Geophysical Journal International

Geophys. J. Int. (2026) **247,** 1–7
https://doi.org/10.1093/gji/ggag295
Advance Access publication 2026 August 08
Express Letter

Massive-scale unlabelled field and labelled synthetic seismic data sets of global shelf-edge clinothems

Hui Gao,[1,2] Xinming Wu ,[2] Jintao Li ,[3] Xiaoming Sun[4] and Jiarun Yang[1]

[1] *School of Artificial Intelligence, Anhui University of Science and Technology, Hefei 231131, China*
[2] *School of Earth and Space Sciences, State Key Laboratory of Precision Geodesy, Mengcheng National Geophysical Observatory, University of Science and Technology of China, Hefei 230026, China. E-mail: xinmwu@ustc.edu.cn*
[3] *State Key Laboratory of Ocean Sensing & Ocean College, Zhejiang University, Zhoushan, Zhejiang 316021, China*
[4] *Institute of Advanced Technology, University of Science and Technology of China, Hefei 230088, China*

Accepted 2026 July 21. Received 2026 June 21; in original form 2026 April 20

SUMMARY

Seismic stratigraphic interpretation of shelf-edge clinothems is essential for revealing tectonic evolution, paleoclimate change, depositional dynamic conditions and hydrocarbon generation and accumulation during basin filling. However, traditional interpretation methods remain labour-intensive, time-consuming and highly subjective. Although AI-based method offer a potential solution for automating this task, its development has been limited by the scarcity of comprehensive and representative benchmark data sets for shelf-edge clinothems. This limitation primarily arises from limited field data availability, the scarcity of reliable geological labels and the structural complexity and strong variability of clinothem-dominated systems. To address this gap, we develop a hybrid benchmark data set through two complementary strategies of field data curation and geological and geophysical forward modelling, ultimately generating 3000 unlabelled field and 4000 labelled synthetic seismic data, respectively. We further evaluate several representative baseline deep learning models on these data sets, and the accurate results demonstrate that the curated data set provides an effective and representative basis for model training, quantitative assessment and practical application. Finally, we have publicly released this hybrid benchmark data set (https://doi.org/10.5281/zenodo.20769762,) to facilitate the development, validation and assessment of deep learning methods for automated seismic stratigraphic interpretation.

Key words: Image processing; Machine learning; Sedimentary basin processes.

1 INTRODUCTION

Clinothem-dominated systems exhibit substantial variability in geometry, stacking pattern and geological structures across different sedimentary basins (I. Anell & I. Midtkandal 2017; S. Patruno & W. Helland-Hansen 2018), making manual interpretation difficult and posing a major challenge for developing robust and generalizable AI-based interpretation methods. Recently, several geophysical benchmark data sets, such as STanford EArthquake Dataset (STEAD, S.M. Mousavi *et al.* 2019), OpenFWI (C. Deng *et al.* 2022), cigChannels (G. Wang *et al.* 2024), cigFacies (H. Gao *et al.* 2025b), OpenSWI (F. Liu *et al.* 2025) and GeoFWI (C. Li *et al.* 2026), have been developed to advance artificial intelligence (AI) methods in various geophysical scenarios. However, such comprehensive and representative benchmark data set remain highly scarce for shelf-edge clinothems of sedimentary basin. Moreover, most existing publicly available data sets of exploration geophysics primarily consist of synthetic data, while field data sets are typically limited in scale, lack geological labels and are mainly utilized for testing and validation. Although synthetic data sets have proven effective for model training and have been successfully applied in many geophysical tasks (X. Wu *et al.* 2020; I. Ferreira & A. Koeshidayatullah 2024), relying solely on synthetic data is insufficient for networks to fully capture the complexity and variability of field clinothems systems. This limitation is particularly pronounced in sedimentary basins, where numerous depositional hiatuses, erosional surfaces, unconformities, complex sequence architectures and tectonic deformation make realistic forward modelling extremely challenging. Moreover, recent advances in self-supervised learning and large-scale pretraining have further increased the demand for diverse unlabelled field data, which provide realistic structural patterns and geological variability that are difficult to realistically simulate in synthetic data. In addition, field data sets spanning multiple basins are indispensable for evaluating model effectiveness, robustness and generalization across diverse geological settings. Therefore, constructing a hybrid benchmark data set that combines realistic field characteristics with high-quality geological labels is essential for effective modelling training, standardized baseline evaluation and improved generalization in practical field applications.

