COMPARISON-AWARE ACTIVE LEARNING FOR SIMULATION-EFFICIENT GLOBAL CHANGE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Global change analysis investigates how complex systems respond under alternative socioeconomic, technological, and policy scenarios. It relies on computationally expensive global change models and draws conclusions from both individual scenario outcomes and differences between scenarios. To improve the efficiency of analyzing large numbers of scenarios, surrogate models provide a more efficient alternative to direct model simulation. However, their construction still depends on costly original simulator evaluations for training data, and low average prediction error from these surrogate models does not guarantee sufficient accuracy for individual scenarios or present their differences in scientific analysis. To address these limitations, we formulate simulation-efficient global change analysis as reducing original-model evaluations while maintaining accuracy for both scenario outcomes and scientifically relevant comparisons. Particularly, we propose a comparison-aware active learning method, which first constructs an input-derived comparison graph, and then estimates both outcome- and comparison-level uncertainty with an ensemble of surrogate models. Based on these estimates, a calibrated fallback strategy finally identifies scenarios that may require original-model evaluation. We further propose one acquisition strategy, which can select scenarios according to their expected reduction in future original-model evaluations. Experiments on a public energy dataset and a newly constructed dietary dataset show that our method improves performance over existing approaches, reducing the number of evaluations of the original model by 85.71% and 67.29%, respectively, while maintaining high edge sign accuracies of 96.65% and 90.22%. Overall, our method provides an efficient and reliable solution for surrogate modeling in global change analysis.

1 INTRODUCTION

Global change analysis investigates how environmental, socioeconomic, technological, and policy changes propagate through coupled human–earth systems. It supports decision-making in areas such as energy, land use, food, and climate policy, where changes in one sector can affect multiple regions and time periods Calvin et al. (2019); Pauliuk et al. (2017); Pastor et al. (2019). Integrated global change models provide a way to represent these interactions and evaluate their consequences under alternative assumptions Van Vuuren et al. (2019); Calvin et al. (2019).

Such analyses are commonly organized around a predefined set of socioeconomic, technological, and policy scenarios. Scenario frameworks such as the Representative Concentration Pathways and Shared Socioeconomic Pathways support systematic exploration of alternative futures Van Vuuren et al. (2011); Riahi et al. (2017); Rogelj et al. (2018). Large scenario ensembles are increasingly used to characterize uncertainty and evaluate the robustness of scientific conclusions O'Neill et al. (2020); Guivarch et al. (2022); Wessel et al. (2024). However, evaluating each scenario with a global change model can be computationally expensive, especially when the model produces high-dimensional outputs over regions, time periods, and variables Fujimori et al. (2025); Calvin et al. (2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Surrogate models provide an efficient alternative to repeated evaluations of the original model. Recent studies have shown that neural networks can emulate global change models with high predictive accuracy, especially for energy-system applications Holmes et al. (2026); Li et al. (2025); Van Katwyk et al. (2026). Nevertheless, constructing a reliable surrogate still requires a sufficiently large collection of original-model evaluations, and uncertainty estimation remains important when the surrogate is applied to scenarios that were not evaluated directly Gal et al. (2022); Price et al. (2025). Moreover, low average prediction error from these models does not necessarily guarantee reliable predictions for individual scenarios or for differences between scenarios, even though such differences are often central to scientific interpretation Pastor et al. (2020); Dekker et al. (2023b;a).

To address these limitations, we cast simulation-efficient global change analysis as selecting a few scenarios for original-model evaluation, while preserving accuracy on individual outcomes and scientifically relevant cross-scenario comparisons, as illustrated in Fig. 1. Specifically, we develop a comparison-aware active learning framework that constructs an input-derived comparison graph and jointly accounts for uncertainty in scenario outcomes and graph-connected comparisons. A calibrated fallback rule combines these uncertainties with input coverage to identify scenarios that may still require original-model evaluation. The acquisition strategy uses a look-ahead update to estimate how each query changes the fallback set, prioritizes candidates with positive expected net savings, and performs uncertainty- and diversity-aware exploration when direct savings are unavailable.

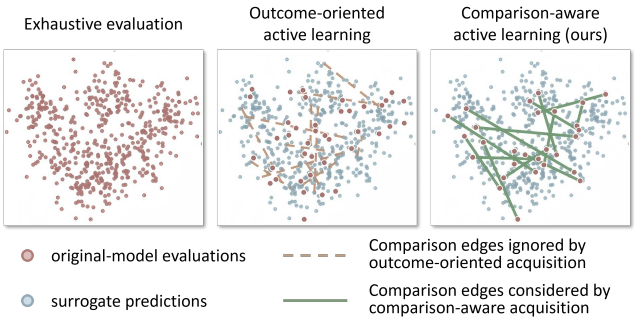


Figure 1: **Motivation for comparison-aware active learning.** Outcome-only acquisition uses individual scenario outcomes, whereas comparison-aware acquisition additionally incorporates cross-scenario relationships when selecting scenarios for original-model evaluation.

We evaluate the proposed method on two global change analysis datasets. The first is a publicly available energy-system dataset derived from GCAM emulation experiments Holmes et al. (2026). The second is a dietary-system dataset constructed in this study, whose scenario inputs include region-specific perturbations and therefore represent regional heterogeneity Rodés-Bachs et al. (2026). Across the two tasks, the method improves performance over existing approaches, reducing the number of evaluations of the original model by 85.71% and 67.29%, while achieving high edge sign accuracies of 96.65% and 90.22%, respectively. Overall, it provides an efficient and reliable solution for surrogate modeling in global change analysis.

The main contributions of our work are as follows:

- We formulate global change analysis as a selective original-model evaluation problem with accuracy required on both scenario outcomes and cross-scenario comparisons.
- We develop a comparison-aware active learning method for scenario selection. This method integrates an input-derived comparison graph, a calibrated fallback strategy, and an acquisition mechanism to maximize the expected reduction in subsequent original-model evaluations, while accounting for uncertainty in both individual scenario outcomes and cross-scenario differences.
- We demonstrate the superiority of the proposed method on both global change analysis tasks in energy systems and a newly constructed regionally heterogeneous dietary dataset.

2 Related Work

Integrated Assessment Models (IAMs) IAMs like GCAM and IMAGE are widely used to study interactions among energy, economy, land, water, agriculture, and climate systems Calvin et al. (2019); Rogelj et al. (2018); Van Vuuren et al. (2019). They provide a core modeling foundation

for global environmental assessments and are commonly used to explore alternative climatic and socioeconomic futures represented by frameworks such as the Representative Concentration Pathways (RCPs) and Shared Socioeconomic Pathways (SSPs) Van Vuuren et al. (2011); Riahi et al. (2017); Huppmann et al. (2018). Rather than relying on a single simulation, IAM studies typically derive scientific conclusions by comparing baseline and intervention scenarios, varying uncertain assumptions, and examining robustness across models and plausible futures Marangoni et al. (2017); Dekker et al. (2023b); Lempert & Collins (2007). Therefore, the scientific value of IAM simulations lies not only in the outputs of individual scenarios, but also in the differences and relationships among them.

Machine Learning for Large Scientific Simulation As uncertainty spaces expand, global-change analysis is increasingly shifting from a few representative pathways toward large scenario ensembles spanning socioeconomic, technological, and policy uncertainties Guivarch et al. (2022); Beath et al. (2026); Wessel et al. (2024). Because high-fidelity simulation of such ensembles is computationally expensive, machine learning has been used to approximate costly simulations and reduce the required number of simulator evaluations Wang et al. (2023). Physics-informed learning, neural operators, and data-driven systems such as Pangu-Weather, GraphCast, and NeuralGCM demonstrate that complex simulation mappings can be learned efficiently Karniadakis et al. (2021); Lu et al. (2021); Li et al. (2021); Bi et al. (2023); Lam et al. (2023); Kochkov et al. (2024). Recent studies extend this idea to integrated-assessment and mitigation scenarios through generative scenario expansion and neural surrogates for IAMs such as GCAM Li et al. (2025); Holmes et al. (2026), while active learning selects informative simulator runs to improve surrogate construction Musekamp et al. (2025); Kim et al. (2025). However, these methods primarily optimize point-wise predictive accuracy, uncertainty, or information gain. Such objectives do not necessarily ensure that a surrogate preserves the simulator's response to changes in policy, technology, or socioeconomic assumptions Holmes et al. (2026), even though these cross-scenario responses are often central to scientific analysis. We therefore study finite, pre-specified scenario campaigns that minimize high-fidelity simulations while reliably completing every scenario and preserving relevant cross-scenario relationships.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Global change analysis typically requires researchers to predefine a set of socioeconomic, technological, and policy scenarios. For each scenario, a global change analysis model is run to produce high-dimensional simulation outputs. These outputs are then used both to analyze individual scenarios and to compare alternative scenarios. Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ denote this scenario set, where $\mathbf{x}_i \in \mathbb{R}^d$ is the parameter vector of scenario i. We denote the original global change model by f, with $\mathbf{y}_i = f(\mathbf{x}_i)$ representing its output for scenario i. Although the output may naturally be represented as a tensor over regions, time periods, and variables, we write it as a vector $\mathbf{y}_i \in \mathbb{R}^p$ for notational simplicity. Let $\mathcal{C}$ denote the set of scenario pairs of scientific interest. For each $(i, j) \in \mathcal{C}$, the corresponding comparison outcome is $\mathbf{\Delta}_{ij} = \mathbf{y}_j - \mathbf{y}_i$.

