GUST: GRAPH-GUIDED CONTINUAL ADAPTATION OF WEATHER AND CLIMATE FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Weather and climate foundation models must accommodate evolving prediction needs, yet existing adaptation typically assumes a static lifecycle in which each task is learned independently from the same pretrained checkpoint. We systematically investigate continual learning (CL) under sequential access to task data across three settings: variable CL expands atmospheric targets, regional CL extends geographical coverage, and climate-projection CL introduces new response objectives. We find that forgetting is strongly scenario-dependent. Direct sequential adaptation remains effective under compatible task shifts but severely degrades earlier capabilities when prediction objectives change. Cross-task transfer is also uneven and directional. Motivated by these findings, we propose Graph-guided Update Selection and Transfer (**GUST**), which stores acquired knowledge in frozen low-rank adapters and selectively composes their updates through a causal task graph guided by structured relations, learned task geometry, and online transfer feedback. GUST achieves the strongest overall acquisition-retention trade-off among CL methods. In variable CL, it obtains an average relative wRMSE of 1.10 and reduces relative forgetting from 39.55% to 9.67%. In regional CL, it achieves a relative wRMSE of 1.04 with zero measured forgetting. In climate-projection CL, it achieves the lowest oracle-normalized error on three of four metrics. These results identify when dedicated CL is necessary and provide an effective avenue for long-term evolution of weather and climate foundation models.

1 INTRODUCTION

Data-driven weather and climate modeling is advancing from high-performing task-specific systems toward general-purpose foundation models Chen et al. (2023). GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), and GenCast (Price et al., 2025) have achieved competitive performance in global medium-range forecasting. ClimaX (Nguyen et al., 2023) and Aurora (Bodnar et al., 2025) learn reusable representations across variables, resolutions, datasets, and prediction tasks, while NeuralGCM (Kochkov et al., 2024) integrates learned components with physical dynamics. Despite this progress, these models are developed under a *static* lifecycle, in which each downstream task is adapted independently from the same pretrained checkpoint. In practice, their operating scope can expand as new variables become available, forecasting services extend to new geographical regions, and climate studies introduce new response targets. Retraining on accumulated data becomes increasingly costly, while maintaining a finetuned model for each task limits cross-task knowledge reuse and incurs growing storage costs. A shared foundation model that can continually acquire, preserve, and reuse capabilities is therefore important for scalable long-term deployment.

We study continual learning (CL) of weather and climate foundation models across three representative forms of task expansion. In *variable CL*, the model sequentially acquires new atmospheric prediction targets. In *regional CL*, it extends a fixed set of prediction capabilities across geographical domains. In *climate-projection CL*, it acquires new climate-response targets. At each stage, only data from the current task are available, and the model is evaluated on all tasks encountered. By varying the atmospheric target, spatial domain, or climate-response target under a common sequential protocol, these settings capture complementary ways in which the capabilities of a foundation model may evolve. They allow us to address a basic but underexplored question: *when is direct sequential adaptation sufficient, and when is a dedicated CL mechanism required?*

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.

Our empirical investigation shows that forgetting depends strongly on how the task evolves. Under compatible task shifts, sequential adaptation causes little degradation and remains a simple, competitive solution. When incoming tasks change the prediction target and require distinct representations, however, the same strategy substantially overwrites acquired capabilities. Historical tasks can also support new ones, but such transfer is uneven and directional. Related variables, regions, or climate responses may share useful predictive structure, whereas indiscriminate reuse can introduce interference. These findings distinguish settings in which direct adaptation is sufficient from those requiring explicit knowledge protection, and establish two complementary objectives for the challenging regimes: stable retention and selective transfer.

Motivated by these findings, we propose Graph-guided Update Selection and Transfer (**GUST**), a parameter-efficient framework that addresses both objectives. Each task is represented by a lightweight low-rank adapter that is frozen after acquisition. GUST organizes the stored adapters into a causal directed task graph, where each edge models the potential contribution of a historical adapter to the current task. Edge weights integrate structured task relations, learned task geometry, and online estimates of transfer utility and interference. For each incoming task, GUST learns a new adapter and selectively composes historical updates, promoting beneficial sources while suppressing irrelevant or conflicting ones. By decoupling stable knowledge storage from adaptive knowledge reuse, GUST supports continual capability expansion with controlled parameter growth and without replaying historical data.

