https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Global Scenario Reference Datasets for Climate Change Integrated Assessment with Machine Learning

Yi-Ming Wei[1, 2, 3, 4, *], Li-Jing Liu[1, 2, 3, 4, *], Shu-Xin Zhang[1, 2, 3], Jia-Xin Liu[1, 2, 3], Luo Zhao[1, 2, 3], Rong-Gang Cong [1, 2, 3, 4, *], Qiao-Mei Liang[1, 2, 3, 4, *], Te Han[1, 2, 3, 4], Xiao-Chen Yuan [1, 2, 3, 4], Yi-Ming Chen [1, 2, 3], Yu-Xiang Hu[1, 2, 3], Ze-Qi Xu [1, 2, 3], Hong-Bo Chen[1, 2, 3], Yu-Xuan Xiao[1, 2, 3], Peng Wang[1, 2, 3], Song-Yang Yan[1, 2, 3], Xiao-Ling Huang[1, 2, 3], Tian-Yuan Wang[1, 2, 3], Xiao-Qi Li[1, 2, 3], Hao-Ran Xu[1, 2, 3], Wen-Chang Zhao[1, 2, 3], Biying Yu[1, 2, 3, 4], Baojun Tang[1, 2, 3, 4], Lan-Cui Liu[5], Hua Liao[1, 2, 3, 4], Zhi-Qiang Li[6], Rui Cui[6]

[1] Center for Energy and Environmental Policy Research, Beijing Institute of Technology, Beijing, China
[2] School of Management, Beijing Institute of Technology, Beijing, China
[3] Beijing Lab for System Engineering of Carbon Neutrality, Beijing, China
[4] NSFC Basic Science Center for Energy and Climate Change, Beijing, China
[5] Business School, Beijing Normal University, Beijing, China
[6] Information Technology Center, Beijing Institute of Technology, Beijing, China

Correspondence:
Yi-Ming Wei (wei@bit.edu.cn)
Li-Jing Liu (liulijing@bit.edu.cn)
Rong-Gang Cong (leocongbit@bit.edu.cn)
Qiao-Mei Liang (liangqiaomei@bit.edu.cn)

Abstract. The deepening of global climate change research and increasingly complex integrated assessment methods generate large amounts of heterogeneous data. The rapid development of artificial intelligence (AI) models, particularly large language models (LLMs) and deep learning techniques, has enhanced the ability to handle vast data, providing new approaches and perspectives for climate analysis. To address the demand for multi-dimensional and comparable scenario design in climate change prediction and policy simulation, this study employs hybrid machine learning techniques to collect and process scenario data from existing literature, developing the Global Climate Scenario Reference datasets (GCSR). The GCSR incorporates data from approximately 90,000 articles across multiple temporal and spatial scales and extracts approximately 53,185 scenarios. With its large scale, extensive coverage, and detailed classification, the GCSR provides a robust foundation for climate change prediction, risk assessment, mitigation policy, and adaptation strategy planning, supporting scenario design in related fields.

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

1. Introduction

The global climate crisis poses the greatest threat to both humanity and ecosystems (Zabala, 2021). Since the Stockholm Declaration was signed in 1972, global attention to climate change has continued to rise. The international community has successively adopted several important initiatives, such as the 1997 Kyoto Protocol and the 2015 Paris Agreement, which established national mitigation targets and emission reduction mechanisms (Masood, 1997; Tollefson and Weiss, 2015). Assessing climate change and managing climate governance have become crucial challenges that require urgent attention.

In recent years, Machine Learning (ML) and Artificial Intelligence (AI) have emerged as powerful tools driving technological advancements. How to more effectively apply these tools to tackle climate change has garnered widespread attention from both the ML community and climate experts (Taddeo et al., 2021). Considering the exceptional capabilities of AI, particularly Large Language Models (LLMs), in handling large-scale heterogeneous data, extracting latent variables, and managing incomplete and uncertain data (Toetzke et al., 2022; Wu et al., 2024), AI technology is deemed to have significant advantages in assisting data collection to support decision-making and guiding future development plans (Rolnick et al., 2022). By extending to climate change governance goals, AI technology offers researchers new analytical frameworks and tools, enhancing climate prediction, policy simulation, risk assessment, and adaptation strategy planning. At the same time, it supports policymakers in formulating more scientific and effective response measures.

Global climate change and greenhouse gas emissions involve both climate and economic systems, and simulation approaches represented by integrated assessment models have become a powerful tool for global climate governance (Wei et al., 2023). Specifically, human economic activities generate carbon emissions, which enter the carbon cycle and alter greenhouse gas concentration and temperature, thereby affecting the climate system. Changes in the climate system in turn feed back to the economic system by impacting agricultural production (Ren et al., 2023; Bousfield et al., 2024), water resource distribution (Padrón et al., 2020; Utsumi and Kim, 2022), and energy supply and demand (van Ruijven et al., 2019; Perera et al., 2020), etc. In climate change assessment and response research, scenario design and simulation are the most critical means used to describe the uncertainties between the climate and economic systems (O'Neill et al., 2020). The significant rise in climate change research (Kallbekken, 2015) and the escalating complexity of

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

corresponding simulation methods (Cointe et al., 2019) have resulted in the generation of a large number of heterogeneous scenario data.

Climate scenarios not only provide core hypotheses for exploring future development pathways and assessing the impacts of current decisions (Rogelj et al., 2019), but also facilitate the identification of blind spots and the generation of creative solutions (Finch et al., 2024). The adequacy of current scenario settings and their ability to accurately reflect the volatile realities of the future have become one of the widely focused issues in the academic community (Weber et al., 2018; Anderson and Jewell, 2019; O'Neill et al., 2020; Guivarch et al., 2022). Institutions and researchers from diverse backgrounds (such as knowledge systems, disciplines, and experiences) may develop different scenarios (Cornell et al., 2013; Trutnevyte et al., 2016; Carrington and Stephenson, 2018), leading to significant variations in research results (Kriegler et al., 2016; van Vuuren et al., 2020). Therefore, it is essential to develop a more objective, detailed, and comprehensive climate change scenario reference dataset. This dataset focuses on the climate change-related literature that carries out simulation analysis and extracts relevant descriptive information for scenario design. The scenario descriptions of this dataset help to enhance scholars' understanding of the simulation of climate and economic-related variables and provide a reference for designing appropriate scenarios for research objectives, thus it can serve as a complement to existing scenario databases that focus on simulation results under different scenarios (such as the IPCC AR6 database).