Evaluating every scenario with the original model can be computationally expensive. Active learning instead evaluates a selected subset of scenarios with the original model and uses the resulting outputs to train a surrogate model. Let $\mathcal{Q}_t$ denote the scenarios evaluated by the original simulator at iteration t, and let $\mathcal{U}_t = \{1, \ldots, N\} \setminus \mathcal{Q}_t$ denote the remaining scenarios. A surrogate model $\widehat{f}_t$ trained on the available evaluations provides predictions $\widehat{\mathbf{y}}_i^{(t)} = \widehat{f}_t(\mathbf{x}_i)$ for $i \in \mathcal{U}_t$. The corresponding comparison predictions are obtained as $\widehat{\mathbf{\Delta}}_{ij}^{(t)} = \widehat{\mathbf{y}}_j^{(t)} - \widehat{\mathbf{y}}_i^{(t)}$.

At each iteration, the method selects a batch $\mathcal{B}_t \subseteq \mathcal{U}_t$ for evaluation by the original model and updates the surrogate model with the newly obtained outputs. The surrogate model is then used to predict the outcomes of the remaining scenarios and their corresponding comparisons. The objective is to use as few original-model evaluations as possible while maintaining accurate predictions of both individual scenario outcomes and scientifically relevant scenario differences.

Algorithm 1 Comparison-aware active learning framework

Require: Scenario inputs $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, initial dataset $\mathcal{D}_0$, initial queried set $\mathcal{Q}_0$, original model f, comparison set $\mathcal{C}$, evaluation budget K
Ensure: Original-model outputs and surrogate predictions for all scenarios
Construct $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from $\mathcal{X}$ and $\mathcal{C}$
Set $\mathcal{I} \leftarrow \{1, \ldots, N\}$ and $\mathcal{U}_0 \leftarrow \mathcal{I} \setminus \mathcal{Q}_0$
Train $\widehat{f}_0$ on $\mathcal{D}_0$; set $t \leftarrow 0$
while $|\mathcal{Q}_t| < K$ and $\mathcal{U}_t \neq \emptyset$ **do**
    Estimate node and edge uncertainties using $\widehat{f}_t$
    Construct $\mathcal{F}_t$ according to Equation 1
    if $\mathcal{F}_t = \emptyset$ **then**
        break
    end if
    $\mathcal{B}_t \leftarrow \textsc{AcquireBatch}(()\mathcal{F}_t, \mathcal{G}, \widehat{f}_t)$
    if $\mathcal{B}_t = \emptyset$ **then**
        break
    end if
    Evaluate $\mathbf{y}_i = f(\mathbf{x}_i)$ for all $i \in \mathcal{B}_t$
    $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\mathbf{x}_i, \mathbf{y}_i) : i \in \mathcal{B}_t\}$
    $\mathcal{Q}_{t+1} \leftarrow \mathcal{Q}_t \cup \mathcal{B}_t$
    $\mathcal{U}_{t+1} \leftarrow \mathcal{I} \setminus \mathcal{Q}_{t+1}$
    Train $\widehat{f}_{t+1}$ on $\mathcal{D}_{t+1}$; $t \leftarrow t + 1$
end while
Evaluate the remaining fallback scenarios in $\mathcal{U}_t$ with f
Predict all other unresolved scenarios in $\mathcal{U}_t$ with $\widehat{f}_t$
return combined original-model outputs and surrogate predictions

3.2 Comparison-Aware Active Learning Framework

We present a comparison-aware active learning framework for selecting scenarios to be evaluated by the original model. Algorithm 1 summarizes the overall procedure. Starting from the initial dataset $\mathcal{D}_0$, the framework constructs a fixed comparison graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, N\}$ and $\mathcal{E} \subseteq \mathcal{C}$. At round t, an ensemble of M surrogate members $\{\widehat{f}_{m,t}\}_{m=1}^{M}$ is trained or updated using the available original-model evaluations. Their aggregated predictions and uncertainties are then used for both individual scenario outcomes and graph-connected comparisons. The M surrogate members only affect surrogate prediction and uncertainty estimation; querying a scenario still requires one original-model evaluation.

The acquisition target is defined through a fallback set rather than through individual scenario uncertainty alone. Let $u_{i,t}^{\text{node}}$ and $u_{ij,t}^{\text{edge}}$ denote the ensemble-aggregated uncertainties for scenario i and comparison (i, j), respectively. A calibrated joint-risk rule ψ_t maps these uncertainties to a fallback decision:

$$\mathcal{F}_t = \left\{ i \in \mathcal{U}_t : \psi_t\Big(u_{i,t}^{\text{node}}, \Big\{ u_{ij,t}^{\text{edge}} : (i,j) \in \mathcal{E} \Big\}\Big) = 1 \right\}. \tag{1}$$

Thus, $\mathcal{F}_t$ contains the remaining scenarios whose outcomes or incident comparisons have excessive calibrated risk. A query can consequently reduce the fallback status of several scenarios through the comparison edges connected to it.

For a candidate $i \in \mathcal{U}_t$, the acquisition strategy performs a look-ahead update for all M surrogate members and obtains an estimated fallback set $\widehat{\mathcal{F}}_t^{(i)}$. Let $\rho_t \in [0, 1]$ denote the empirical reliability of this estimate. The expected net saving of querying i is

$$g_t(i) = \rho_t \left(|\mathcal{F}_t| - \mathbb{E}\Big[|\widehat{\mathcal{F}}_t^{(i)}| \Big] \right) - 1, \tag{2}$$

where the final term is the cost of one original-model evaluation. Candidates with positive net saving are selected sequentially. If $\mathcal{P}_{t,k}$ denotes the first k candidates in this sequence, the retained prefix maximizes

$$k_t^\star = \arg\max_k \left\{ G_t(k) = \rho_t \left(|\mathcal{F}_t| - \mathbb{E}\left[|\widehat{\mathcal{F}}_t^{(\mathcal{P}_{t,k})}| \right] \right) - k \right\}, \qquad G_t(k_t^\star) > 0. \tag{3}$$

When no positive-saving prefix exists, the framework may construct a cost-neutral exploration batch $\mathcal{E}_t$ from the current fallback scenarios. This batch prioritizes ensemble-uncertainty reduction and graph-aware diversity. It is accepted only when the calibrated uncertainty or the history-adjusted expected future benefit is sufficiently high; otherwise, $\mathcal{B}_t = \emptyset$ and the active-learning loop terminates.

For an accepted batch $\mathcal{B}_t$, the framework evaluates the original model once for each $i \in \mathcal{B}_t$, updates $\mathcal{D}_{t+1}$ and $\mathcal{Q}_{t+1}$, and retrains the M surrogate members. The loop stops when $\mathcal{F}_t = \emptyset$, no admissible batch is found, the candidate set is exhausted, or the evaluation budget is reached. At termination, the remaining fallback scenarios are evaluated by the original model, while the other unresolved scenarios use the ensemble prediction.