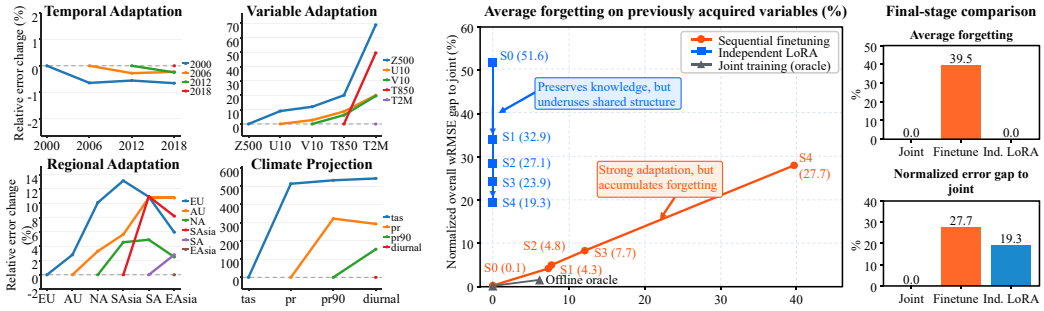
We evaluate GUST across all three settings on weather forecasting and climate projection benchmarks, comparing it with sequential finetuning, independent task adapters, offline joint training, and representative CL methods. GUST achieves the best overall balance between acquisition and retention among continual methods. In variable CL, it obtains an average relative wRMSE of 1.10 and reduces relative forgetting from 39.55% to 9.67%. In regional CL, it achieves a relative wRMSE of 1.04 and an average ACC of 0.55 with zero measured forgetting. In climate-projection CL, it achieves the lowest oracle-normalized error on three of four metrics. Ablation studies confirm that these gains extend beyond parameter isolation, while routing analyses reveal structured and directional knowledge transfer across tasks.

Our contributions are fourfold. (1) We formulate CL for weather and climate foundation models across atmospheric variables, geographical regions, and climate-response targets. (2) We reveal scenario-dependent forgetting and directional transfer, distinguishing when direct adaptation suffices from when explicit protection and reuse are needed. (3) We introduce GUST, which organizes frozen task adapters into a causal task graph and selectively composes historical updates using task relations and online feedback. (4) Extensive experiments show that GUST achieves the best overall balance between acquisition and retention among CL methods, substantially reducing forgetting while revealing interpretable cross-task transfer.

2 Related Work

Weather and climate foundation models. Data-driven weather and climate modeling has developed along two complementary directions. One focuses on specialized global forecasting systems based on spectral operators, three-dimensional architectures, graph representations, and probabilistic generative modeling (Kurth et al., 2023; Bi et al., 2023; Lam et al., 2023; Price et al., 2025). Collectively, these systems have improved the accuracy and efficiency of medium-range forecasting and extended deterministic prediction to probabilistic ensembles. The other develops general-purpose models pretrained across heterogeneous variables, resolutions, datasets, and prediction domains, enabling transfer across Earth-system applications (Nguyen et al., 2023; Schmude et al., 2024; Bodnar et al., 2025). Related efforts couple learned components with differentiable physical dynamics (Kochkov et al., 2024) and provide standardized benchmarks for climate emulation and projection (Watson-Parris et al., 2022; Kaltenborn et al., 2023). Despite their broader applicability, these models are typically evaluated through static pretraining followed by independent adaptation to individual tasks. Their sequential expansion across atmospheric variables, geographical regions, and climate-response targets remains underexplored.

Continual learning with foundation models. Continual learning mitigates catastrophic forgetting through several complementary strategies. Classical methods regularize parameters important to previous tasks, distill earlier model predictions, or replay representative samples (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Rebuffi et al., 2017). With foundation models, increasing attention has



(a) Forgetting behavior under different continual adaptation settings. (b) Stability-plasticity trade-off under variable continual learning.

Figure 1: **Empirical investigation of CL in weather and climate foundation models. (a)** Forgetting across temporal, regional, variable, and climate-projection adaptation, revealing strong scenario dependence. **(b)** Stability-plasticity trade-off in variable CL. Parameter isolation eliminates forgetting but leaves a substantial gap to joint training, motivating selective knowledge reuse.

shifted toward freezing the pretrained backbone and adapting lightweight task-dependent parameters. Prompt-based methods retrieve or compose learned prompts for incoming samples (Wang et al., 2022b;a; Smith et al., 2023), while adapter-based methods isolate task updates in expandable subspaces or low-rank components to reduce interference and control parameter growth (Zhou et al., 2024; Liang & Li, 2024; Wu et al., 2025). These approaches have been studied primarily in class-incremental vision and language settings, where tasks are defined by label-space or data-distribution shifts. Their behavior across distinct forms of scientific task expansion remains unclear, and existing methods provide limited mechanisms for exploiting structured and directional relationships among atmospheric variables, spatial domains, and climate responses. Our work addresses these gaps by jointly characterizing three forms of CL and guiding replay-free knowledge reuse through structured task relations, learned task geometry, and online transfer feedback.