Here, this study leverages AI's capabilities in processing large-scale heterogeneous data and efficiently extracting value information from complex and diverse data (Toetzke et al., 2022; Wu et al., 2024), attempts to employ hybrid ML methods to collect and process scenario data from existing literature, and develops the Global Climate Scenario Reference datasets (GCSR). Approximately 90,000 climate change-related studies, spanning different temporal and spatial scales, have been collected. The content covers multiple key dimensions: cause analysis, impact assessment, prediction methods, and governance strategies of climate change. Drawing on these extensive literature resources, this study uses an LLM to extract and organize around 53,185 scenario data points. Employing ML techniques such as keyword recognition and topic extraction, this study categorizes the scenarios more precisely. Researchers can then use issue-specific keywords to swiftly identify related scenarios in the GCSR, thereby guiding their exploration of uncertainties in scenario design research. Policymakers, after locating relevant scenarios, can further filter related

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

96 literature and scenario results to inform policy design. It provides powerful data support for scenario
97 design in research areas such as global climate change prediction, risk assessment, mitigation policy
98 simulation, and adaptation strategy planning.

99 2. Methods

100 This study provides a complete process from literature identification, literature collection,
101 scenario extraction, scenario keyword recognition, and topic classification, to the development of
102 GCSR (see Figure 1). Based on a thematic analysis of the IPCC reports (IPCC, 2022), the major
103 issues related to climate change can be grouped into four key categories: its underlying causes,
104 associated impacts and risks, future emission and temperature projections, and governance and
105 response strategies. Accordingly, the literature can be classified into four main categories: climate
106 change causes, impacts, predictions, and governance. The Web of Science core database is used to
107 check and supplement the search formula of various literature, ultimately finalizing the search
108 formula and collecting the relevant studies. Subsequently, this study utilizes the LLM to extract
109 scenarios and reclassifies each scenario into four categories. To facilitate easy access and use by
110 researchers, natural language processing techniques are applied to the initial scenario dataset. This
111 is followed by high-frequency word analysis, keyword extraction, and topic classification using
112 BERTopic semantic recognition, allowing for more specific subcategories of scenarios. Finally, the
113 scenario data is organized and stored according to these refined categories.

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

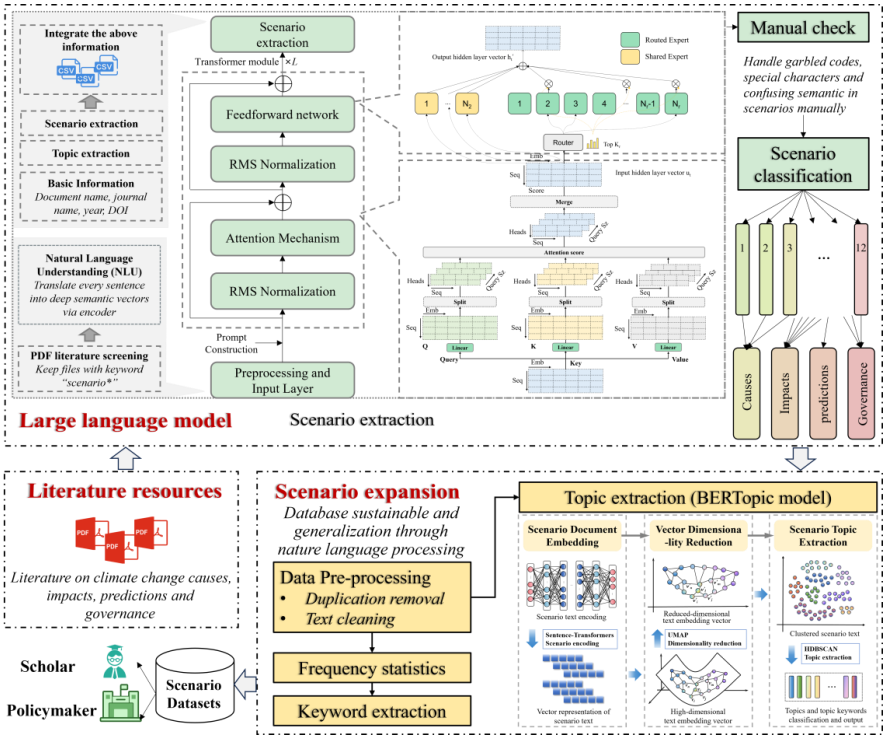

Figure 1: Overview of the methodological framework for developing the GCSR.

2.1 Literature classification and collection

The subcategories of the four issues are shown in Figure 2(a). Climate change causes can be attributed to both natural and human factors. In terms of human factors, economic activities, lifestyles, and energy use generate greenhouse gas emissions; changes in land use patterns affect carbon sequestration. The impacts of climate change are observed across both natural and human systems. In natural systems, climate change alters the structure and function of ecosystems and exacerbates extreme weather events. In the human system, climate change threatens water and food security and human health and has cascade effects on key economic sectors. Climate change predictions mainly involve estimating future trends in key climate elements such as temperature rise and precipitation. Climate change governance mainly covers two core strategies: mitigation and adaptation. Mitigation, beyond the application of emission reduction technologies, relies on a series of emission reduction policies, including mainstream policy measures such as international cooperative emission reduction mechanisms, carbon pricing mechanisms, and power market

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

reforms.