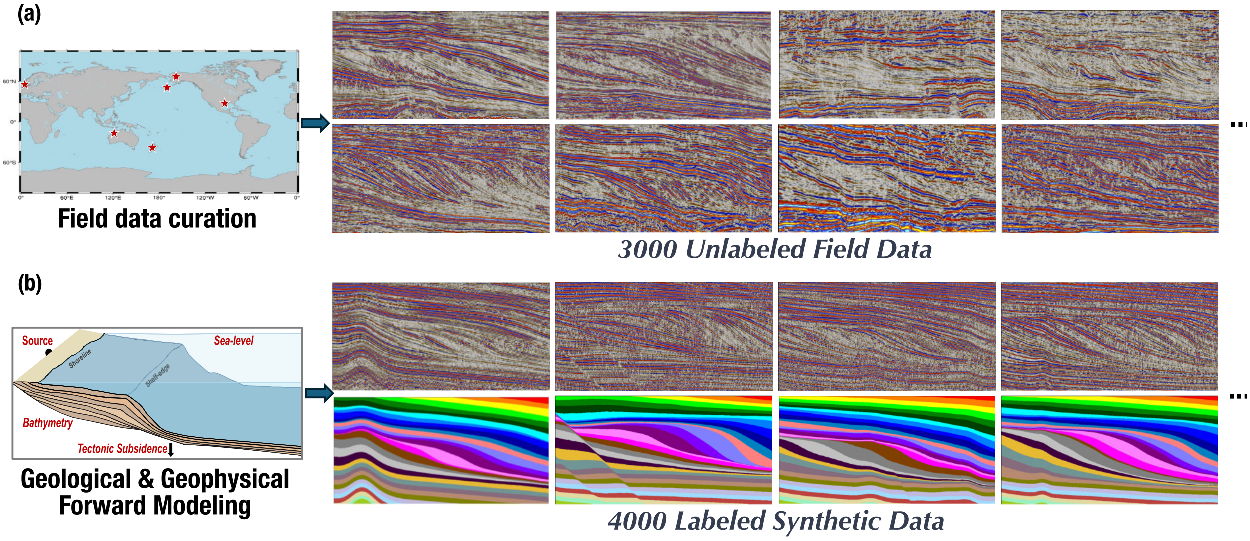

Figure 1. Construction of the unlabelled field and labelled synthetic seismic data sets. (a) We construct the unlabelled field data set through 2-D and 3-D raw data collection, manual selection, cropping, resampling and data post-processing, resulting in 3000 normalized 2-D field seismic images with diverse clinothems architectures. (b) We then perform stratigraphic forward modelling, geological structural deformation and geophysical forward modelling to generate various realistic synthetic seismic data sets and corresponding RGT labels.

To address these limitations, we construct a hybrid benchmark data set consisting of an unlabelled field seismic data set derived from field data curation and a labelled synthetic seismic data set generated from geological and geophysical forward modelling (see Fig. 1). Specifically, we curate publicly available field data through collection, manual selection, cropping, resampling and standardization, ultimately building 3000 unlabelled samples that preserve the realism, diversity and structural complexity of clinothem-dominated systems. We then generate a labelled synthetic seismic data set through a three-step workflow of stratigraphic forward modelling, geological structural deformation and geophysical forward modelling, ultimately generating 4000 geologically plausible synthetic samples with corresponding relative geologic time (RGT; T.J. Stark 2003, 2004) labels. These RGT labels provide a dense stratigraphic representation that preserves the full stratigraphic succession and relative temporal relationships, which is critical for characterizing clinothem architectures and supporting seismic stratigraphic interpretation. The field data set provides realistic benchmark samples for field testing and application, whereas the synthetic data set provides geologically reasonable and label-complete samples for supervised training and quantitative evaluation. Finally, we further evaluate several representative baseline models on these data sets to demonstrate their comprehensiveness, representativeness, geological plausibility and effectiveness for model training and assessment.

2 BENCHMARK DATA SET CONSTRUCTION

RGT estimation is a continuous regression task that is highly sensitive to stratigraphic continuity and structural consistency. Its performance depends not only on network architecture and optimization strategy, but also critically on the representativeness of the training data and the quality of the target labels. In this study, we develop two complementary strategies to construct massive-scale training data sets with diverse clinothems architectures shown in Fig. 1. The first strategy is to build 3000 unlabelled field seismic data through field data curation (Fig. 1a). This field data set preserves the realism, diversity, structural complexity and representativeness of field clinothem-dominated systems, thereby providing an essential basis for model evaluation and field application. The second strategy is to build 4000 labelled synthetic seismic data through geological and geophysical forward modelling (Fig. 1b). This synthetic data set provides geologically consistent samples with corresponding RGT labels, enabling the model to learn a stable mapping from seismic data to RGT.