3.3 Surrogate Model

At each active-learning round, we train an ensemble of M surrogate members $\{\widehat{f}_{m,t}\}_{m=1}^M$ using the currently available dataset $\mathcal{D}_t$. To handle high-dimensional outputs, each member predicts a low-dimensional latent representation and maps it back to the original output space:

$$\widehat{\mathbf{y}}_{i,m}^{(t)} = \mathcal{D}_t\big(\mathbf{W}_{m,t}^\mathsf{T} \boldsymbol{\phi}_{m,t}(\mathbf{x}_i) + \mathbf{b}_{m,t}\big), \qquad \widehat{\mathbf{y}}_i^{(t)} = \frac{1}{M} \sum_{m=1}^M \widehat{\mathbf{y}}_{i,m}^{(t)}, \tag{4}$$

where $\boldsymbol{\phi}_{m,t}(\mathbf{x}_i)$ is the feature representation of scenario i and $\mathcal{D}_t$ denotes the decoder from the latent output space to the original output space. Each surrogate member provides a feature-space covariance $\boldsymbol{\Sigma}_{m,t}$. The latent uncertainties for a scenario and a comparison edge are computed as $v_{i,m}^{\text{node}} = \boldsymbol{\phi}_{m,t}(\mathbf{x}_i)^\mathsf{T} \boldsymbol{\Sigma}_{m,t} \boldsymbol{\phi}_{m,t}(\mathbf{x}_i)$ and $v_{ij,m}^{\text{edge}} = \boldsymbol{\phi}_{ij,m,t}^\mathsf{T} \boldsymbol{\Sigma}_{m,t} \boldsymbol{\phi}_{ij,m,t}$, where $\boldsymbol{\phi}_{ij,m,t} = \boldsymbol{\phi}_{m,t}(\mathbf{x}_j) - \boldsymbol{\phi}_{m,t}(\mathbf{x}_i)$. The node and edge uncertainties used in the following sections combine the posterior uncertainties of the M members with their prediction disagreement. They are then used to construct the fallback set and evaluate candidate queries.

3.4 Scenario Comparison Graph

To represent the scenario differences considered during active learning, we construct a fixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, N\}$ and $\mathcal{E} \subseteq \mathcal{C}$. The graph is built from the normalized scenario inputs $\widetilde{\mathbf{x}}_i$ and remains unchanged throughout active learning. Its edge set combines four types of comparisons $\mathcal{E} = \mathcal{E}_{\text{mst}} \cup \mathcal{E}_{\text{local}} \cup \mathcal{E}_{\text{global}} \cup \mathcal{E}_{\text{dom}}$.

The set $\mathcal{E}_{\text{mst}}$ is the minimum spanning tree of the scenario inputs, while $\mathcal{E}_{\text{local}}$ connects each scenario to its nearest neighbors. The set $\mathcal{E}_{\text{global}}$ contains pairs sampled from different input-distance strata. For a pair (i, j), let $d_{ij} = \|\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_j\|_2$. Parameter m is considered dominant when its absolute change is sufficiently large and its relative contribution $\omega_{ijm} = d(\widetilde{x}_{im} - \widetilde{x}_{jm})^2 / d_{ij}^2$ exceeds a prescribed level, where d is the number of input parameters. Such pairs form $\mathcal{E}_{\text{dom}}$.The planning graph retains all structural edges and a parameter-balanced coreset of the dominant edges, with duplicate pairs merged. Each edge $(i, j) \in \mathcal{E}$ represents the comparison $\boldsymbol{\Delta}_{ij}$.

3.5 Fallback Strategy

The fallback strategy converts the uncertainty of the M-member surrogate ensemble into a scenario-level decision. The uncertainties used below, $u_{i,t}^{\text{node}}$ and $u_{ij,t}^{\text{edge}}$, are the aggregated node and edge uncertainties defined in Section 3.3. For an unresolved scenario i, let $d_{i,t} = \min_{q \in \mathcal{Q}_t} \|\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{x}}_q\|_2$ be its distance to the queried set, and let $\bar{d}_{ij,t} = (d_{i,t} + d_{j,t})/2$ for edge (i, j). The corresponding uncertainty proxies are

$$a_{i,t}^{\text{node}} = \frac{u_{i,t}^{\text{node}}}{s_{\text{node}}^u} + \frac{d_{i,t}}{s_{\text{node}}^d}, \qquad a_{ij,t}^{\text{edge}} = \frac{u_{ij,t}^{\text{edge}}}{s_{\text{edge}}^u} + \frac{\bar{d}_{ij,t}}{s_{\text{edge}}^d}, \tag{5}$$

where the scale terms are estimated from the fixed calibration subset $\mathcal{D}_{\mathrm{cal}}$. Let e_k^{node} and e_{kl}^{edge} denote the errors of the ensemble-mean predictions on this subset. The calibration multiplier is

$$\kappa = \mathrm{median}_{k \in \mathcal{D}_{\mathrm{cal}}} \frac{\max\left\{e_k^{\mathrm{node}}/\tau_{\mathrm{node}}, \max_{(k,l)\in\mathcal{E}} e_{kl}^{\mathrm{edge}}/\tau_{\mathrm{edge}}\right\}}{\max\left\{a_k^{\mathrm{node}}/\tau_{\mathrm{node}}, \max_{(k,l)\in\mathcal{E}} a_{kl}^{\mathrm{edge}}/\tau_{\mathrm{edge}}\right\} + \epsilon}. \tag{6}$$

The joint severity of an unresolved scenario at round t is then

$$r_{i,t} = \kappa \max\left\{\frac{a_{i,t}^{\mathrm{node}}}{\tau_{\mathrm{node}}}, \max_{(i,j)\in\mathcal{E}} \frac{a_{ij,t}^{\mathrm{edge}}}{\tau_{\mathrm{edge}}}\right\}. \tag{7}$$

The fallback set is $\mathcal{F}_t = \{i \in \mathcal{U}_t : r_{i,t} > 1\}$. Thus, a scenario enters the fallback set when either its own outcome or an incident comparison has excessive calibrated severity. Queried scenarios are excluded because their original-model outputs are already available.

3.6 Active Acquisition Strategy

At round t, each candidate is evaluated through a zero-innovation look-ahead update applied independently to the M surrogate members. For candidate i, the covariance of member m becomes

$$\boldsymbol{\Sigma}_{m,t}^{(i)} = \boldsymbol{\Sigma}_{m,t} - \frac{\boldsymbol{\Sigma}_{m,t}\boldsymbol{\phi}_{m,t}(\mathbf{x}_i)\boldsymbol{\phi}_{m,t}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\Sigma}_{m,t}}{\sigma_m^2 + \boldsymbol{\phi}_{m,t}(\mathbf{x}_i)^{\mathsf{T}}\boldsymbol{\Sigma}_{m,t}\boldsymbol{\phi}_{m,t}(\mathbf{x}_i)}, \qquad m = 1, \ldots, M. \tag{8}$$

For edge (j, k), member m uses the feature difference $\boldsymbol{\phi}_{jk,m,t} = \boldsymbol{\phi}_{m,t}(\mathbf{x}_j) - \boldsymbol{\phi}_{m,t}(\mathbf{x}_k)$. The updated node and edge uncertainties are aggregated across members to estimate the hypothetical fallback set $\widehat{\mathcal{F}}_t^{(i)}$. The expected saving from querying i is $g_t(i) = \rho_t\big(|\mathcal{F}_t| - \mathbb{E}[|\widehat{\mathcal{F}}_t^{(i)}|]\big) - 1$, where ρ_t denotes the empirical reliability of the look-ahead estimate.

Candidates with positive expected savings are selected sequentially. After each virtual selection, the member covariances, queried-status mask, and nearest-query distances are updated before rescoring the remaining candidates. The executed batch is the prefix maximizing the positive batch saving $G_t(k)$ in Equation 3; ties are resolved by uncertainty reduction and then by fallback-severity reduction.

If no positive-saving prefix exists, the method considers a cost-neutral exploration batch from fallback candidates that do not increase the immediate evaluation cost. Let $\Delta_{i,t}^{\mathrm{info}}$ and $\Delta_{i,t}^{\mathrm{sev}}$ denote the reductions in common-support uncertainty and fallback severity, respectively. Candidate quality is defined as $q_{i,t} = \Delta_{i,t}^{\mathrm{info}}/(\max_j \Delta_{j,t}^{\mathrm{info}} + \epsilon) + \epsilon_{\mathrm{sev}}\Delta_{i,t}^{\mathrm{sev}}/(\max_j \Delta_{j,t}^{\mathrm{sev}} + \epsilon)$, and the exploration score is

$$A_{i,t}^{\mathrm{explore}} = q_{i,t}\, D_{i,t}^{\mathrm{post}}\, D_{i,t}^{\mathrm{input}}, \tag{9}$$

where $D_{i,t}^{\mathrm{post}}$ penalizes posterior redundancy and $D_{i,t}^{\mathrm{input}}$ promotes input-space diversity. Candidates are added greedily, and the batch is accepted only when the calibrated ensemble uncertainty remains high or its history-adjusted expected benefit is positive. After each accepted batch, the original model is evaluated and all surrogate members are retrained. The procedure terminates when no batch is accepted, the fallback set is empty, the candidate set is exhausted, or the evaluation budget is reached.