3 Preliminaries

In this section, we first describe the problem formulation for continual adaptation of weather and climate foundation models, and then present an in-depth empirical investigation.

3.1 Problem Formulation

Weather and climate prediction. We consider foundation models that support prediction tasks with different targets and spatial domains. Let $\mathcal{T}_k$ denote the k-th task and $\mathcal{D}_k = \{(\mathcal{X}_i^{(k)}, \mathcal{Y}_i^{(k)})\}_{i=1}^{n_k}$ its training set, where $\mathcal{X}_i^{(k)}$ and $\mathcal{Y}_i^{(k)}$ are the input and target fields. Given $\mathcal{X}$, a model with parameters θ produces $\widehat{\mathcal{Y}} = f_\theta(\mathcal{X}; \mathcal{T}_k)$, where $\mathcal{T}_k$ specifies the prediction family, target, and spatial domain but need not be explicitly encoded as model input. We denote a weather forecasting task by $\mathcal{T}^{\mathrm{WF}}(\mathcal{V}, \Omega)$, where $\mathcal{V}$ is the target-variable set and Ω the spatial domain. Its input contains recent atmospheric states and its target contains future values of $\mathcal{V}$ over Ω. We denote a climate-projection task by $\mathcal{T}^{\mathrm{CP}}(r, \Omega)$, where the input contains external forcing trajectories and r is the resulting climate-response target. Other configurations, including input variables, forecast horizons, forcing channels, and spatial resolution, remain fixed within each scenario and are omitted.

Continual adaptation. We consider a sequence of tasks $\mathcal{T}_1, \ldots, \mathcal{T}_K$ that arrive over time. At stage k, the learner accesses only $\mathcal{D}_k$ and the state retained from previous stages, without replaying historical raw data. Full-parameter tuning repeatedly modifies the pretrained parameters θ_0, which is computationally expensive and can overwrite shared representations. We instead focus on parameter-efficient adaptation, which freezes θ_0 and introduces lightweight parameters $\Delta\theta_k$ for task $\mathcal{T}_k$, as specified in Sec. 4. After each stage, the model is evaluated on the current and all previously observed tasks, requiring it to acquire new capabilities, retain earlier ones, and exploit transferable knowledge. Independent adaptation and offline joint training serve as references for task isolation and unrestricted data access, respectively.

Continual-learning scenarios. We instantiate three task sequences by varying one component of weather or climate prediction. In *variable CL*, $\mathcal{T}_k = \mathcal{T}^{\mathrm{WF}}(\{v_k\}, \Omega)$, so the model acquires atmospheric variables $v_1, \ldots, v_K$ over a fixed domain; these variables are physically coupled but may

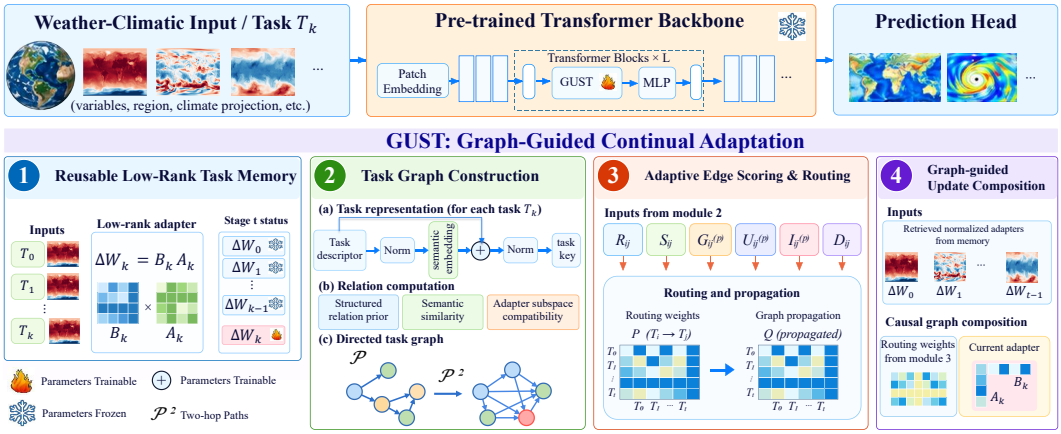


Figure 2: **Overview of GUST.** GUST stores acquired tasks as frozen low-rank adapters, organizes them into a directed task graph, and scores transfer edges using structured relations, learned task geometry, and online feedback. Graph propagation then composes selected historical adapters with the current update, which is applied to the frozen backbone without replaying historical data.

require distinct representations. In *regional CL*, $\mathcal{T}_k = \mathcal{T}^{\mathrm{WF}}(\mathcal{V}, \Omega_k)$, so a fixed variable set is predicted across geographical domains $\Omega_1, \ldots, \Omega_K$ that differ in regional climatology and data distribution. In *climate-projection CL*, $\mathcal{T}_k = \mathcal{T}^{\mathrm{CP}}(r_k, \Omega)$, so the model acquires climate-response targets $r_1, \ldots, r_K$ under fixed forcings and spatial domain; these targets share forcing mechanisms but exhibit different responses. The three scenarios therefore capture continual changes in atmospheric targets, geographical domains, and climate-response mappings under a sequential protocol.