2.1 Unlabelled field seismic data sets

To construct a large-scale field data set containing diverse clinothem architectures across different sedimentary basins, we develop a systematic field data curation workflow consisting of raw data collection, manual selection and cropping, resampling and data post-processing. Specifically, we first collect more than 600 3-D seismic volumes and 2000 2-D seismic profiles from multiple publicly available data set repositories, such as the Society of Exploration Geophysicists (SEG), Netherlands Oil and Gas Portal (NLOG), New Zealand Petroleum & Minerals (NZPAM), the United States Geological Survey (USGS), the South Australian Resources Information Gateway (SARIG) and Western Australian Petroleum and Geothermal Information Management System (WAPIMS). We then manually select and crop these data sets to retain approximately 60 3-D seismic volumes and 500 2-D seismic profiles that exhibit

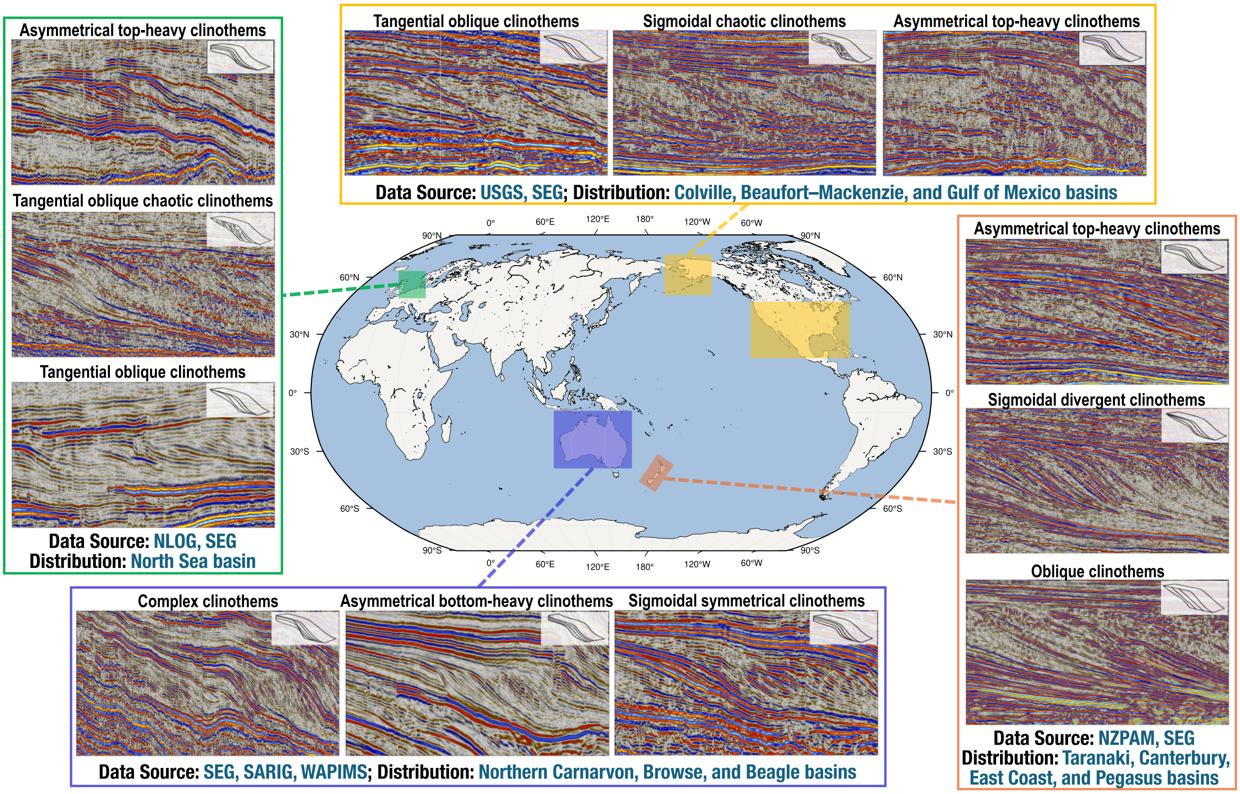

Figure 2. Spatial distribution and representative patterns of field seismic clinothem data set. These field samples cover several representative sedimentary basins worldwide and contain diverse clinothems geometric types. Some representative field samples shown in the field-data curation workflow in Fig. 1(a) are presented again here to facilitate the illustration of their spatial distributions and corresponding geometric types. These geometric cartoon diagrams are modified from I. Anell & I. Midtkandal (2017).