This study first identifies initial search formulas for various types of literature. On August 6, 2024, a search was performed using the Web of Science Core Collection database, which encompasses the Science Citation Index Expanded (SCI-E) and Social Science Citation Index (SSCI). During the search process, keywords from highly relevant literature are reviewed to refine initial search formulas. Subsequently, search formulas are revised, and the search is conducted again. After multiple iterations of verification and modification, the study finalizes search formulas for each type of literature (see Supporting Information 1). In total, approximately 90,000 articles are collected.

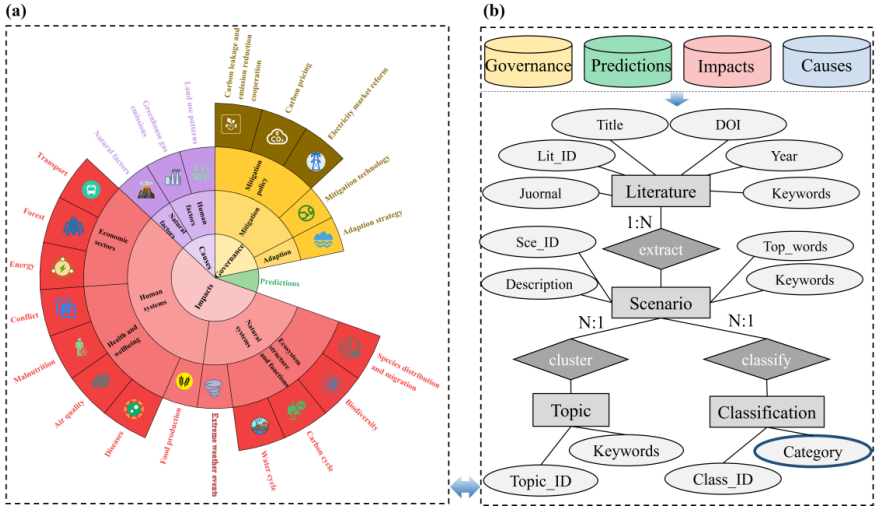

Figure 2: Datasets structure and number of scenarios. (a) Literature classification standards. **(b)** The structure of the datasets.

2.2 Scenario extraction using LLM

This study utilizes the excellent natural language deep semantic understanding ability of the LLM to efficiently identify and extract key information from complex texts (Ray, 2023; Qu and Wang, 2024). After multiple rounds of testing and evaluation, this study selects the DeepSeek-V2 model for scenario extraction from the literature, as it outperforms others in terms of accuracy and operational efficiency. DeepSeek-V2 supports processing context information up to 128,000 tokens, making it excellent at handling long documents and performing complex scenario extraction tasks with high accuracy in capturing key scenario-related information (DeepSeek-AI, 2024a; DeepSeek-

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

AI, 2024b). To precisely identify scenarios, this study optimizes the design of input prompts for LLM, developing an optimal prompt structure capable of guiding LLM to accurately capture scenarios. The specific scenario extraction process includes the following four main steps:

(1) Processing of Input Literature Data

The literature data undergoes text preprocessing and segmentation, transforming it into a sequence of tokens $X=[x_1,x_2,...,x_n]$. n represents the number of tokens in the literature.

(2) Generation of Latent Vectors

The Multi-head Latent Attention (MLA) mechanism is employed to transform the input tokens into compact latent vectors, denoted as z_i, and can be expressed as Eq. (1).

$$MLA(X)=Concat\left(Head_1,Head_2,\ldots,Head_h\right)W^O \tag{1}$$

Where $Head_i$ represents the output of the *i-th* attention head. The final multi-head attention result is obtained by concatenating the outputs of all heads and combining them with the output weight matrix W^O. Each attention head $Head_i$ is computed as Eq. (2):

$$Head_i=softmax\left(\frac{QW_i^Q\left(KW_i^K\right)^T}{\sqrt{d_k}}\right)\left(VW_i^V\right) \tag{2}$$

Where Q is the matrix of Queries, K is the matrix of Keys, V is the matrix of Values, W_i^Q,W_i^K,W_i^V are learnable weight matrices for the *i-th* head that project the input queries, keys, and values into appropriate spaces. The $softmax$ operation normalizes the attention scores. Through this mechanism, the model captures diverse information via multi-head parallel processing, enabling it to integrate the most representative features and generate refined latent vectors z_i (Han et al., 2024). These refined vectors provide an efficient input representation for subsequent scenario recognition and extraction tasks.

(3) Scenario Recognition and Extraction

The core of scenario extraction lies in accurately recognizing and outputting scenarios, which is achieved through the Mixture-of-Experts (MoE). The latent vectors z_i are input into multiple experts, and the activation weights of each expert are computed based on a soft selection mechanism. The experts relevant to the scenario extraction task are selectively activated for computation, and the corresponding scenario representations are generated, as shown in Eq.(3).

$$z_i^*=\sum_{e=1}^{k}\alpha_{i,e}\cdot E_e\left(z_i\right) \tag{3}$$

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Where z_i^* denotes the final latent vector, including scenario information extracted from different experts. $\alpha_{i,e}$ is the selection weight of the *e*-th expert for the latent vector z_i, indicating the expert's contribution to the input. $E_e(z_i)$ is the processing result of the *e*-th expert on the latent vector z_i, typically a transformation or mapping function used to generate scenario-related representations. Through the MoE mechanism, DeepSeek-V2 can selectively activate appropriate experts, ensuring that only the scenario-relevant contents are processed, thereby improving the efficiency and accuracy of scenario extraction.

(4) Output Generation

The latent vectors z_i processed by the expert network are further analyzed and aggregated to generate the final scenario data y. The model outputs the extracted scenario information as structured text, as shown in Eq.(3).

$$y = DeepSeekMoE\left(z_1^*, z_2^*, \ldots, z_n^*\right) \tag{4}$$

Where $\left(z_1^*, z_2^*, \ldots, z_n^*\right)$ denotes the processed latent vector, and y represents the extracted scenario data. The design of this model perfectly aligns with the requirements of scenario extraction tasks, efficiently handling multimodal data and diverse corpora. These output results consist of paragraphs, sentences, or key information about the literature's scenarios, ultimately summarizing the scenario descriptions and hypotheses.