3.2 Empirical Investigation

Before introducing our method, we empirically examine how weather and climate foundation models behave under different forms of sequential adaptation. We focus on two questions: how the form of task evolution affects forgetting, and whether retaining previous capabilities is sufficient for effective continual adaptation.

Forgetting depends on the form of task evolution. Using the relative forgetting metric defined in Sec. 5.1, we investigate how task evolution affects knowledge retention. Besides the three scenarios in Sec. 3.1, we include temporal adaptation as a diagnostic reference, in which the prediction target and spatial domain remain fixed while successive training sets are ordered chronologically. Fig. 1(a) reveals markedly different forgetting behavior across scenarios. Temporal adaptation causes little degradation and sometimes improves earlier tasks, while regional adaptation produces moderate interference. Changing prediction targets is substantially more disruptive: the relative error on early tasks reaches approximately 70% in variable CL and exceeds 500% in climate-projection CL. The severity of forgetting is therefore not determined by sequential optimization alone. It also depends on the structural compatibility of successive tasks and the predictive representations they require.

Retention alone is insufficient for effective adaptation. We next use variable CL to examine whether eliminating interference is sufficient. Fig. 1(b) compares sequential finetuning, independent LoRA adapters, and offline joint training. Sequential finetuning readily adapts to incoming variables through shared parameters but progressively overwrites earlier capabilities. At the final stage, it incurs 39.5% average forgetting and a 27.7% normalized wRMSE gap to joint training. Independent adapters eliminate measured forgetting by isolating task-specific updates, but still exhibit a 19.3% gap to the joint-training oracle. This contrast separates two complementary objectives: parameter isolation preserves acquired capabilities, while knowledge sharing is needed to exploit predictive structure across tasks. Effective continual adaptation must therefore protect historical knowledge while selectively transferring useful information to incoming tasks.

4 Method

The preceding analysis identifies two complementary requirements for continual adaptation: preserving acquired capabilities and transferring useful knowledge across tasks. Shared updates enable

transfer but risk overwriting previous capabilities, while independent adapters prevent interference but leave cross-task structure underexploited. We address this tension with Graph-guided Update Selection and Transfer (**GUST**), which separates stable knowledge storage from adaptive knowledge reuse. GUST stores each acquired task as a frozen low-rank adapter and organizes the stored adapters into a causal directed graph. For each incoming task, the graph selects and composes historical updates using structured task relations, learned task geometry, and online estimates of transfer utility and interference. Graph propagation further enables direct and multi-hop transfer without replaying historical data. Following the formulation in Sec. 3.1, we use variable CL as the running example. The same storage, routing, and composition procedures apply to regional and climate-projection CL, with scenario-specific task descriptors and relation priors.

4.1 Reusable Low-Rank Task Memory

To preserve task-specific knowledge without modifying the pretrained representation, GUST stores each acquired task as a lightweight low-rank adapter. Let $\mathcal{P}$ denote the set of adapted linear projections, and let $W_0^{(p)} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be the frozen pretrained weight of projection $p \in \mathcal{P}$. GUST instantiates the lightweight parameters $\Delta\theta_k$ introduced in Sec. 3.1 using

$$\Delta W_k^{(p)} = B_k^{(p)} A_k^{(p)}, \tag{1}$$

where $B_k^{(p)} \in \mathbb{R}^{d_{\text{out}} \times r_{\text{L}}}$ and $A_k^{(p)} \in \mathbb{R}^{r_{\text{L}} \times d_{\text{in}}}$ are trainable factors, with $r_{\text{L}} \ll \min(d_{\text{in}}, d_{\text{out}})$. The complete adapter for task $\mathcal{T}_k$ is $\Delta\theta_k = \{\Delta W_k^{(p)}\}_{p \in \mathcal{P}}$.

At stage k, the pretrained backbone and all historical adapters remain frozen, while $\Delta\theta_k$ is optimized on $\mathcal{D}_k$ together with the routing components in Sec. 4.3. Once task $\mathcal{T}_k$ is acquired, its adapter is frozen and added to the task memory, preventing subsequent stages from overwriting the stored update. Independent adaptation uses each adapter only for its originating task. GUST instead retains historical adapters as candidate knowledge sources and selectively combines them through the task graph in Sec. 4.2. The memory therefore grows by one low-rank adapter per task and stores no historical examples.