Table 1. A summary of the field and synthetic seismic data sets.

Data set	Category	Number	Total
Field	North Sea Basin	1982	3000
Data set	New Zealand East Coast, Raukumara, Pegasus, Taranaki, Canterbury Basins	596	
	Australia Northern Carnarvon, Browse, Beagle Basins	180	
	Gulf of Mexico, Colville, Beaufort–Mackenzie Basins	242	
Synthetic	One third-order sea-level cycle	780	4000
Dataset	Two third-order sea-level cycles	1370	
	Three third-order sea-level cycles	1714	
	Four third-order sea-level cycles	136	

representative clinothems geometries. These selected field data are primarily distributed across the Dutch North Sea, Australia, New Zealand, Alaska North Slope, the Beaufort Sea and Gulf of Mexico, encompassing diverse sedimentary environments and stratigraphic patterns (see Fig. 2). We further randomly slice the 3-D volumes into 2-D profiles from inline or crossline directions with at least 25 profiles between adjacent slices to reduce redundancy. Finally, we perform resampling and mean–variance normalization process to each profile, yielding 3000 unlabelled 2-D field seismic data, as illustrated in Figs 1(a) and 2. _Some representative samples used to illustrate the field-data curation workflow in Fig. 1(a) are presented again in Fig. 2, where their spatial distributions and corresponding geometric types are further illustrated.

The field clinothem data set provides broad coverage of sedimentary settings and clinothem geometries, highlighting its diversity, complexity and representativeness. In terms of sedimentary environments, such data sets cover multiple representative basin surveys, including North Sea, Northern Carnarvon, Browse, Beagle, New Zealand East Coast, Raukumara, Pegasus, Taranaki, Canterbury, Gulf of Mexico, Colville, Beaufort–Mackenzie basins (Table 1), spanning a wide range of sedimentary dynamic conditions, long-term tectonic and climate changes and depositional environments. In terms of clinothem geometric complexity, these data sets contain a variety of typical clinothem architectures (Fig. 2), such as oblique, tangential oblique, tangential oblique chaotic, asymmetric top-heavy, asymmetric bottom-heavy, sigmoidal symmetric, sigmoidal divergent, sigmoidal chaotic, complex clinothems types (I. Anell & I. Midtkandal 2017). These geometric variations reflect diverse stratigraphic stacking patterns formed under different conditions of sediment supply,