In total, 136,754 scenarios are extracted from the literature. To improve the accuracy of the categorization, this study reclassifies scenarios using the LLM to avoid misalignment with their assigned categories or the omission of scenarios that exhibit characteristics of multiple categories. Additionally, a small sample of scenarios from each category is manually verified to assess the accuracy of the model's classification results. This verified subset is then used as a training set to reclassify the remaining scenarios. Ultimately, based on the four main categories, this study further subdivides the scenarios into 12 specific subcategories, as shown in Figure 2(a).

2.3 Semantic check

The initial scenarios contain issues such as unrecognizable special characters, misrecognized letters, and semantic confusion. To resolve these problems, this study systematically screens and corrects the data, addressing character recognition errors and ensuring the integrity and readability of the scenario datasets. To tackle the issues of semantic confusion, this study utilizes the Python-based LanguageTool (Jumanov and Karshiev, 2020), which combines predefined grammatical rules,

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

language patterns, and statistical models to achieve efficient grammatical checking.

Following the automated error detection and suggestions, a manual review is conducted to ensure detailed modifications, ensuring that each scenario description is semantically accurate and logically coherent. Moreover, the extracted data source may not have real climate scenarios. The scenario mentioned in the original paper may have other meanings. So, we use a supervised learning method to judge the extracted data, and finally, only retain the data with real climate scenarios.

2.4 Scenario expansion using BERTopic

Based on the initially obtained scenario datasets, this study applies a series of methods, including scenario coding, word frequency statistics, similarity calculation, and topic extraction, to conduct in-depth processing on each scenario dataset separately (Berio Fortini et al., 2023). These steps remove duplicate scenarios and assign clear thematic tags to each group of scenarios, thereby enhancing the structural clarity and operability of the scenario datasets. The optimization of the scenario datasets can provide a clearer and more efficient retrieval environment for subsequent scholars' research, which helps to promote in-depth research in related fields.

(1) Duplication removal

To enhance the retrieval efficiency of the scenario datasets, this study removes duplicate scenarios, effectively reducing data redundancy. Through this processing, scenario datasets become more concise and clearer, significantly improving the efficiency of users during retrieval and usage, allowing them to quickly and accurately locate the required key information.

(2) Text cleaning

This study tokenizes the text data, divides each document into independent lexical units, and converts all words to lowercase, eliminating the impact of case differences on the statistical results. Furthermore, to more accurately focus on the core content of the scenarios, this study removes punctuation, special characters, and stop words that appear frequently but contribute little to actual semantics, such as "the", "and", etc.

(3) High-frequency word statistics

To provide basic data support for subsequent analysis of the scenario datasets, this study uses Natural Language Toolkit (NLTK) (Bird and Loper, 2004) to count high-frequency words in the text

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

data. Specifically, the program counts the word frequency in each scenario and selects the most representative high-frequency words. Finally, this study outputs the top five high-frequency words in each scenario.

(4) Keyword extraction

To identify the core keywords in the scenario description, this study uses the Best Matching 25 (BM25) algorithm. Compared to traditional methods, BM25 can more accurately identify keywords that are representative of scenario texts, and it avoids the common issue of term frequency bias in long documents, thereby improving the accuracy and effectiveness of keyword extraction (Amati and Rijsbergen, 2002). The weight calculation for each term in the scenario y is shown in Eq.(5).

$$BM\ 25(t,y)=\sum_{i=1}^{n} IDF\left(t_i\right)\cdot\frac{\mathrm{TF}\left(t_i,y\right)\cdot\left(k_1+1\right)}{TF\left(t_i,y\right)+k_1\cdot\left(1-b+b\cdot\frac{|y|}{\mathrm{avgdl}}\right)} \tag{5}$$

$$IDF(t)=log\left(\frac{N-df(t)+0.5}{df(t)+0.5}+1\right) \tag{6}$$

Inverse Document Frequency (IDF) is used to measure the scarcity of a term t_i across the entire scenario dataset. N is the total number of scenarios in the datasets, and $df(t)$ is the number of scenarios containing the term. $f\left(t_i,y\right)$ represents the term frequency (TF) of t_i in the scenario y. k_1 and b are adjustment parameters to balance the influence of term frequency and document length on the score. $|y|$ is the length of the scenario y, and $avgdl$ is the average length of all scenario texts. The BM25 algorithm not only optimizes the recognition of common words but also comprehensively considers TF and IDF, making the extracted keywords more aligned with the core content of the scenario.

(5) Topic extraction and classification

To effectively extract and classify the groups of scenarios, this study adopts BERTopic, a topic modeling method based on Transformer. Traditional topic classification methods typically focus on the surface information of words, neglecting the contextual semantics and deep relationships between words. When dealing with semantically complex or syntactically flexible texts, this limitation becomes particularly prominent, often resulting in information loss or misjudgment. However, BERTopic is a topic modeling method based on Transformer, which combines deep learning and clustering techniques. It is capable of handling large-scale text, automatically and efficiently identifying topics within the text (Grootendorst, 2022). Furthermore, BERTopic supports

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

dynamic updates and extensions, allowing it to adapt to the continuous evolution of scenario datasets.

$$\boldsymbol{v} = \mathrm{ST}(y), v' = UMAP(v), Cluster_i = \left\{ \boldsymbol{v}' \mid Density\left(\boldsymbol{v}'\right) \geq Threshold \right\} \quad (7)$$

The BERTopic model consists of three main steps. Firstly, it uses a pre-trained model, sentence-transformers (all-MiniLM-L6-v2), to encode the scenario text into a high-dimensional vector v, allowing the semantic relationships between texts to be effectively reflected. $\mathrm{ST}(y)$ represents the process of encoding scenario text using sentence-transformers. Furthermore, the Uniform Manifold Approximation and Projection (UMAP) algorithm is used to reduce the dimension of the semantic vector, while retaining the location information between the documents (McInnes et al., 2018). Finally, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is employed to cluster the dimensionality-reduced scenario text vectors. $Density\left(\boldsymbol{v}'\right)$ denotes the final scenario classifications obtained through HDBSCAN clustering. $Threshold$ serves as the minimum density threshold to filter out text clusters with higher density.