4 Experiments

4.1 Dataset

To evaluate the applicability of the proposed method across a broad range of global change scenarios, we use two datasets. Besides the publicly available energy-scenario dataset (Energy) Holmes et al. (2026), which provides an established test case in a distinct global change context, we develop a new regionally heterogeneous dietary-transition dataset (Diet) to broaden the evaluation while explicitly representing regional heterogeneity.

Following the dietary-transition setup of Rodés-Bachs et al. (Rodés-Bachs et al., 2026), we extend the framework to account for regional heterogeneity by allowing plant-protein transition intensities to vary independently across 32 regions, and construct the dataset by running the GCAM model. All scenarios follow the SSP2 socioeconomic pathway without additional climate-mitigation policies. Each scenario is described by 34 inputs, $\mathbf{x} = (\Delta s_1, \Delta s_2, \ldots, \Delta s_{32}, x_0, k)$, where Δs_r denotes the increase in the plant-protein share in region r, x_0 is the peak transition year, and k controls the transition rate. Since exhaustive simulation over the full parameter space is infeasible, we generate a balanced design of 5,184 scenarios combining homogeneous anchor scenarios with heterogeneous regional combinations; 5,180 are successfully solved and used in subsequent analyses. Each valid scenario spans 32 regions, six five-year time points from 2025 to 2050, and 34 output indicators covering food, land, water, and greenhouse-gas variables, and is treated as an independent sample. The dataset is summarized in Table 1, with further details in Appendix A.

Table 1: Summary of the Diet dataset.

Characteristic	Value
Number of regions	32
Transition-intensity levels	9
Peak-year levels	3
Transition-rate levels	4
Input dimension	34
Valid scenarios	5,180
Time points	6
Output indicators	34
Values per scenario	6,528

4.2 Experimental setup

Baselines We compare our method against six active learning baselines. **Random** draws scenarios uniformly without replacement; **QbC** Seung et al. (1992) selects those with maximal ensemble disagreement; **LCMD** Holzmüller et al. (2023) prioritizes poorly covered clusters in feature space; **SBAL** Kirsch et al. (2023) samples proportionally to temperature-scaled QbC scores; **BIG** MacKay (1992) greedily maximizes mutual information with Bayesian linear-head parameters; and **STAP** Kim et al. (2025) selects uncertain scenarios that cover other uncertain regions.

Evaluation Metrics We report node- and edge-level RMSE and R^2 to assess predicted scenario outcomes and pairwise differences, respectively, and edge sign accuracy, the proportion of comparisons with matching predicted and real-model signs. Sampling efficiency is measured by the proportion of real-model evaluations relative to the full candidate set. All experiments are repeated with five random seeds.

4.3 Main Results

Overall Performance Table 2 compares all methods using the same number of training samples on each dataset. Our method achieves the best performance across all metrics on Energy and most metrics on Diet, demonstrating effective sample selection for both scenario prediction and cross-scenario comparison. The consistent gains in sign accuracy suggest better preservation of the directional relationships needed for scientific interpretation. Random slightly outperforms our method in edge R^2 on Diet. This may reflect a trade-off in our acquisition objective, which considers both outcome and comparison uncertainty to reduce future simulator evaluations, without directly optimizing edge R^2. Nevertheless, our method achieves lower edge RMSE and higher sign accuracy on Diet, supporting its overall comparison quality.

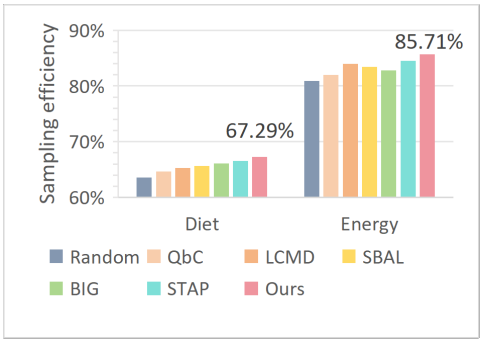


Figure 2: Percentage of original-model evaluations saved under the same stopping conditions.

Simulation Efficiency We apply the same stopping conditions to methods and compare the reduc-

Table 2: Performance with the same number of training scenarios. Within each dataset, all baselines use the same number of training scenarios as our method.

Dataset	Method	Node RMSE ↓	Node R^2 ↑	Edge RMSE ↓	Edge R^2 ↑	Sign accuracy ↑
Diet	Random	0.0437	0.9043	0.0618	**0.9002**	89.33%
	QbC	0.0431	0.8999	0.0609	0.8840	89.03%
	LCMD	0.0447	0.8967	0.0632	0.8987	89.27%
	SBAL	0.0428	0.8988	0.0606	0.8922	89.13%
	BIG	0.0428	0.9021	0.0605	0.8870	89.03%
	STAP	0.0431	0.8972	0.0610	0.8869	89.06%
	Ours	**0.0407**	**0.9099**	**0.0576**	0.8977	**90.22%**
Energy	Random	0.0334	0.9830	0.0471	0.9889	95.06%
	QbC	0.0318	0.9851	0.0448	0.9895	95.16%
	LCMD	0.0322	0.9836	0.0455	0.9895	95.16%
	SBAL	0.0327	0.9853	0.0463	0.9896	95.20%
	BIG	0.0338	0.9823	0.0477	0.9882	94.95%
	STAP	0.0304	0.9843	0.0430	0.9902	95.36%
	Ours	**0.0302**	**0.9903**	**0.0426**	**0.9903**	**96.65%**

Table 3: Absolute-error quantiles on Diet with the same number of training scenarios. All values are expressed in units of 10^{-2}.

Method	Node P50	Node P95	Node P99	Edge P50	Edge P95	Edge P99
Random	0.0728	7.3001	18.2058	0.1287	9.9440	25.7624
QbC	0.0743	7.5067	18.5076	0.1293	10.0069	25.6933
LCMD	0.0721	7.2698	18.0915	0.1268	9.8516	25.6754
SBAL	0.0737	7.3825	18.1225	0.1289	9.8695	25.6421
BIG	0.0746	7.4362	18.1070	0.1292	9.8584	25.3828
STAP	0.0733	7.3952	18.0987	0.1279	9.8504	25.1575
Ours	**0.0715**	**7.2220**	**17.8110**	**0.1255**	**9.6860**	**24.9980**

tion in original-model evaluations. As shown in Figure 2, our method achieves the largest reduction on both datasets, indicating its acquisition strategy reduces simulation requirements more effectively.

Error Distribution Table 3 further shows that the advantage of our method extends across the error distribution rather than being limited to average performance. Under equal training budgets on Diet, our method achieves the lowest mean errors across all reported node and edge quantiles. This improvement extends from the median to the upper tail, suggesting that the benefits are not limited to typical predictions but also include cases with larger errors. The lower upper-tail edge errors further support the method's ability to preserve cross-scenario differences.

Computational Complexity The comparison graph is constructed once in $O(N^2 d)$ time. At round t, evaluating all candidate queries costs $O(M|\mathcal{U}_t|(|\mathcal{U}_t| + |\mathcal{E}|))$. Sequential construction of $\mathcal{B}_t$ requires at most $O(M|\mathcal{B}_t||\mathcal{U}_t|(|\mathcal{U}_t| + |\mathcal{E}|))$ time, whereas exploration reuses a single candidate scan. Since the planning graph is sparse, the acquisition complexity is quadratic in N when M and the batch size are bounded.

4.4 Ablation Study

We evaluate three ablations on the energy dataset: *No-Acquisition* replaces our acquisition strategy with uniform sampling over unresolved scenarios, *No-Edges* removes edge information from acquisition and fallback, and *No-Fallback* keeps the node–edge threshold but drops the net-call-saving stopping criterion. Figure 3 reports five-seed means with shaded standard deviations. *No-Acquisition* improves more slowly, confirming informed selection helps; *No-Edges* stops early after meeting node criteria but leaves comparison errors unaddressed; *No-Fallback* samples longer without surpassing *Full*. Overall, *Full* attains the lowest final node and edge log RMSE and highest sign accuracy, supporting comparison-aware acquisition and cost-aware stopping.