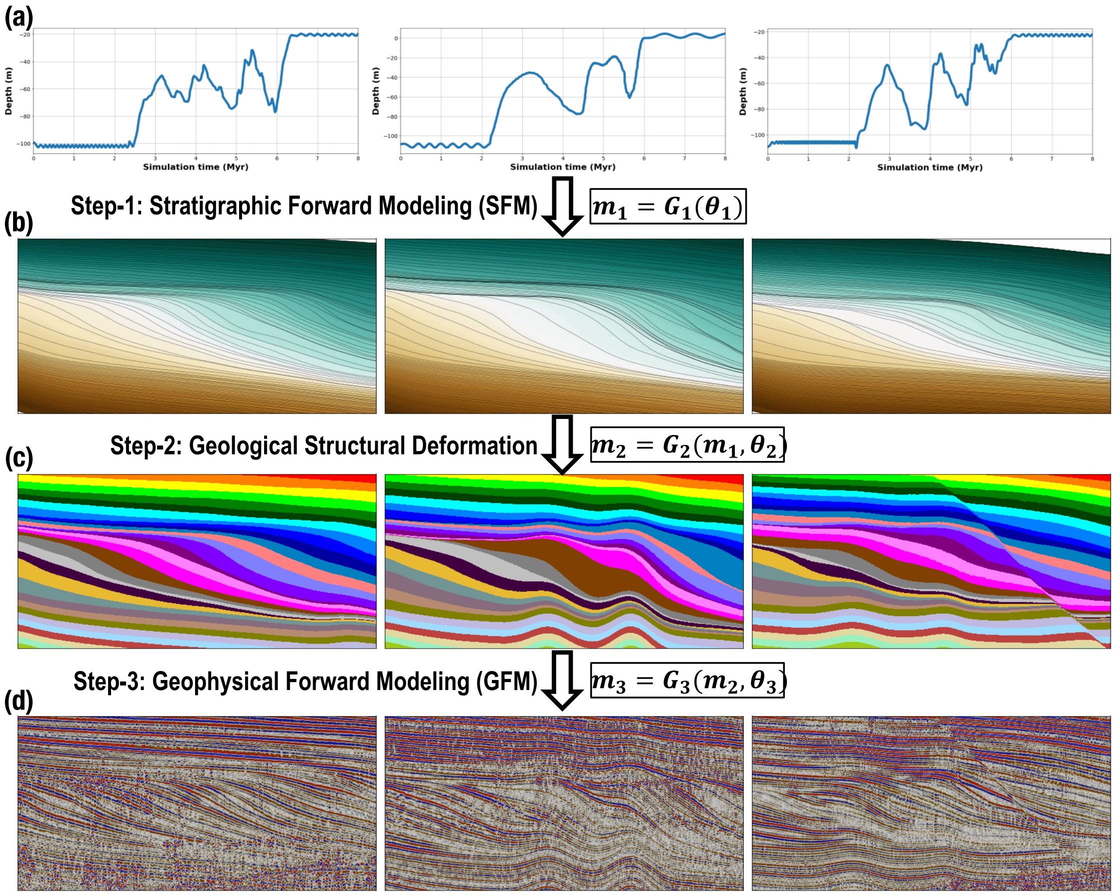

Figure 3. Geological and geophysical forward modelling workflow for constructing labelled synthetic seismic data set. (a) Relative sea level curves, (b) stratigraphic layers simulating from stratigraphic forward modelling, (c) geological models after geological structural deformation and (d) synthetic seismic images generated from geophysical forward modelling.

accommodation, tectonic processes and climate changes. Although this data set does not contain RGT labels, it preserves the geometric variability and realistic stratigraphic and structural features, thus providing a valuable and comprehensive benchmark and application data set for assessing model performance across different sedimentary basin. Besides, this unlabelled field data set can be used for unsupervised or self-supervised learning, enabling networks to learn representative field stratigraphic and structural features from diverse geological settings, thus potentially reducing the domain gap between synthetic and field data.

2.2 Labelled synthetic seismic data sets

To address the scarcity of high-quality RGT labels in field seismic data, we further develop a comprehensive and geologically plausible three-step forward modelling workflow modified from H. Gao *et al.* (2025a) to generate realistic synthetic seismic data and corresponding labels. This workflow consists of stratigraphic forward modelling, geological structural deformation and geophysical forward modelling. We first simulate sediment transport and depositional evolution in fluvial–deltaic systems using sedimentary process simulation, and generate various shelf-edge stratigraphic clinoforms model ($\boldsymbol{m}_1$) with realistic and diverse stratigraphic stacking patterns and complex geometric architectures shown in Fig. 3(b). This process can be expressed as

$$\boldsymbol{m}_1 = \boldsymbol{G}_1(\theta_1), \tag{1}$$

where $\boldsymbol{G}_1$ represents the stratigraphic forward modelling operator implemented with a landscape evolution model (T. Salles *et al.* 2018), and θ_1 denotes a set of geologically reasonable forcing conditions, such as initial bathymetry, river discharge and distribution, sediment flux, sea-level fluctuations (Fig. 3a and Table 1), tectonic subsidence and climate changes. To further improve the realism and diversity of geological models under complex post-depositional geological structural deformations, we then add faulting and folding structures into stratigraphic models ($\boldsymbol{m}_1$) to simulate realistic subsurface geological models ($\boldsymbol{m}_2$) during post-depositional burial (Fig. 3c). This deformation process is formulated as:

$$\boldsymbol{m}_2 = \boldsymbol{G}_2(\boldsymbol{m}_1, \theta_2), \tag{2}$$

where $\boldsymbol{G}_2$ represents the operator of simulating geological structural deformation proposed by X. Wu *et al.* (2020), and θ_2 denotes a set of parameters controlling fault and fold geometries, such as fault spatial distributions, strikes, dips and throws, as well as fold locations, extents and shapes. To simulate realistic fault characteristics within field shelf-edge regions, we restrict its strike to primarily subparallel with the basin margin, thus aligning with the typical normal fault systems commonly developed in extensional tectonic environments. Finally, we perform geophysical forward modelling ($\boldsymbol{G}_3$) to generate the corresponding realistic synthetic seismic data ($\boldsymbol{d}$), which is given by