Through this process, this study can automatically and efficiently identify topics in scenarios and output keywords for each topic category through Class-based TF-IDF, providing effective support for the efficient categorization and dynamic evolution of scenarios. This ensures the long-term adaptability and sustainability of the model.

3. Results and discussion

Figure 3 illustrates the number of scenarios for four categories across different years, as well as the proportion of scenarios in each sub-category. Overall, after 2010, there was a significant increase in the number of scenarios related to the climate change causes and impacts. In contrast, the number of scenarios concerning prediction and governance did not show a slow growth trend until 2015. Regarding the climate change causes, approximately 70% - 80% of the scenarios focused on greenhouse gas emissions, around 10% - 20% concentrated on land use patterns, and the remaining scenarios (less than 10%) paid attention to natural factors. In terms of the climate change impacts, the number of scenarios centered on ecological structure and function impacts ranked first. However, over time, the quantity of such scenarios exhibited a decreasing trend. Meanwhile, the focus on the impacts on economic sectors gradually increased, with their proportion approaching 25% in recent years. With respect to climate change governance, mitigation policies received the most attention, with the proportion of scenarios in this category remaining at around 60%. It was followed by scenarios related to mitigation technology (about 30%), and the remaining 10% focused

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

on adaptation measures.

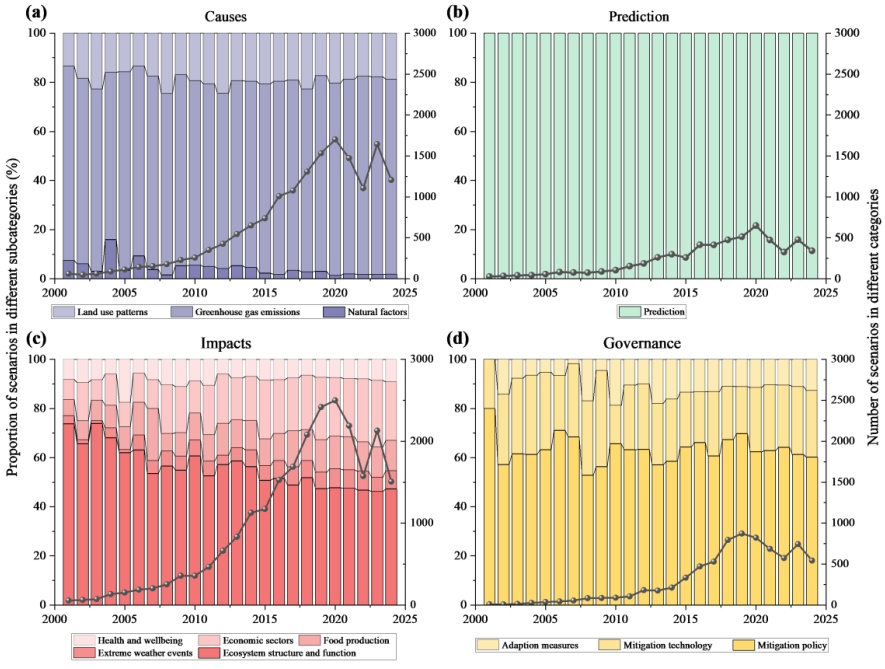

Figure 3: Proportion and number of different scenario categories. (a) scenarios in climate change causes. **(b)** scenarios in climate change prediction. **(c)** scenarios in climate change impacts. **(d)** scenarios in climate change governance. The bar chart represents the proportion of scenarios in each subcategory, and the dot-and-line graph represents the total number of scenarios in each category.

Figure 4 displays the word frequency statistics results of top words under different scenario categories. The statistics indicate that in scenarios focusing on the climate change causes, words such as "carbon", "emission", "CO_2" appear with relatively high frequency. This reflects that relevant studies primarily concentrate on the core factor that emission activities trigger climate change. In scenarios centered around climate change prediction, words like "future", "temperature", and "model" occur frequently. This demonstrates a research trend of utilizing various models to predict indicators such as future temperature rise. In addition, studies focusing on the climate change impacts tend to emphasize areas like "energy", "temperature", and "water". As for research on climate change governance, scenarios are often designed around elements such as "tax", "electricity", and "price".

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

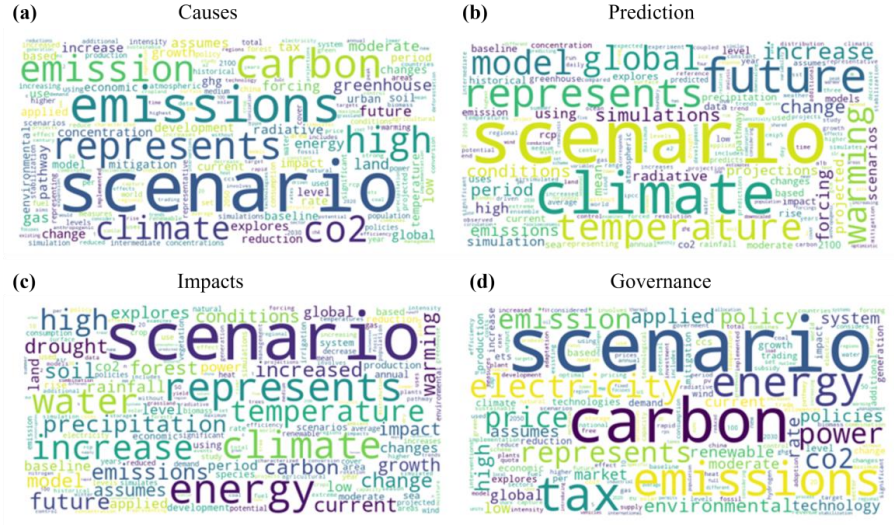

Figure 4: Word cloud of top words for different scenario categories. (a) scenarios in climate change causes. **(b)** scenarios in climate change prediction. **(c)** scenarios in climate change impacts. **(d)** scenarios in climate change governance.