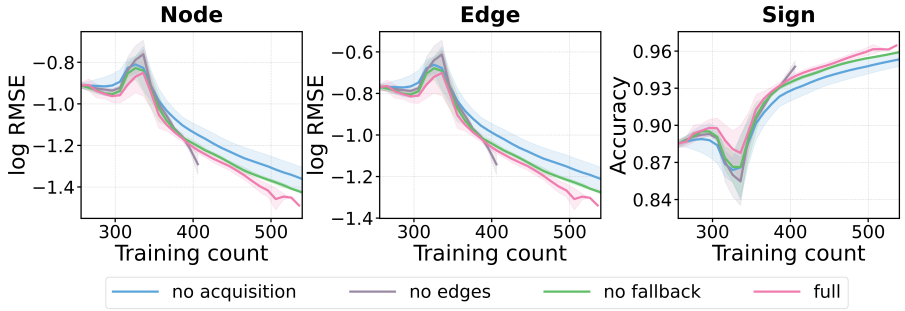


Figure 3: Ablation results on the energy dataset. Node log RMSE, edge log RMSE, and edge sign accuracy are shown as the number of training scenarios increases.

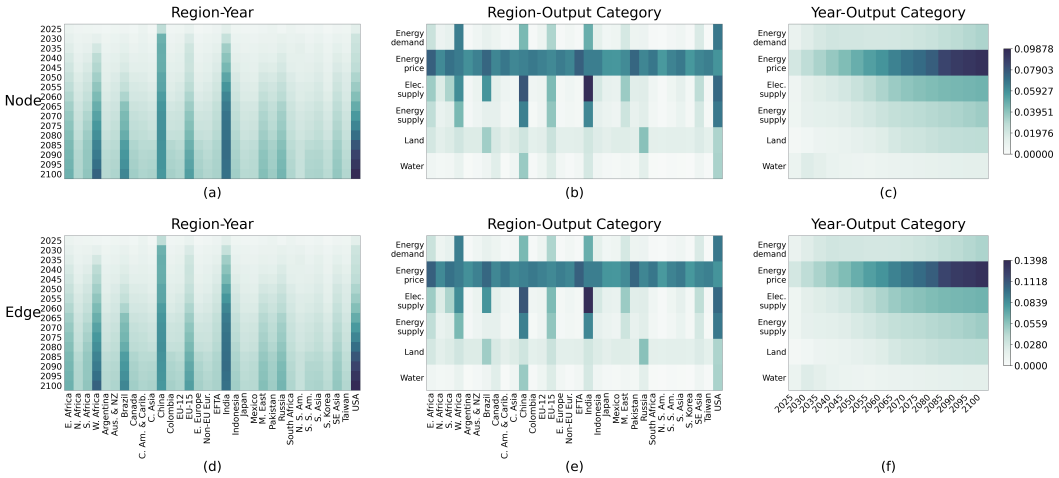


Figure 4: RMSE on the energy dataset for nodes (top) and edges (bottom). Columns show region–year, region–output category, and year–output category aggregations.

4.5 Quantitative Analysis

Figure 4 presents node- and edge-level RMSE on the energy dataset, averaged over five random seeds and aggregated by region–year, region–output category, and year–output category. Node and edge errors show broadly consistent patterns, generally increasing over time. Energy prices exhibit the largest errors and the clearest temporal increase, while electricity and energy supply show stronger regional heterogeneity. Land and water outputs generally have lower errors. These patterns suggest that scenario prediction and cross-scenario comparison share similar sources of difficulty. Corresponding results on the Diet dataset are provided in the Appendix C.3.

5 Conclusion

In this work, we developed a comparison-aware active learning framework for global change analysis, which selectively acquires original-model evaluations while preserving the outcome patterns and cross-scenario relationships needed for scientific interpretation. By representing scenario relationships explicitly and coordinating uncertainty assessment with adaptive scenario selection, the framework provides a principled way to explore large scenario spaces under limited computational resources. Future work could incorporate more domain-specific and dynamically evolving comparison structures, and investigate the framework under stronger distribution shifts, higher-dimensional outputs, and richer spatial and temporal dependencies.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets or cleaning and reformatting datasets, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, draft parts of a research paper, summarize or analyse existing literature, source or search for information, edit a research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. All AI-assisted text, code, and literature summaries were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code implementing our proposed method in the supplementary materials.

REFERENCES

Hamish Beath, Chris Smith, Jarmo S Kikstra, Mark M Dekker, Matthew J Gidden, and Joeri Rogelj. A weighting framework to improve the use of emissions scenario ensembles of opportunity. *Nature Climate Change*, 16(3):305–312, 2026.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Katherine Calvin, Pralit Patel, Leon Clarke, Ghassem Asrar, Ben Bond-Lamberty, Ryna Yiyun Cui, Alan Di Vittorio, Kalyn Dorheim, Jae Edmonds, Corinne Hartin, et al. Gcam v5. 1: representing the linkages between energy, water, land, climate, and economic systems. *Geoscientific Model Development*, 12(2):677–698, 2019.

Mark M Dekker, Vassilis Daioglou, Robert Pietzcker, Renato Rodrigues, Harmen-Sytze De Boer, Francesco Dalla Longa, Laurent Drouet, Johannes Emmerling, Amir Fattahi, Theofano Fotiou, et al. Identifying energy model fingerprints in mitigation scenarios. *Nature Energy*, 8(12):1395–1404, 2023a.

Mark M Dekker, Andries F Hof, Maarten van den Berg, Vassilis Daioglou, Rik van Heerden, Kaj-Ivar van der Wijst, and Detlef P van Vuuren. Spread in climate policy scenarios unravelled. *Nature*, 624(7991):309–316, 2023b.

Shinichiro Fujimori, Volker Krey, Keywan Riahi, Masahiro Sugiyama, Tomoko Hasegawa, James Edmonds, Celine Guivarch, Sergey Paltsev, Steven Rose, Roberto Schaeffer, et al. Towards an open model intercomparison platform for integrated assessment models scenarios. *Nature Climate Change*, 15(11):1156–1164, 2025.

Yarin Gal, Petros Koumoutsakos, Francois Lanusse, Gilles Louppe, and Costas Papadimitriou. Bayesian uncertainty quantification for machine-learned models in physics. *Nature Reviews Physics*, 4(9):573–577, 2022.

Céline Guivarch, Thomas Le Gallic, Nico Bauer, Panagiotis Fragkos, Daniel Huppmann, Marc Jaxa-Rozen, Ilkka Keppo, Elmar Kriegler, Tamás Krisztin, Giacomo Marangoni, et al. Using large ensembles of climate change mitigation scenarios for robust insights. *Nature Climate Change*, 12 (5):428–435, 2022.

Andrew Holmes, Hidemi Mitani Shen, Matt Jensen, Sarah Coffland, Logan Sizemore, Seth Bassetti, Brenna Nieva, Claudia Tebaldi, Abigail Snyder, and Brian Hutchinson. Emulating the global change analysis model with deep learning: An energy sector case study. *Environmental Modelling & Software*, pp. 106945, 2026.

David Holzmüller, Viktor Zaverkin, Johannes Kästner, and Ingo Steinwart. A framework and benchmark for deep batch active learning for regression. *Journal of Machine Learning Research*, 24 (164):1–81, 2023.

Daniel Huppmann, Joeri Rogelj, Elmar Kriegler, Volker Krey, and Keywan Riahi. A new scenario resource for integrated 1.5 c research. *Nature climate change*, 8(12):1027–1030, 2018.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Yegon Kim, Hyunsu Kim, Gyeonghoon Ko, and Juho Lee. Active learning with selective time-step acquisition for PDEs. In *Forty-second International Conference on Machine Learning*, 2025.

Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frédéric Branchaud-Charron, and Yarin Gal. Stochastic batch acquisition: A simple baseline for deep active learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Robert J Lempert and Myles T Collins. Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches. *Risk Analysis: An International Journal*, 27(4):1009–1026, 2007.

Peijin Li, Rongqi Zhu, Haewon McJeon, Edward Byers, Peijie Zhou, and Yang Ou. Using deep learning to generate key variables in global mitigation scenarios. *Nature Climate Change*, 15(7): 760–768, 2025.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Giacomo Marangoni, Massimo Tavoni, Valentina Bosetti, Emanuele Borgonovo, Pantelis Capros, Oliver Fricko, David EHJ Gernaat, Céline Guivarch, Petr Havlik, Daniel Huppmann, et al. Sensitivity of projected long-term co2 emissions across the shared socioeconomic pathways. *Nature Climate Change*, 7(2):113–117, 2017.

Daniel Musekamp, Marimuthu Kalimuthu, David Holzmüller, Makoto Takamoto, and Mathias Niepert. Active learning for neural pde solvers. In *International Conference on Learning Representations*, volume 2025, pp. 38777–38825, 2025.