$$\boldsymbol{d} = \boldsymbol{G}_3(\boldsymbol{m}_2, \theta_3), \tag{3}$$

where $\boldsymbol{G}_3$ represents geophysical forward modelling operator, including realistic impedance model construction, depth-time conversion, wavelet convolution and real or random noise addition, and θ_3 denotes a set of parameters controlling the geophysical properties, wavelet type and dominant frequency, noise characteristics and signal-to-noise ratio. Through geological and geophysical forward modelling workflow, we ultimately generate 4000 synthetic seismic data and corresponding RGT labels, as shown in Figs 1(b) and 3(d). Such synthetic data set maintains both geological plausibility and physical consistency while providing accurate and high-quality RGT labels, thereby establishing a reliable supervision foundation for AI models to learn a stable seismic-to-RGT mapping.

3 BASELINE MODEL EVALUATIONS

To evaluate the effectiveness of the curated benchmark data sets, we further perform a comprehensive performance evaluation using several representative deep learning models, including U-Net (O. Ronneberger *et al.* 2015), High-Resolution Network (HRNet; J. Wang *et al.* 2020), LocalBins (S.F. Bhat *et al.* 2022) and Global-Local Path Networks (GLP; D. Kim *et al.* 2022). These baseline models are trained and evaluated within a geologically informed and data-driven framework modified from H. Gao *et al.* (2025a), which integrates synthetic supervision and geological prior constraints for training. Specifically, the 4000 labelled synthetic samples are split into 3200 training and 800 validation samples, and all four models are trained for 400 epochs using the hybrid loss proposed in (H. Gao *et al.* 2025a), which combines mean square error, multiscale structural similarity (Z. Wang *et al.* 2003, 2004), IsochronLoss and NormalLoss.

We then select six samples from field data set and apply the trained baseline models for RGT estimation, and extract multiple iso-surfaces at equal geologic time intervals, as illustrated in Fig. 4. Some of these selected samples, which exhibit representative clinothem geometries and stratigraphic characteristics, are also shown in Figs 1(a) and 2 and are presented again here as typical cases for comparing baseline-model predictions. To further quantitatively evaluate model performance on the field data set without labels, we further compute two unsupervised geologically informed metrics, $\mathcal{M}_{\mathrm{Isochron}}$ and $\mathcal{M}_{\mathrm{Normal}}$ (H. Gao *et al.* 2025a), which respectively assess the isochronism consistency along locally tracked horizon segments and the structural consistency between the predicted RGT gradient and the seismic normal vector, with lower values indicating better performance. The extracted iso-surfaces exhibit strongly alignment with the seismic reflectors and provide a reliable representation of clinothems geometries, stratigraphic stacking patterns, geological structues and unconformities distribution across various field surveys.

These precise results demonstrate that the curated benchmark data set provides an effective and representative basis for model training, practical application and quantitative evaluation. However, some fine-scale weak stratigraphic fittings are still observed in local regions with stratigraphic pinch-out, unconformities, low signal-to-noise ratios and complex tectonic structures, as indicated by the orange arrows in Fig. 4. Besides, the performance comparison reveals that transformer-based networks (GLP and LocalBins) and multiscale feature-fusion network (HRNet) outperform the simpler convolutional U-Net. This performance gap highlights the critical importance of extracting and sensing global features for RGT estimation in shelf-edge clinothem data, which commonly contain numerous depositional hiatuses and stratigraphic discontinuities.

4 CONCLUSION

In this paper, we construct a comprehensive hybrid benchmark data set specifically designed for shelf-edge clinothems, comprising 3000 unlabelled field seismic samples and 4000 labelled synthetic seismic samples with corresponding RGT labels. Baseline model evaluations demonstrate that these two complementary benchmark data sets provide an effective and representative foundation for model training, field application, quantitative evaluation, development and deployment of AI method for seismic stratigraphic interpretation in clinothem-dominated systems. Finally, we have made these comprehensive clinothem data sets and representative deep learning baseline models publicly available (H. Gao *et al.* 2026), facilitating further model development, comparison and validation for high-resolution seismic stratigraphic interpretation of sedimentary basins.

ACKNOWLEDGMENTS

This research was supported the National Science Foundation of China (grant no. 42374127), China Postdoctoral Science Foundation (grant no. 2026M790497), and Postdoctoral Fellowship Program of CPSF (grant no. GZB20260111). The authors thank SEG, NLOG, NZPAM, USGS, SARIG and WAPIMS for providing public available data sets. We also thank the USTC supercomputing center for providing computational resources.