GCSR stores a wealth of scenario design cases. Its detailed classification and labeling enable researchers and policymakers to efficiently identify and utilize scenario information relevant to specific issues. For instance, if researchers are studying the impact of climate change on air pollution, they can use keywords like “climate change impact,” “human systems,” and “health and wellbeing” to find scenarios related to “air”, thereby easily obtaining closely related scenario cases. These cases contain detailed scenario descriptions, serving as invaluable reference materials for researchers. Policymakers who have screened the GCSR for scenarios relevant to their policy issues can further utilize the literature information provided by the GCSR to directly access the original literature. By comparing simulation results under different scenarios, policymakers can assess the potential impacts and feasibility of various policy options. Hence, the significance of the GCSR dataset lies in the following: ①Enhancing Research Efficiency: The GCSR significantly reduces the time researchers and policymakers spend searching for relevant cases and literature through efficient keyword searching and scenario matching functions, thereby improving their work efficiency and research quality. ②Promoting Knowledge Sharing: The scenario cases in the GCSR originate from global researchers and practitioners, fostering knowledge sharing and exchange within the field of scenario design research and contributing to a more comprehensive and in-depth understanding. ③

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Enhancing Scientific Decision-Making: Based on scenario cases in the GCSR, policymakers can conduct more scientific and objective decision-making analysis, reducing the blindness and uncertainty in decision-making and enhancing the relevance and effectiveness of policies. ④ Driving Disciplinary Development: The establishment and use of the GCSR also promote the continuous improvement and development of scenario design methods, providing new perspectives and ideas for research in this field.

4. Data availability

To facilitate the use by researchers, decision-makers, industry professionals, and other relevant personnel, the GCSR provides two forms of data storage: MySQL and EXCEL. The final results are uploaded through Zenodo: https://doi.org/10.5281/zenodo.15536298 (Wei et al., 2025). The structure of the GCSR core data is shown in Figure 2 (b).

(1) Literature: Stores information related to the source literature of scenarios, including the title of the literature, the name of the journal published, the year of publication, the DOI number, and keywords. This information provides comprehensive background support for the source of scenarios.

(2) Scenario: Contains extracted and summarized scenario descriptions and assumptions. Each document may contain multiple scenarios, with each scenario stored as a separate piece of content. In addition, it also includes the high-frequency words of the scenario, keywords extracted by the Best Matching 25 (BM25), and the scenario topics based on BERTopic clustering.

(3) Topic: Includes scenario topic categories generated by BERTopic clustering, the name of each topic category, and its core keywords, facilitating further organization and classification of scenario content.

(4) Classification: Provides scenario classification based on LLM categorization, and hierarchy for labeling scenarios (including the first, second, and third-level categories shown in Figure 2(a)), aiding in structured analysis from broad categories to specific scenarios.

5. Technical Validation

As machine learning increasingly relies on large datasets, there is a growing acknowledgment

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

of the importance of data quality (Gero et al., 2023). The data quality of the GCSR is affected by several factors. The primary factor lies in whether the data sources are comprehensive and extensive, capable of covering a majority of climate change scenarios, which directly impacts the training capability of the model. Secondly, the effectiveness of scenario extraction is also crucial. It requires the effective extraction of large-scale and heterogeneous scenarios while preserving the original meaning of the scenarios, which helps enhance the model's understanding capability. Furthermore, the accuracy of scenario classification is a key step in ensuring the overall reliability of the scenario dataset, directly influencing the efficient organization and retrieval of information within the dataset. Therefore, to improve data quality, this study adopts the following specific measures.

5.1 Ensure more comprehensive literature sources

In this study, the search query about climate change research was repeatedly modified and validated according to the identified relevant literature topics and keywords, and the final search query ensured comprehensive coverage of the literature. Additionally, literature not automatically marked as successfully downloaded by the program was manually retrieved as a supplement.

5.2 Check the extracted scenarios' semantics

On the one hand, the scenarios extracted by the original LLM may contain issues such as incorrect letter recognition and semantic confusion. This study utilizes a language tool to help identify potential grammatical errors and unclear expressions in the initial scenarios. On the other hand, the extracted data might not represent real climate scenarios. The term "scenario" mentioned in the original papers may carry other meanings. To address this, we employ a supervised learning approach to evaluate the extracted data. By carefully reading the original texts, we classify 1,862 data samples, assigning a label of 1 to those that truly contain climate scenarios and 0 to those that do not. Distinct from manual assessment, there might contain some generalization errors of such a big data analysis. However, after conducting five-fold stratified cross-validation based on these samples, the accuracy rate of the model reaches over 90% (see Table 1).

Table 1 Comparison between supervised learning and manual verification of scenario results

Fold	Precision	Recall	F1
1	0.961111111	0.940217391	0.950549451
2	0.941489362	0.961956522	0.951612903
3	0.972826087	0.978142077	0.975476839

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

4	0.928205128	0.989071038	0.957671958
5	0.977142857	0.929347826	0.95264624

5.3 Verify the scenario classification's training set

To improve the classification accuracy, a small subset of scenarios is first classified using the LLM, and these preliminary classifications are then further verified by assessing the rationality of the features, content, and assigned categories. Inconsistencies are corrected to develop a high-quality training set, which is subsequently used to classify additional scenarios. Additionally, to account for scenarios outside the four predefined categories, the overall recall rate is calculated by comparing the number of categorized scenarios to the total (see Table 2). The results show that the quality of the GCSR is reliable.

Table 2 Comparison between LLM and manual verification of classification results.