Brian C O'Neill, Timothy R Carter, Kristie Ebi, Paula A Harrison, Eric Kemp-Benedict, Kasper Kok, Elmar Kriegler, Benjamin L Preston, Keywan Riahi, Jana Sillmann, et al. Achievements and needs for the climate change scenario framework. *Nature climate change*, 10(12):1074–1084, 2020.

Amandine V Pastor, DCS Vieira, Floor H Soudijn, and Oreane Yasmin Edelenbosch. How uncertainties are tackled in multi-disciplinary science? a review of integrated assessments under global change. *Catena*, 186:104305, 2020.

Amandine Valérie Pastor, Amanda Palazzo, Petr Havlik, Hester Biemans, Yoshihide Wada, M Obersteiner, Pavel Kabat, and Fulco Ludwig. The global nexus of food–trade–water sustaining environmental flows by 2050. *Nature Sustainability*, 2(6):499–507, 2019.

Stefan Pauliuk, Anders Arvesen, Konstantin Stadler, and Edgar G Hertwich. Industrial ecology in integrated assessment models. *Nature Climate Change*, 7(1):13–20, 2017.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Keywan Riahi, Detlef P Van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C O'neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, et al. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global environmental change*, 42:153–168, 2017.

Clàudia Rodés-Bachs, Jon Sampedro, Dirk-Jan Van de Ven, Russell Horowitz, Guillermo Pardo, and Xin Zhao. Region-specific and nutritionally adequate dietary transitions can bolster sustainability and socioeconomic benefits. *Nature Food*, 7(3):247–259, 2026.

Joeri Rogelj, Alexander Popp, Katherine V Calvin, Gunnar Luderer, Johannes Emmerling, David Gernaat, Shinichiro Fujimori, Jessica Strefler, Tomoko Hasegawa, Giacomo Marangoni, et al. Scenarios towards limiting global mean temperature increase below 1.5 c. *Nature climate change*, 8(4):325–332, 2018.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Peter Van Katwyk, Baylor Fox-Kemper, Helene T Hewitt, and Karianne J Bergen. Rewiring climate modeling with machine learning emulators. *Communications Earth & Environment*, 7(1):107, 2026.

Detlef P Van Vuuren, Jae Edmonds, Mikiko Kainuma, Keywan Riahi, Allison Thomson, Kathy Hibbard, George C Hurtt, Tom Kram, Volker Krey, Jean-Francois Lamarque, et al. The representative concentration pathways: an overview. *Climatic change*, 109(1):5, 2011.

Detlef P Van Vuuren, David L Bijl, Patrick Bogaart, Elke Stehfest, Hester Biemans, Stefan C Dekker, Jonathan C Doelman, David EHJ Gernaat, and Mathijs Harmsen. Integrated scenarios to support analysis of the food–energy–water nexus. *Nature Sustainability*, 2(12):1132–1141, 2019.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Jacob Wessel, Gokul Iyer, Thomas Wild, Yang Ou, Haewon McJeon, and Jonathan Lamontagne. Large ensemble exploration of global energy transitions under national emissions pledges. *Earth's Future*, 12(10):e2024EF004754, 2024.

A Supplementary Details on Datasets

A.1 Construction of the Regionally Heterogeneous Dietary-Transition Dataset

A.1.1 Model setting and input parameterization

The dietary-transition dataset was constructed by extending the dietary-transition setting of Rodés-Bachs et al. (Rodés-Bachs et al., 2026) to represent regional heterogeneity. We used the GCAM model to run all scenarios under the SSP2 socioeconomic pathway without additional climate-mitigation policies. The dataset contains 32 regions, and the plant-protein transition intensity is allowed to vary independently across regions. Each scenario is defined by 34 inputs, $\mathbf{x} = (\Delta s_1, \Delta s_2, \ldots, s_{32}, x_0, k)$, where Δs_r denotes the increase in the plant-protein share in region r,

Table 4: Output composition of the dietary-transition dataset.

Category	Number	Description
Food consumption	21	Crop, legume, nut, dairy, meat, and fish consumption indicators
Food demand and prices	4	Staple-food demand, non-staple-food demand, and corresponding prices
Land use	4	Cropland, pasture, forest, and other natural land
Water use	1	Regional water withdrawals
Greenhouse-gas emissions	4	CO_2, land-use-change CO_2, CH_4, and N_2O

x_0 is the year in which the transition rate reaches its maximum, and k controls the transition rate. The regional transition intensity takes nine levels, $\Delta s_r \in \{10, 20, \ldots, 90\}$, measured in percentage points. The peak year takes the values 2025, 2035, and 2045, while k takes the values 0.005, 0.405, 0.805, and 1.205.

For each region, the target plant-protein share was obtained by adding Δs_r to the corresponding SSP2 reference value and capping the result at 100%. The transition pathway from the historical value to this target was determined by x_0 and k. These parameters modify the preference weights associated with plant-protein consumption rather than directly prescribing final food consumption; food demand, prices, trade, land allocation, water use, and greenhouse-gas emissions are therefore determined jointly by the GCAM model.

A.1.2 Scenario design

The complete parameter space contains $3 \times 4 \times 9^{32} \approx 4.1 \times 10^{31}$ possible combinations and cannot be evaluated exhaustively. We therefore constructed a balanced design of 5,184 scenarios. The three peak years and four transition-rate levels form 12 temporal configurations, each containing 432 scenarios so that all peak-year and transition-rate combinations are equally represented. Within each temporal configuration, we included nine homogeneous anchor scenarios in which all 32 regions use the same transition intensity, resulting in 108 homogeneous anchor scenarios overall; the remaining 5,076 scenarios contain heterogeneous regional combinations.

To ensure balanced coverage of regional transition intensities, we imposed the constraint $N(\Delta s_r = l \mid x_0, k) = 48$ for each region r, each temporal configuration (x_0, k), and each transition level $l \in \{10, 20, \ldots, 90\}$. Thus, within each temporal configuration, every transition level is assigned to region r in exactly 48 scenarios. Across the 12 temporal configurations, each level consequently occurs 576 times for each region. Because the assignments are made independently across regions, this marginal balance preserves substantial regional heterogeneity within individual scenarios.

A.1.3 Model execution and outputs

Each designed scenario was converted into a GCAM input configuration and run as a complete model case. Of the 5,184 designed scenarios, 5,180 were successfully solved; the four unsuccessful cases were excluded because of timeouts or incomplete model outputs. Each valid scenario contains 32 regions, six five-year time points from 2025 to 2050, and 34 output indicators covering food consumption, food demand and prices, land use, water use, and greenhouse-gas emissions. The output structure of one valid scenario is therefore $32 \times 6 \times 34 = 6{,}528$ region–year–indicator values. The 34 output indicators are grouped as shown in Table 4.

The complete scenario, rather than an individual region–year record, is treated as the independent sample. The 5,180 valid scenarios are divided into training and test sets at a ratio of 9:1, resulting in 4,662 training scenarios and 518 test scenarios. All records belonging to the same scenario are assigned to the same subset to prevent information leakage across the two sets.

A.2 Public Energy-Scenario Dataset

To evaluate the method in a different global-change context, we used the publicly available energy-scenario dataset developed by Holmes et al. (Holmes et al., 2026). The dataset contains 12 scenario-

Table 5: Comparison of the two datasets used in this study.

Characteristic	Dietary-transition dataset	Energy-scenario dataset
Global change context	Regional dietary transition	Energy technology and socioeconomic change
Data source	GCAM runs conducted in this study	Public dataset of Holmes et al.
Sampling strategy	Balanced regional-combination design	Latin hypercube sampling
Input dimension	34	12
Regional input structure	Independent transition intensity across 32 regions	Scenario-level factors jointly define each global scenario
Valid scenarios	5,180	4,010
Number of regions	32	32
Time range	2025–2050	2025–2100
Number of time points	6	16
Output indicators	34	44
Values per scenario	6,528	22,528

level input variables describing changes in backup capacity, bioenergy constraints, carbon capture and storage, electrification, emissions constraints, socioeconomic development, fossil-fuel use, nuclear energy, and solar and wind technologies. Nine input variables are continuously interpolated between their low and high settings, whereas the remaining three variables take binary values. The published dataset includes a Latin hypercube subset designed to provide broad coverage of the continuous input space, which we used directly together with its published input variables, data partition, and model outputs.