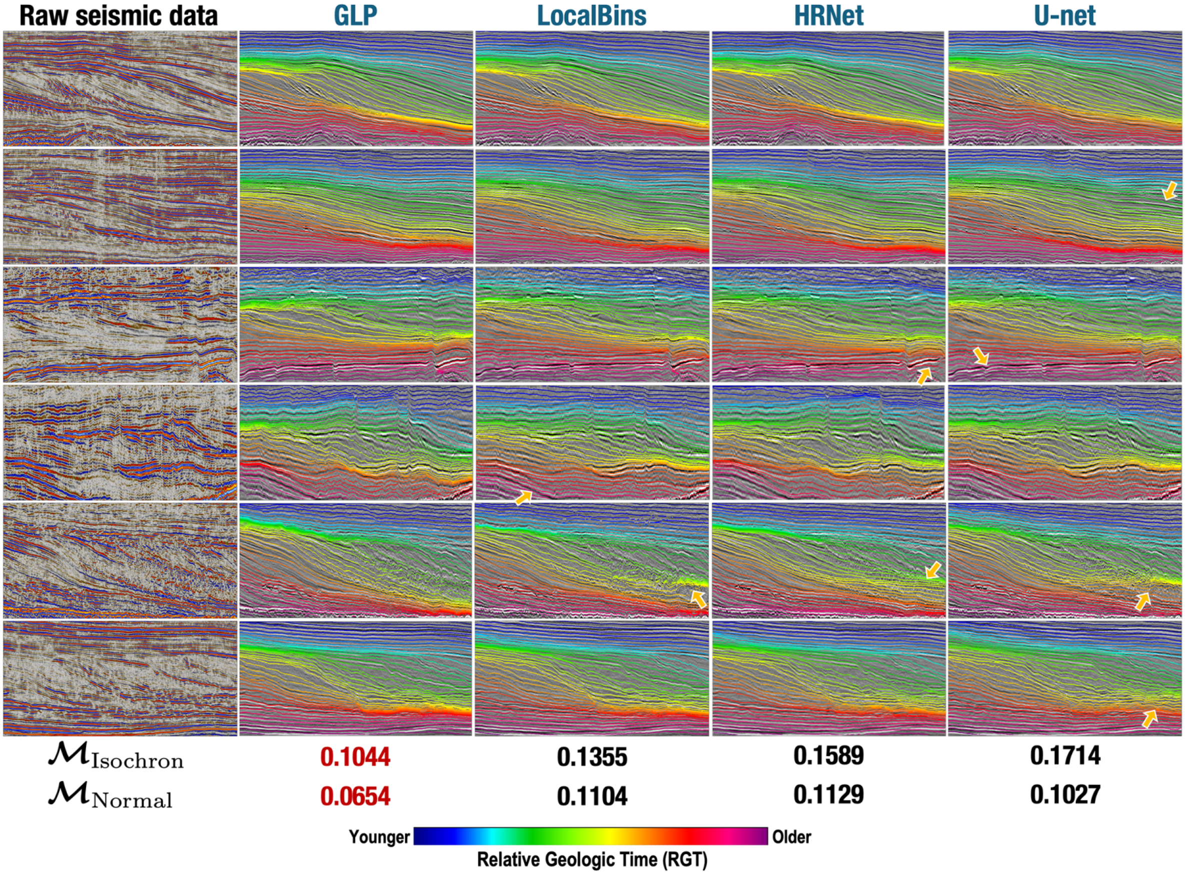

Figure 4. Comparison of RGT predicted results across different baseline models (GLP, LocalBins, HRNet and U-Net). Multiple iso-surfaces are extracted from the RGT predictions and overlaid on the seismic image. Some of these selected samples, which exhibit representative clinothem geometries and stratigraphic characteristics, are also shown in Figs 1(a) and 2 and are presented again here as typical cases for comparing baseline-model predictions. The two unsupervised geologically informed metrics ($\mathcal{M}_{\text{Isochron}}$ and $\mathcal{M}_{\text{Normal}}$) (H. Gao *et al.* 2025a), are computed on field samples to quantitatively evaluate model performance, where lower values indicate better stratigraphic isochronism and structural consistency, respectively.

DATA AVAILABILITY

The 3000 unlabelled field data and 4000 labelled synthetic data of global shelf-edge clinothems has been uploaded to Zenodo and is freely available at https://doi.org/10.5281/zenodo.20769762 (H. Gao *et al.* 2026).