Classification		LLM classification vs. manual check		
		Precision	Recall	F1
Climate change causes	Climate change caused by natural factors	0.99	0.94	0.96
	Greenhouse gas emissions	0.92		0.93
	Land use patterns	0.94		0.94
Climate change impacts	Ecosystem structure and function	0.80		0.86
	Extreme weather events	0.98		0.96
	Food production	0.97		0.96
	Health and wellbeing	0.92		0.93
	Economic sectors	0.92		0.93
Climate change predictions		0.90		0.92
Climate policies and governance	Mitigation policy	0.92		0.93
	Mitigation technology	0.93		0.93
	Adaption measures	0.93		0.93

6. Code availability

Code, dataset, and some intermediate results are presented in Supporting Information 2.

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

Reference

Amati, G. and Rijsbergen, C. J. V.: Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Trans. Inf. Syst., 20, 357–389, https://doi.org/10.1145/582415.582416, 2002.

Anderson, K. and Jewell, J.: Debating the bedrock of climate-change mitigation scenarios, Nature, 573, 348–349, https://doi.org/10.1038/d41586-019-02744-9, 2019.

Berio Fortini, L., Kaiser, L. R., Frazier, A. G., and Giambelluca, T. W.: Examining current bias and future projection consistency of globally downscaled climate projections commonly used in climate impact studies, Climatic Change, 176, 169, https://doi.org/10.1007/s10584-023-03623-z, 2023.

Bird, S. and Loper, E.: NLTK: The Natural Language Toolkit, In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 214–217, Barcelona, Spain. Association for Computational Linguistics, 2004.

Bousfield, C. G., Morton, O., and Edwards, D. P.: Climate change will exacerbate land conflict between agriculture and timber production, Nature Climate Change, https://doi.org/10.1038/s41558-024-02113-z, 2024.

Carrington, G. and Stephenson, J.: The politics of energy scenarios: Are International Energy Agency and other conservative projections hampering the renewable energy transition?, Energy Research & Social Science, 46, 103-113, https://doi.org/10.1016/j.erss.2018.07.011, 2018.

Cointe, B., Cassen, C., and Nadaï, A.: Organising Policy-Relevant Knowledge for Climate Action: Integrated Assessment Modelling, the IPCC, and the Emergence of a Collective Expertise on Socioeconomic Emission Scenarios, Science & Technology Studies, 32, 36-57, https://doi.org/10.23987/sts.65031, 2019.

Cornell, S., Berkhout, F., Tuinstra, W., Tàbara, J. D., Jäger, J., Chabay, I., de Wit, B., Langlais, R., Mills, D., Moll, P., Otto, I. M., Petersen, A., Pohl, C., and van Kerkhoff, L.: Opening up knowledge systems for better responses to global environmental change, Environmental Science & Policy, 28, 60-70, https://doi.org/10.1016/j.envsci.2012.11.008, 2013.

DeepSeek-AI: DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model, https://doi.org/10.48550/arXiv.2405.04434, 2024a.

DeepSeek-AI: DeepSeek LLM Scaling Open-Source Language Models with Longtermism, https://doi.org/10.48550/arXiv.2401.02954, 2024b.

Finch, M., Older, M., Mahon, M., and Robertson, D.: Climate action and the vantage point of imagined futures: a scenario-based conversation, npj Climate Action, 3, 45, https://doi.org/10.1038/s44168-024-00123-3, 2024.

Gero, K. I., Das, P., Dognin, P., Padhi, I., Sattigeri, P., and Varshney, K. R.: The incentive gap in data work in the era of large models, Nature Machine Intelligence, 5, 565-567, https://doi.org/10.1038/s42256-023-00673-x, 2023.

Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure, https://doi.org/10.48550/arXiv.2203.05794, 2022.

Guivarch, C., Le Gallic, T., Bauer, N., Fragkos, P., Huppmann, D., Jaxa-Rozen, M., Keppo, I., Kriegler, E., Krisztin, T., Marangoni, G., Pye, S., Riahi, K., Schaeffer, R., Tavoni, M., Trutnevyte, E., van Vuuren, D., and Wagner, F.: Using large ensembles of climate change mitigation scenarios for robust insights, Nature Climate Change, 12, 428-435, https://doi.org/10.1038/s41558-022-01349-x, 2022.

Han, T., Cong, R.-G., Yu, B., Tang, B., and Wei, Y.-M.: Integrating local knowledge with ChatGPT-like large-scale language models for enhanced societal comprehension of carbon neutrality, Energy and AI, 18, 100440, https://doi.org/10.1016/j.egyai.2024.100440, 2024.

IPCC: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)], Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 3056, https://doi.org/10.1017/9781009325844, 2022.

Jumanov, I. I. and Karshiev, K. B.: Mechanisms for optimization of detection and correction of text errors based on combining multilevel morphological analysis with n-gram models, Journal of Physics: Conference Series, 1546, 012082, https://doi.org/10.1088/1742-6596/1546/1/012082, 2020.

Kallbekken, S.: National climate ambition must match international targets, Nature,