After unsuccessful model runs were removed, the subset contained 4,010 valid scenarios. The published data partition includes 3,209 training scenarios, 401 validation scenarios, and 400 test scenarios, with the partition defined at the scenario level so that all outputs associated with a given scenario remain in the same subset. Each scenario contains 32 regions, 44 energy, land, and water indicators, and 16 five-year time points from 2025 to 2100. The output dimension of one scenario is therefore $32 \times 44 \times 16 = 22{,}528$ values. A comparison of the main characteristics of the two datasets is provided in Table 5.

B Supplementary Experimental Details

B.1 Baselines

To isolate the effect of the acquisition strategy, all baseline methods use the same surrogate architecture, training procedure, initial labeled set, and acquisition budget. The methods differ only in how they select new scenarios from the remaining unlabeled training pool.

Random Sampling Random Sampling uniformly selects scenarios from the unlabeled pool without replacement. It does not use surrogate predictions, uncertainty estimates, or learned representations, and therefore provides a model-independent reference for evaluating the benefits of informed acquisition.

Query-by-CommitteeSeung et al. (1992) Query-by-Committee (QbC) selects scenarios for which the members of the surrogate ensemble produce the most inconsistent predictions. The disagreement is measured in the latent output space and aggregated across output dimensions. Scenarios with greater committee disagreement are regarded as having higher epistemic uncertainty and are prioritized for acquisition.

Largest Cluster Maximum DistanceHolzmüller et al. (2023) Largest Cluster Maximum Distance (LCMD) performs acquisition in the feature space learned by the surrogate ensemble. Previously acquired scenarios are treated as cluster centers, while each unlabeled scenario is assigned to

its nearest center. LCMD repeatedly selects a distant scenario from the most poorly covered cluster and adds it as a new center, thereby promoting both coverage of the input space and diversity within each acquisition batch.

Stochastic Batch Active LearningKirsch et al. (2023) Stochastic Batch Active Learning (SBAL) converts an acquisition function into a probability distribution over the unlabeled pool. In our implementation, ensemble disagreement from QbC is used as the underlying acquisition score. Scenarios are sampled without replacement according to their temperature-scaled scores, retaining a preference for uncertain scenarios while reducing the redundancy that can arise from deterministic top-score selection.

Bayesian Information GainMacKay (1992) Bayesian Information Gain (BIG) selects scenarios according to the expected information they provide about the parameters of the Bayesian linear output heads. Candidates with high posterior predictive variance provide greater information gain and are selected preferentially. The posterior covariance is updated after each selected scenario during batch construction, which discourages the acquisition of scenarios that provide redundant information.

STAP-Inspired Scenario AcquisitionKim et al. (2025) Selective Time-Step Acquisition for PDEs (STAP) was originally developed to select informative time steps within dynamically simulated trajectories. Because GCAM scenarios are acquired as complete simulations rather than independently queryable time steps, we employ a scenario-level adaptation denoted STAP-S. STAP-S combines ensemble uncertainty with similarity in the learned representation space and progressively discounts regions already covered by the current batch, encouraging the selection of uncertain yet mutually complementary scenarios. This method is a task-specific adaptation of STAP rather than an exact reproduction of its original time-step acquisition procedure.

B.2 Detailed Evaluation Metrics

Let $\mathcal{S}$ denote the set of evaluated scenarios, and let $\mathbf{y}_i \in \mathbb{R}^D$ and $\hat{\mathbf{y}}_i \in \mathbb{R}^D$ denote the real-model and surrogate-model outputs for scenario i, respectively. Here, D is the total number of scalar output coordinates across all reported variables, regions, and time steps.

Because the output variables have different numerical scales, the real and predicted outputs are standardized using the same per-variable normalization statistics before calculating the aggregated regression metrics. For simplicity, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ below denote these standardized outputs.

The node-level root mean squared error is calculated as

$$\mathrm{RMSE}_{\mathrm{node}} = \sqrt{\frac{1}{|\mathcal{S}|D} \sum_{i \in \mathcal{S}} \sum_{d=1}^{D} \left(\hat{y}_{i,d} - y_{i,d}\right)^2}. \tag{10}$$

To prevent output coordinates with large between-coordinate scale differences from dominating the coefficient of determination, R^2 is first calculated separately for each output coordinate. For coordinate d, the node-level coefficient of determination is

$$R^2_{\mathrm{node},d} = 1 - \frac{\sum_{i \in \mathcal{S}} \left(\hat{y}_{i,d} - y_{i,d}\right)^2}{\sum_{i \in \mathcal{S}} \left(y_{i,d} - \bar{y}_d\right)^2}, \tag{11}$$

where

$$\bar{y}_d = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} y_{i,d}. \tag{12}$$

Let $\mathcal{D}_{\mathrm{node}}$ denote the set of output coordinates with nonzero real-model variance over the evaluated scenarios:

$$\mathcal{D}_{\mathrm{node}} = \left\{ d : \sum_{i \in \mathcal{S}} \left(y_{i,d} - \bar{y}_d\right)^2 > 0 \right\}. \tag{13}$$

The reported node-level R^2 is the macro-average over these valid output coordinates:

$$R^2_{\text{node}} = \frac{1}{|\mathcal{D}_{\text{node}}|} \sum_{d \in \mathcal{D}_{\text{node}}} R^2_{\text{node},d}. \tag{14}$$

For edge-level evaluation, let $\mathcal{E}$ denote the set of evaluated scenario pairs. For each edge $(i, j) \in \mathcal{E}$, the real and predicted pairwise differences are defined as

$$\Delta y_{ij,d} = y_{j,d} - y_{i,d}, \qquad \Delta \hat{y}_{ij,d} = \hat{y}_{j,d} - \hat{y}_{i,d}. \tag{15}$$

The edge-level root mean squared error is calculated as

$$\text{RMSE}_{\text{edge}} = \sqrt{\frac{1}{|\mathcal{E}|D} \sum_{(i,j)\in\mathcal{E}} \sum_{d=1}^{D} \left(\Delta \hat{y}_{ij,d} - \Delta y_{ij,d}\right)^2}. \tag{16}$$

For each output coordinate d, the edge-level coefficient of determination is

$$R^2_{\text{edge},d} = 1 - \frac{\sum_{(i,j)\in\mathcal{E}} \left(\Delta \hat{y}_{ij,d} - \Delta y_{ij,d}\right)^2}{\sum_{(i,j)\in\mathcal{E}} \left(\Delta y_{ij,d} - \overline{\Delta y}_d\right)^2}, \tag{17}$$

where

$$\overline{\Delta y}_d = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \Delta y_{ij,d}. \tag{18}$$

Let $\mathcal{D}_{\text{edge}}$ denote the set of output coordinates with nonzero variance in their real-model pairwise differences:

$$\mathcal{D}_{\text{edge}} = \left\{ d : \sum_{(i,j)\in\mathcal{E}} \left(\Delta y_{ij,d} - \overline{\Delta y}_d\right)^2 > 0 \right\}. \tag{19}$$

The reported edge-level R^2 is the corresponding macro-average:

$$R^2_{\text{edge}} = \frac{1}{|\mathcal{D}_{\text{edge}}|} \sum_{d \in \mathcal{D}_{\text{edge}}} R^2_{\text{edge},d}. \tag{20}$$

Edge sign accuracy measures the proportion of scalar pairwise differences for which the surrogate predicts the same sign as the real model:

$$\text{Acc}_{\text{sign}} = \frac{1}{|\mathcal{E}|D} \sum_{(i,j)\in\mathcal{E}} \sum_{d=1}^{D} \mathbb{I}\left[\text{sgn}\left(\Delta \hat{y}_{ij,d}\right) = \text{sgn}\left(\Delta y_{ij,d}\right)\right], \tag{21}$$

where $\mathbb{I}[\cdot]$ is the indicator function and

$$\text{sgn}(x) = \begin{cases} +1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases} \tag{22}$$

A prediction is therefore counted as correct when the predicted and real-model differences are both positive, both negative, or both exactly zero. Edge sign accuracy is equivalent to one minus the edge sign error rate.

Sampling efficiency is quantified using both the number of scenarios evaluated by the real model, N_{real}, and its proportion relative to exhaustive sampling:

$$\rho_{\text{real}} = \frac{N_{\text{real}}}{N_{\text{full}}}, \tag{23}$$

where N_{real} is the total number of scenarios evaluated using the real model, including any fixed reference scenarios when applicable, and N_{full} is the number of real-model evaluations required by exhaustive sampling. A smaller value of ρ_{real} indicates greater simulation savings.

Table 6: Number of training scenarios used by the four ablation variants on the energy dataset at the end of active learning. Values are reported as mean ± standard deviation over five random seeds.