REFERENCES

Anell, I. & Midtkandal, I., 2017. The quantifiable clinothem–types, shapes and geometric relationships in the plio-pleistocene giant foresets formation, taranaki basin, new zealand, *Basin Res.,* **29**(S1), 277–297.

Bhat, S.F., Alhashim, I. & Wonka, P., 2022. Localbins: improving depth estimation by learning local distributions, *European Conference on Computer Vision,* pp. 480–496, Springer-Verlag.

Deng, C. *et al.*, 2022. Openfwi: large-scale multi-structural benchmark datasets for full waveform inversion, *Adv. Neural Inf. Process. Syst.,* **35,** 6007–6020.

Ferreira, I. & Koeshidayatullah, A., 2024. Synthetic data in geosciences: challenges and opportunities, *85th EAGE Annual Conference and Exhibition, Vol. 2024,* pp. 1–5, European Association of Geoscientists and Engineers.

Gao, H., Wu, X. & Ding, X., 2025a. A geologically-informed and data-driven ai workflow for fully seismic stratigraphic interpretation of sedimentary basin, *IEEE Trans. Geosci. Remote Sens*. **63,** 1–13.

Gao, H., Wu, X., Sun, X., Hou, M., Gao, H., Wang, G. & Sheng, H., 2025b. cigfacies: a massive-scale benchmark dataset of seismic facies and its application, *Earth Syst. Sci. Data,* **17**(2), 595–609.

Gao, H., Wu, X., Li, J., Sun, X. & Yang, J., 2026. The datasets for "massive-scale unlabeled field and labeled synthetic seismic datasets of global shelf-edge scale clinothems".

Kim, D., Ka, W., Ahn, P., Joo, D., Chun, S. & Kim, J., 2022. Global-local path networks for monocular depth estimation with vertical cutdepth, *arXiv preprint arXiv:2201.07436*.

Li, C., Shen, Y., Fomel, S., Waheed, U.B., Savvaidis, A. & Chen, Y., 2026. Geofwi: a large velocity model data set for benchmarking full waveform inversion using deep learning, *J. Geophys. Res.: Machine Learning and Computation,* **3**(2), e2025JH001037.

Liu, F. *et al.*, 2025. Openswi: a massive-scale benchmark dataset for surface wave dispersion curve inversion, *Earth Syst. Sci. Data Discuss.,* **2025,** 1–37.

Mousavi, S.M., Sheng, Y., Zhu, W. & Beroza, G.C., 2019. Stanford earthquake dataset (STEAD): a global data set of seismic signals for ai, *IEEE Access,* **7,** 179 464–179 476.

Patruno, S. & Helland-Hansen, W., 2018. Clinoforms and clinoform systems: Review and dynamic classification scheme for shorelines, subaqueous deltas, shelf edges and continental margins, *Earth Sci. Rev.,* **185,** 202–233.

Ronneberger, O., Fischer, P. & Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-assisted Intervention,* pp. 234–241, Springer-Verlag.

Salles, T., Ding, X. & Brocard, G., 2018. pybadlands: a framework to simulate sediment transport, landscape dynamics and basin stratigraphic evolution through space and time, *PloS one,* **13**(4), e0195557.

Stark, T.J., 2003. Unwrapping instantaneous phase to generate a relative geologic time volume, *SEG International Exposition and Annual Meeting,* SEG–2003 pp., SEG.

Stark, T.J., 2004. Relative geologic time (age) volumes—relating every seismic sample to a geologically reasonable horizon, *Leading Edge,* **23**(9), 928–932.

Wang, G., Wu, X. & Zhang, W., 2024. cigchannel: A massive-scale 3D seismic dataset with labeled paleochannels for advancing deep learning in seismic interpretation, *Earth Syst. Sci. Data,* **2024,** 1–27.

Wang, J. *et al.*, 2020. Deep high-resolution representation learning for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.,* **43**(10), 3349–3364.

Wang, Z., Simoncelli, E.P. & Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment, *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003, Vol. 2,* pp. 1398–1402, IEEE.

Wang, Z., Bovik, A.C., Sheikh, H.R. & Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.,* **13**(4), 600–612.

Wu, X., Geng, Z., Shi, Y., Pham, N., Fomel, S. & Caumon, G., 2020. Building realistic structure models to train convolutional neural networks for seismic structural interpretation, *Geophysics,* **85**(4), WA27–WA39.