4.2 Structured and Adaptive Task Graph

GUST organizes the stored adapters into a causal directed graph. At stage k, each observed task $\mathcal{T}_i$, where $i \leq k$, forms a node associated with its adapter $\Delta\theta_i$ and task key $\mathbf{z}_i$. An edge $j \to i$, where $j < i$, represents potential transfer from source task $\mathcal{T}_j$ to recipient task $\mathcal{T}_i$. Because semantic similarity does not guarantee positive transfer, GUST characterizes each candidate edge using both structured domain knowledge and adaptive task geometry.

Structured task relation. Weather and climate tasks exhibit known dependencies across atmospheric variables, geographical regions, and climate responses. We encode these dependencies using a directed prior $R_{ij} \in [0, 1]$, where a larger value indicates greater prior relevance of task $\mathcal{T}_j$ to task $\mathcal{T}_i$. For variable CL, R_{ij} combines physical coupling and statistical association between variables. For regional CL, it reflects geographical and climatological similarity. For climate-projection CL, it captures relationships among response targets under shared forcings. These relations are computed when tasks are observed and retained as lightweight metadata, requiring no replay of historical data. Their scenario-specific construction is detailed in Appendix A.

Adaptive task geometry. Structured relations provide an interpretable prior but may not capture all transferable structure represented by the pretrained model. We therefore embed each task in a shared key space:

$$\mathbf{z}_i = \text{Norm}(e(\mathcal{T}_i) + \mathbf{d}_i), \tag{2}$$

where $e(\mathcal{T}_i)$ is a fixed embedding of the task descriptor, $\mathbf{d}_i$ is a learnable residual initialized to zero, and $\text{Norm}(\cdot)$ denotes ℓ_2 normalization. Among the task keys, only the residual $\mathbf{d}_i$ of the current task is optimized at stage i; the resulting key is frozen after acquisition. The geometric compatibility between recipient $\mathcal{T}_i$ and source $\mathcal{T}_j$ is $S_{ij} = \mathbf{z}_i^\top \mathbf{z}_j$. The structured prior R_{ij} and learned compatibility S_{ij} provide complementary evidence of potential transfer. The routing mechanism in Sec. 4.3 further refines this evidence using online optimization feedback.

4.3 Adaptive Routing and Graph Composition

The structured relation and task geometry defined in Sec. 4.2 estimate potential transfer, but cannot determine how a historical adapter affects optimization of the current task. GUST therefore augments them with online estimates of parameter interference and transfer utility.

Transfer-aware routing. At stage k, we quantify interference between the current adapter and historical adapter $\Delta\theta_j$ by averaging their directional conflict across adapted projections:

$$I_{kj} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \Big[-\cos_F\Big(\Delta W_k^{(p)}, \Delta W_j^{(p)}\Big) \Big]_+, \tag{3}$$

where $[x]_+ = \max(x, 0)$ and $\cos_F(A, B) = \langle A, B \rangle_F / (\|A\|_F \|B\|_F + \epsilon)$ is the stabilized cosine similarity under the Frobenius inner product, with $\epsilon > 0$. A larger I_{kj} indicates stronger conflict between update directions, and we set $I_{kk} = 0$.

For each candidate source $j \le k$, GUST combines the structured relation in Appendix A, geometric compatibility, online utility U_{kj}, and interference in Eq. equation 3 into

$$s_{kj} = \alpha_R R_{kj} + \alpha_S S_{kj} + \alpha_U U_{kj} - \alpha_I I_{kj}, \tag{4}$$

where $\alpha_R, \alpha_S, \alpha_U, \alpha_I \ge 0$ are learned coefficients. The current adapter is included as a self-source with $R_{kk} = S_{kk} = 1$. The routing weights are obtained by normalizing the scores in Eq. equation 4:

$$P_{kj} = \frac{\exp(s_{kj}/\tau)}{\sum_{m=1}^{k} \exp(s_{km}/\tau)}, \tag{5}$$

where $\tau > 0$ controls routing sharpness.

The utility U_{kj} estimates whether increasing the contribution of adapter j reduces the current prediction loss. It is initialized to zero and updated after each optimization step as

$$\widehat{U}_{kj} = -\frac{\partial \mathcal{L}_{\text{task}}^{(k)}}{\partial P_{kj}}, \qquad U_{kj} \leftarrow \gamma U_{kj} + (1 - \gamma)\widehat{U}_{kj}, \tag{6}$$

where $\gamma \in [0, 1)$ is the smoothing coefficient. The updated utility is treated as a fixed statistic when computing Eq. equation 4 at the next step. Positive utility promotes sources that improve the current objective, while the interference term suppresses conflicting updates.