Method	Training scenarios
No Acquisition	652.8 ± 97.8
No Edges	410.0 ± 4.8
No Fallback	603.6 ± 17.5
Full	516.2 ± 13.9

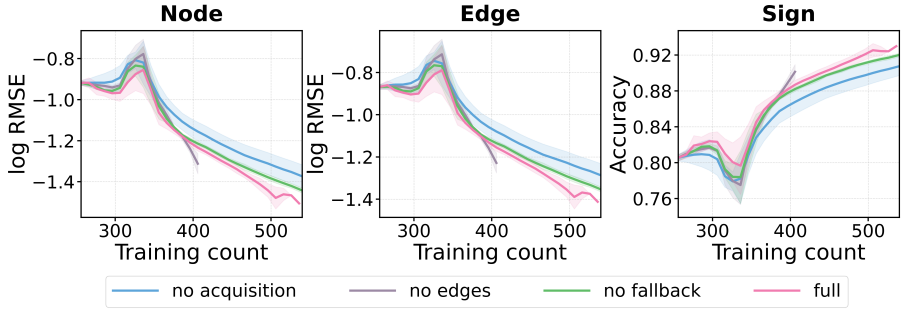


Figure 5: Ablation results on the remaining training set of the energy dataset. From left to right, the panels show node log RMSE, edge log RMSE, and sign accuracy as a function of the training-scenario count. Curves show the mean over five random seeds, and shaded regions indicate the standard deviation. Lower values are better for node and edge log RMSE, whereas higher values are better for sign accuracy.

C Extended Experiments

C.1 Additional Ablation Results

This appendix reports the number of training scenarios used by each ablation variant and its performance on the remaining training scenarios of the Energy dataset. The results are averaged over five random seeds. Table 6 summarizes the number of selected training scenarios at the end of active learning as mean ± standard deviation.

Figure 5 shows the corresponding performance on the remaining training scenarios. The three panels report node log RMSE, edge log RMSE, and edge sign accuracy as functions of the number of training scenarios. Curves show means over five random seeds, and shaded regions show standard deviations. Lower values indicate better performance for node and edge log RMSE, whereas higher values indicate better performance for edge sign accuracy.

The *Full* method achieves the best node and edge log RMSE and the highest edge sign accuracy in the late stages. *No-Acquisition* reaches comparable performance only after using substantially more training scenarios. *No-Edges* terminates earlier under the node-based stopping criterion, because edge information is excluded from acquisition and fallback decisions. Consequently, it cannot further address relational errors after the node criteria are satisfied. *No-Fallback* uses more training scenarios than *Full* to reach comparable performance, suggesting that the fallback mechanism improves sample efficiency.

C.2 Classification of Output Metrics for Two Datasets

The output variables were grouped according to their physical and socioeconomic meanings. The Energy dataset was divided into six categories: energy demand, energy price, electricity supply, primary energy supply, land, and water. The Diet dataset was divided into six categories: livestock, crops, food demand and price, land, emissions, and water. Table 7 summarizes these output cate-

gories and their included variables. These categories were used for category-wise aggregation in the quantitative analysis.

Table 7: Classification of output variables in the Energy and Diet datasets.

Dataset and category	Included outputs
Energy: Energy demand	Electricity demand in buildings, industry, and transport; biomass-fuel demand in buildings and industry; and fossil-fuel demand in buildings, industry, and transport.
Energy: Energy price	Coal, electricity, gas, and oil prices.
Energy: Electricity supply	Electricity supplied from biomass, coal, gas, nuclear, oil, other sources, solar, and wind.
Energy: Energy supply	Primary energy supplied from biomass, coal, gas, nuclear, oil, other sources, solar, and wind.
Energy: Land	Land allocation, land demand for food and feed, land prices, and land production by type.
Energy: Water	Water demand from crop production and electricity generation.
Diet: Livestock	Beef, dairy, other meat and fish, pork, poultry, and sheep and goat products.
Diet: Crops	Corn, fiber crops, fruits, legumes, miscellaneous crops, nuts and seeds, oil crops, grains, rice, roots and tubers, soybean, sugar crops, vegetables, and wheat.
Diet: Food demand and price	Staple and non-staple food demand and their corresponding prices.
Diet: Land	Cropland, forest, other natural land, and pasture.
Diet: Emissions	CH_4, CO_2, land-use-change CO_2, and N_2O.
Diet: Water	Water withdrawals.

C.3 Analysis of Diet Results

Table 8 reports node- and edge-level performance on the Diet dataset by output category. Water and land achieve the highest R^2 values, while emissions have the lowest R^2 and edge sign accuracy. Livestock has the largest RMSE, but its R^2 is higher than those of crops and food demand and price; therefore, it is not consistently the least predictable category. The category rankings are broadly consistent between node- and edge-level predictions. Edge sign accuracy remains high across all categories, and the small standard deviations indicate stable performance across random seeds. Because RMSE depends on output scale, cross-category comparisons should primarily rely on R^2 and sign accuracy unless all outputs are standardized.

Because the Diet outputs vary across regions, years, and output categories, RMSE was aggregated by region–year, region–output category, and year–output category. Figure 6 presents node- and edge-level RMSE averaged over five random seeds. The node- and edge-level results show broadly similar spatial, temporal, and output-specific patterns. Prediction errors vary across regions, suggesting that model accuracy is not uniform across the spatial domain. In the reported scale, livestock and emissions tend to have larger RMSE values, whereas land and water tend to have smaller values. Temporal patterns also differ across output categories. The similarity between the node- and edge-level heatmaps suggests that the two prediction tasks share several difficult regions and output categories.

Table 8: Node- and edge-level performance on the diet dataset by output category. Values are reported as the mean ± standard deviation over five random seeds.

Output category	Node RMSE	Node R^2	Edge RMSE	Edge R^2	Edge sign accuracy (%)
Livestock	0.07104 ± 0.00118	0.9130 ± 0.0026	0.10045 ± 0.00166	0.9013 ± 0.0029	90.47 ± 0.12
Crops	0.02202 ± 0.00013	0.9061 ± 0.0017	0.03114 ± 0.00017	0.8952 ± 0.0018	90.12 ± 0.08
Food demand & price	0.04070 ± 0.00068	0.9040 ± 0.0018	0.05754 ± 0.00096	0.8916 ± 0.0021	89.54 ± 0.09
Land	0.00425 ± 0.00003	0.9295 ± 0.0030	0.00601 ± 0.00005	0.9191 ± 0.0034	91.44 ± 0.12
Emissions	0.05496 ± 0.00082	0.8977 ± 0.0007	0.07767 ± 0.00116	0.8777 ± 0.0008	89.29 ± 0.03
Water	0.00526 ± 0.00004	0.9436 ± 0.0033	0.00743 ± 0.00005	0.9339 ± 0.0037	91.90 ± 0.10
Overall	0.04072 ± 0.00045	0.9099 ± 0.0019	0.05757 ± 0.00062	0.8977 ± 0.0021	90.22 ± 0.08

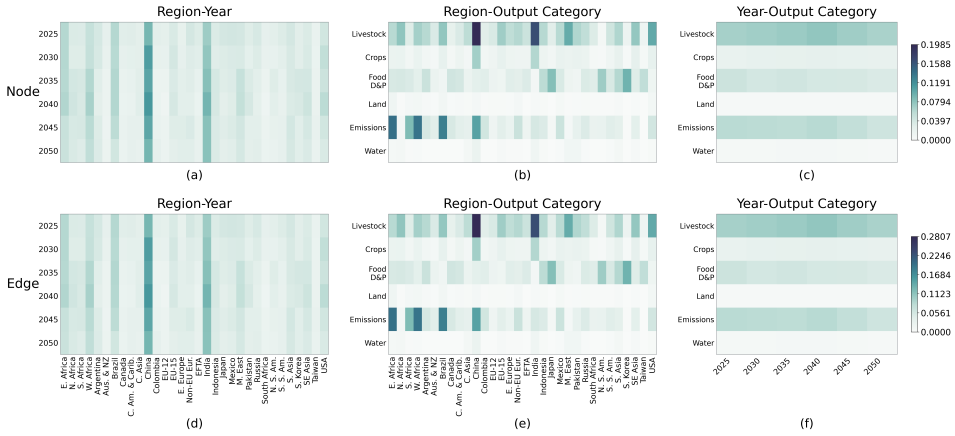


Figure 6: Aggregated node- and edge-level RMSE on the Diet dataset. The top and bottom rows show node- and edge-level results, respectively. The columns present errors aggregated by region–year, region–output category, and year–output category. Results are averaged over five random seeds.