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

https://doi.org/10.1038/nature.2015.19077, 2015.
Kriegler, E., Mouratiadou, I., Luderer, G., Bauer, N., Brecha, R. J., Calvin, K., De Cian, E., Edmonds, J., Jiang, K., Tavoni, M., and Edenhofer, O.: Will economic growth and fossil fuel scarcity help or hinder climate stabilization?, Climatic Change, 136, 7-22, https://doi.org/10.1007/s10584-016-1668-3, 2016.
Masood, E.: Kyoto agreement creates new agenda for climate research, Nature, 390, 649-650, https://doi.org/10.1038/37686, 1997.
McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, https://doi.org/10.48550/arXiv.1802.03426, 2018.
O'Neill, B. C., Carter, T. R., Ebi, K., Harrison, P. A., Kemp-Benedict, E., Kok, K., Kriegler, E., Preston, B. L., Riahi, K., Sillmann, J., van Ruijven, B. J., van Vuuren, D., Carlisle, D., Conde, C., Fuglestvedt, J., Green, C., Hasegawa, T., Leininger, J., Monteith, S., and Pichs-Madruga, R.: Achievements and needs for the climate change scenario framework, Nature Climate Change, 10, 1074-1084, https://doi.org/10.1038/s41558-020-00952-0, 2020.
Padrón, R. S., Gudmundsson, L., Decharme, B., Ducharne, A., Lawrence, D. M., Mao, J., Peano, D., Krinner, G., Kim, H., and Seneviratne, S. I.: Observed changes in dry-season water availability attributed to human-induced climate change, Nature Geoscience, 13, 477-481, https://doi.org/10.1038/s41561-020-0594-1, 2020.
Perera, A. T. D., Nik, V. M., Chen, D., Scartezzini, J.-L., and Hong, T.: Quantifying the impacts of climate change and extreme climate events on energy systems, Nature Energy, 5, 150-159, https://doi.org/10.1038/s41560-020-0558-0, 2020.
Qu, Y. and Wang, J.: Performance and biases of Large Language Models in public opinion simulation, Humanities and Social Sciences Communications, 11, 1095, https://doi.org/10.1057/s41599-024-03609-x, 2024.
Ray, P. P.: ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems, 3, 121-154, https://doi.org/10.1016/j.iotcps.2023.04.003, 2023.
Ren, C., Zhang, X., Reis, S., Wang, S., Jin, J., Xu, J., and Gu, B.: Climate change unequally affects nitrogen use and losses in global croplands, Nature Food, 4, 294-304, https://doi.org/10.1038/s43016-023-00730-z, 2023.
Rogelj, J., Huppmann, D., Krey, V., Riahi, K., Clarke, L., Gidden, M., Nicholls, Z., and Meinshausen, M.: A new scenario logic for the Paris Agreement long-term temperature goal, Nature, 573, 357-363, https://doi.org/10.1038/s41586-019-1541-4, 2019.
Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y.: Tackling Climate Change with Machine Learning, ACM Comput. Surv., 55, 42, https://doi.org/10.1145/3485128, 2022.
Taddeo, M., Tsamados, A., Cowls, J., and Floridi, L.: Artificial intelligence and the climate emergency: Opportunities, challenges, and recommendations, One Earth, 4, 776-779, https://doi.org/10.1016/j.oneear.2021.05.018, 2021.
Toetzke, M., Banholzer, N., and Feuerriegel, S.: Monitoring global development aid with machine learning, Nature Sustainability, 5, 533-541, https://doi.org/10.1038/s41893-022-00874-z, 2022.
Tollefson, J. and Weiss, K. R.: Nations approve historic global climate accord, Nature, 528, 315-316, https://doi.org/10.1038/528315a, 2015.
Trutnevyte, E., McDowall, W., Tomei, J., and Keppo, I.: Energy scenario choices: Insights from a retrospective review of UK energy futures, Renewable and Sustainable Energy Reviews, 55, 326-337, https://doi.org/10.1016/j.rser.2015.10.067, 2016.
Utsumi, N. and Kim, H.: Observed influence of anthropogenic climate change on tropical cyclone heavy rainfall, Nature Climate Change, 12, 436-440, https://doi.org/10.1038/s41558-022-01344-2, 2022.
van Ruijven, B. J., De Cian, E., and Sue Wing, I.: Amplification of future energy demand growth due to climate change, Nature Communications, 10, 2762, 10.1038/s41467-019-10399-3, 2019.
van Vuuren, D. P., van der Wijst, K.-I., Marsman, S., van den Berg, M., Hof, A. F., and Jones, C. D.: The costs of achieving climate targets and the sources of uncertainty, Nature Climate Change, 10, 329-334, https://doi.org/10.1038/s41558-020-0732-1, 2020.
Weber, C., McCollum, D. L., Edmonds, J., Faria, P., Pyanet, A., Rogelj, J., Tavoni, M., Thoma, J., and Kriegler, E.: Mitigation scenarios must cater to new users, Nature Climate Change, 8, 845-848,

https://doi.org/10.5194/essd-2025-299
Preprint. Discussion started: 12 June 2025
© Author(s) 2025. CC BY 4.0 License.

https://doi.org/10.1038/s41558-018-0293-8, 2018.
Wei, Y.-M., Liu, L.-J., Zhang, S.-X., and Liu, J.-X.: Global Climate Scenario Reference Dataset, [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15536298, 2025.
Wei, Y. M., Liang, Q. M., Yu, B. Y., and Liao, H. 2023. Climate Change Integrated Assessment Model and Its Applications, Science Press.
Wu, L., Huang, Z., Zhang, X., and Wang, Y.: Harmonizing existing climate change mitigation policy datasets with a hybrid machine learning approach, Scientific Data, 11, 580, https://doi.org/10.1038/s41597-024-03411-z, 2024.
Zabala, A.: The right to a sound environment, Nature Sustainability, 4, 750-751, https://doi.org/10.1038/s41893-021-00733-3, 2021.

Acknowledgements

We gratefully acknowledge the financial support of the National Natural Science Foundation of China [grant number 72293600, 72293605, 72488101, 72474024 and 72474021]. We thank our colleagues for their support and acknowledge help from CEEP-BIT and Information Technology Center of BIT.

Author contributions

Y.-M. W., L.-J. L., Q.-M. L. conceptualized the paper. S.-X. Z., J.-X. L., L. Z., Y.-M. C., Y.-X. H., Z.-Q. X., H.-B. C., Y.-X. X., P. W., S.-Y. Y., X.-L. H., T.-Y. W., X.-Q. L., H.-R. X., W.-C. Z., Z.-Q. L., and R. C. acquired the data. J.-X. L. and L. Z. carried out the model. L.-J. L., S.-X. Z., J.-X. L., Q.-M. L., and R.-G. C. contributed to the interpretation of the results. S.-X. Z., J.-X. L., Y.-M. C., Y.-X. H. and P. W. implemented the visualization. J.-X. L. and S.-X. Z. contributed to the Supplementary Information. Y.-M. W., L.-J. L., S.-X. Z., J.-X. L., Q.-M. L., R.-G. C., T. H., X.-C. Y., B. Y., B. T., L.-C. L. and H. L. wrote and reviewed the main manuscript.

Competing interests

The authors declare no competing interests.