Causal graph composition. After task $\mathcal{T}_i$ is acquired, its routing row $\{P_{ij}\}_{j=1}^{i}$ is frozen. Stacking the routing rows through stage k gives a lower-triangular row-stochastic matrix $\mathbf{P}_k \in \mathbb{R}^{k \times k}$, with $P_{ij} = 0$ for $j > i$. This causal structure prevents access to future tasks. To capture transfer through intermediate tasks, GUST propagates the routing weights as $\mathbf{Q}_k = (1-\rho)\mathbf{P}_k + \rho\mathbf{P}_k^2$, where $\rho \in [0, 1]$ controls the contribution of two-hop paths. For each adapted projection p, the deployed update after stage k is

$$\widetilde{\Delta W}_k^{(p)} = \sum_{j=1}^{k} Q_{kj} \Delta W_j^{(p)}, \tag{7}$$

where Q_{kj} is the (k, j)-th entry of $\mathbf{Q}_k$. The resulting model applies $W_0^{(p)} + \widetilde{\Delta W}_k^{(p)}$ at each adapted projection and is evaluated on all tasks observed through stage k. Graph propagation acts only on routing weights and never modifies the stored adapters.

Continual optimization. At stage k, GUST optimizes the current adapter $\Delta\theta_k$, geometry residual $\mathbf{d}_k$, routing coefficients, and task-specific prediction head using only $\mathcal{D}_k$. The objective is

$$\mathcal{L}_k = \mathcal{L}_{\text{task}}^{(k)} + \lambda_{\text{geo}} \|\mathbf{d}_k\|_2^2, \tag{8}$$

where $\mathcal{L}_{\text{task}}^{(k)}$ is the prediction loss on $\mathcal{D}_k$ and $\lambda_{\text{geo}} \ge 0$ limits deviation from the initial task descriptor in Eq. equation 2. After optimization, the adapter, task key, and routing row are frozen and retained for subsequent stages. No historical examples are stored or replayed.

5 Experiments

In this section, we first describe the experimental setup and then analyze the experimental results.

Table 1: **Final-stage performance in variable CL.** Results are reported after learning all five variables. Each metric is normalized by the corresponding Joint Training result, and *Avg.* denotes the macro average across variables. *Forget* denotes the average relative degradation of previously learned variables, computed from the acquisition stage to the final stage according to Sec. 5.1.

	Relative wRMSE ↓						Forget	Relative ACC ↑						Forget
Method	z500	u10m	v10m	t850	t2m	Avg.	wRMSE	z500	u10m	v10m	t850	t2m	Avg.	ACC
Joint Training	1.00	1.00	1.00	1.00	1.00	1.00	0.00%	1.00	1.00	1.00	1.00	1.00	1.00	0.00%
Finetuning	1.69	1.20	1.20	1.44	0.86	1.28	39.55%	0.94	0.92	0.92	0.95	1.01	0.95	7.27%
EWC	1.65	1.18	1.18	1.42	0.87	1.26	37.07%	0.95	0.92	0.93	0.95	1.01	0.95	6.71%
SimpleCIL	2.66	1.67	1.71	1.90	1.84	1.95	**0.00%**	0.79	0.55	0.46	0.82	0.86	0.70	**0.00%**
L2P	2.89	1.53	1.44	1.54	1.69	1.82	10.74%	0.85	0.74	0.77	0.92	0.90	0.83	3.04%
DualPrompt	2.78	1.48	1.46	1.58	1.44	1.75	10.90%	0.86	0.74	0.75	0.92	0.94	0.84	2.68%
CODA-Prompt	2.59	1.24	1.24	1.39	1.09	1.31	29.80%	0.88	0.79	0.80	0.94	0.94	0.87	4.64%
CPrompt	2.83	1.57	1.53	1.63	1.51	1.81	36.17%	0.80	0.68	0.69	0.90	0.94	0.80	1.58%
SdLoRA	1.96	1.32	1.29	1.47	0.94	1.39	44.72%	0.93	0.86	0.88	0.94	1.01	0.92	8.85%
InfoLoRA	1.95	1.27	1.24	1.29	1.01	1.35	37.92%	0.96	0.88	0.91	0.97	1.00	0.94	6.35%
Ours	**1.30**	**1.10**	**1.09**	**1.07**	**0.96**	**1.10**	**9.67%**	**0.98**	**0.97**	**0.96**	**1.00**	**1.00**	**0.98**	**1.42%**

