreland K, Pugmire D, Chen J. The Exploitation of Data Reduction for Visualization. Tech. Rep. ORNL/LTR-2022/412, Oak Ridge National Laboratory; Oak Ridge National Laboratory: 2022.

38. Gladman B, Innocente V, Mather J, Zimmermann P. Accuracy of Mathematical Functions in Single, Double, Double Extended, and Quadruple Precision. tech. rep., LORIA; 2024.

39. Dinechin dF, Forget L, Muller JM, Uguen Y. Posits: the good, the bad and the ugly. In: Proc. Conf. Next Generation Arithmetic (CoNGA). 2019.

40. Lindstrom P. MultiPosits: Universal Coding of $\mathbb{R}^n$. In: Proc. Conf. Next Generation Arithmetic (CoNGA). Springer 2022:66–83

41. Sicat R, Ibrahim M, Ageeli A, Mannuss F, Rautek P, Hadwiger M. Real-Time Visualization of Large-Scale Geological Models with Nonlinear Feature-Preserving Levels of Detail. *IEEE Trans. Visual Comput. Graph.*. 2023;29(2):1491–1505. doi: 10.1109/tvcg.2021.3120372

42. Christophe E, Mailhes C, Duhamel P. Hyperspectral Image Compression: Adapting SPIHT and EZW to Anisotropic 3-D Wavelet Coding. *IEEE Trans. Image Process.*. 2008;17(12):2334–2346. doi: 10.1109/tip.2008.2005824

43. Jacques L, Duval L, Chaux C, Peyré G. A panorama on multiscale geometric representations, intertwining spatial, directional and frequency selectivity. *Signal Process.*. 2011;91(12):2699–2730. doi: 10.1016/j.sigpro.2011.04.025

44. Glaws A, King R, Sprague M. Deep learning for *in situ* data compression of large turbulent flow simulations. *Phys. Rev. Fluids.* 2020;5(11):114602. doi: 10.1103/physrevfluids.5.114602

45. Duval L, Payan F, Preux C, Bouard L. How do reduced resolution/precision and companding on permeability data affect flow simulation? Benchmarks on LUNDIsim models. *PREPRINT.* 2024.

SUPPORTING INFORMATION

A companion paper[45] about the impact of reduced resolution/precision and companding with HexaShrink on the proposed simulation is planned (partly presented in[35]).