5.1 Experimental Setup

Benchmarks. Following the scenarios defined in Sec. 3.1, we evaluate three CL benchmarks in which tasks arrive sequentially, only the current-task data are available, and performance is evaluated on all observed tasks without replay. The weather benchmarks use ERA5 data from WeatherBench (Rasp et al., 2024), with ClimaX (Nguyen et al., 2023) as the pretrained backbone. Variable CL follows $Z_{500} \rightarrow U_{10} \rightarrow V_{10} \rightarrow T_{850} \rightarrow T_{2m}$, denoting 500-hPa geopotential height, 10-m zonal and meridional winds, 850-hPa temperature, and 2-m air temperature, respectively. Regional CL predicts the same variable set across $\mathrm{Europe} \rightarrow \mathrm{Australia} \rightarrow$ North America $\rightarrow$ South Asia $\rightarrow$ South America $\rightarrow$ East Asia. Climate-projection CL uses ClimateBench (Watson-Parris et al., 2022), with CO_2, SO_2, CH_4, and black carbon as forcings and $\mathrm{tas} \rightarrow \mathrm{pr} \rightarrow \mathrm{pr90} \rightarrow \mathrm{diurnal}$ as response targets, corresponding to near-surface air temperature, precipitation, 90th-percentile precipitation, and diurnal temperature range.

Baselines. We compare GUST with sequential finetuning, independent task adapters, EWC (Kirkpatrick et al., 2017), prompt-based methods including L2P, DualPrompt, and CODA-Prompt (Wang et al., 2022b;a; Smith et al., 2023), and pretrained-model-based methods including SimpleCIL (Zhou et al., 2025), CPrompt (Gao et al., 2024), EASE (Zhou et al., 2024), InfLoRA (Liang & Li, 2024), and SD-LoRA (Wu et al., 2025). All methods follow the same task sequence without accessing historical raw data. Joint Training simultaneously accesses all task and serves as an offline oracle.

Evaluation metrics. For weather forecasting, we report latitude-weighted RMSE (wRMSE), anomaly correlation coefficient (ACC), and average forgetting on previously learned tasks. For a task $\mathcal{T}_i$ acquired at stage i, with prediction error $E_i^{(t)}$ after adaptation through stage t, we define its relative forgetting as $\Delta_{i \rightarrow t} = (E_i^{(t)} - E_i^{(i)})/E_i^{(i)} \times 100\%$, where positive values indicate degradation and negative values indicate improvement. The reported forgetting is averaged over all previous tasks at the final adaptation stage. We additionally normalize each task's wRMSE and ACC by its Joint Training result and report their macro averages across tasks. For climate projection, we follow ClimateBench and report spatial NRMSE (NRMSE_s), global NRMSE (NRMSE_g), Total NRMSE, and RMSE, together with their oracle-normalized ratios averaged across response targets.

Implementation details. All methods use the same backbone, data splits, training budget, and evaluation protocol. Models are optimized with AdamW for 50 epochs per stage using a learning rate of 5e-7 and batch size 24. GUST uses LoRA rank $r_{\mathrm{L}} = 10$, routing temperature $\tau = 1.0$, propagation coefficient $\rho = 0.5$, utility smoothing coefficient $\gamma = 0.95$, and geometry regularization coefficient $\lambda_{\mathrm{geo}} = 0.1$. Results are averaged over 3 runs and reported as mean only.

5.2 Experimental Results

Variable CL. Table 1 reports the final performance after sequentially learning $Z_{500} \rightarrow U_{10} \rightarrow V_{10} \rightarrow T_{850} \rightarrow T_{2m}$. Sequential finetuning achieves strong adaptation on the final T_{2m} task, obtaining a relative wRMSE of 0.86, but suffers from substantial degradation on previously learned variables, resulting in an average relative wRMSE of 1.28 and an average relative ACC of 0.95.

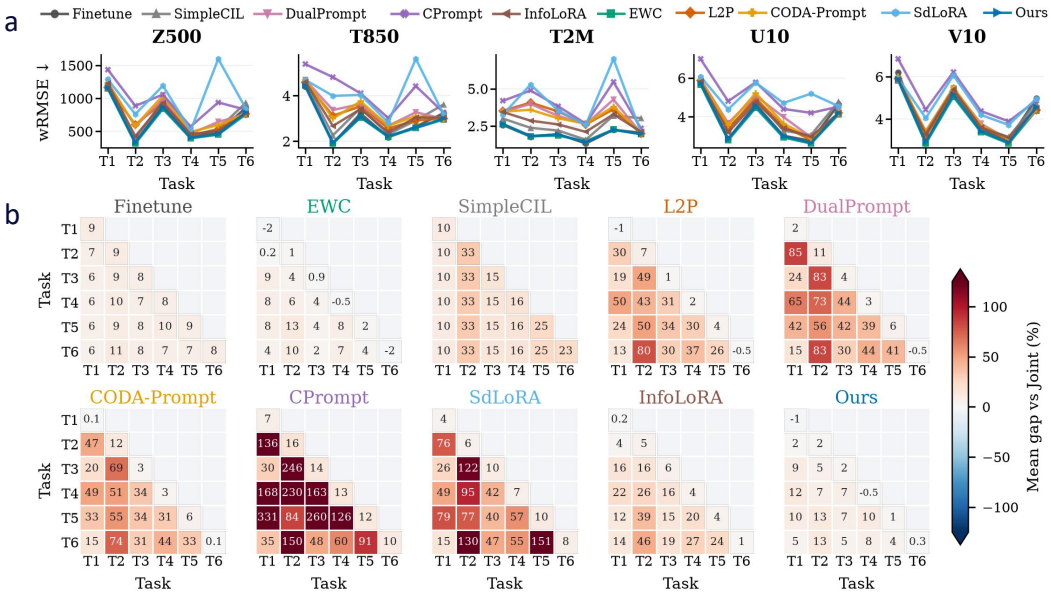


Figure 3: **Performance in regional CL. (a)** Final-stage forecasting performance across six regions for five meteorological variables. T_1–T_6 follow the task order. **(b)** Average relative wRMSE gap to Joint Training across variables at each adaptation stage and evaluation region. Each panel represents one method, and diagonal entries mark initial task acquisition. Positive and negative values indicate higher and lower error than Joint Training, respectively.

GUST achieves the best overall balance between retention and adaptation, with an average relative wRMSE of 1.10 and relative ACC of 0.98, while remaining competitive on the newly learned T_{2m} variable. Compared with the strongest baseline in average wRMSE, GUST reduces the relative wRMSE from 1.26 to 1.10 and improves average relative ACC from 0.95 to 0.98. Furthermore, GUST substantially alleviates forgetting: the relative wRMSE degradation of previously learned variables is reduced to 9.67%, compared with 37.07% for EWC and 39.55% for finetuning. The ACC forgetting is also limited to 1.42%, indicating that the proposed graph-guided update selection effectively preserves previously acquired variable representations while enabling CL.

Regional CL. Figure 3 shows that regional CL is comparatively less susceptible to forgetting, consistent with the empirical analysis in Sec. 3.2. Finetuning and EWC already remain relatively close to Joint Training, and GUST preserves this competitiveness while avoiding the large deviations exhibited by several prompt- and adapter-based baselines. Across the GUST heatmap in Fig. 3(b), the average relative wRMSE gap remains between -1% and 13% throughout adaptation; at the final stage, it averages approximately 5.9% across the six regions. By comparison, CPrompt and SDLoRA exceed a 100% gap in several stage-task pairs. These results show that GUST remains robust when direct adaptation is already effective, without introducing substantial interference through unnecessary knowledge reuse.

Climate Projection CL. Table 2 reports results after sequentially learning $\mathrm{tas} \rightarrow \mathrm{pr} \rightarrow \mathrm{pr90} \rightarrow \mathrm{diurnal}$. GUST achieves the strongest overall oracle-normalized performance, obtaining the lowest NRMSE_s, NRMSE_g, and RMSE ratios of 0.57, 7.23, and 0.17, respectively, while achieving a competitive Total NRMSE ratio of 15.31. These results indicate that GUST maintains more consistent performance across heterogeneous climate-response targets instead of optimizing only a subset of objectives. Compared with sequential finetuning and regularization-based approaches such as EWC, GUST avoids the substantial performance degradation caused by sequential updates, especially under large changes in prediction objectives. The forgetting analysis further shows that GUST introduces moderate degradation on previous targets, with relative forgetting rates of 48.13%, 174.93%, 75.77%, and 39.57% for NRMSE_s, NRMSE_g, Total, and RMSE, respectively. Although some baselines exhibit lower forgetting by limiting adaptation, their final-stage errors remain considerably higher, suggesting that GUST achieves a better balance between retaining previous climate-response knowledge and acquiring new predictive capabilities.

Ablation study. Table 3 reveals that historical adapter reuse is beneficial only when accompanied by mechanisms that control transfer. Naively enabling reuse increases the Z_{500} wRMSE from 368.77

Table 2: **Final-stage performance and forgetting in climate projection CL.** Results are reported after learning all four response targets. *Value* denotes the final metric, while *Forget* denotes the average relative degradation of previously learned targets from their acquisition stage to the final stage, following the forgetting definition in Sec. 5.1. Lower values indicate better performance.

	NRMSEs ↓		NRMSEg ↓		Total ↓		RMSE ↓	
Method	Value	Forget	Value	Forget	Value	Forget	Value	Forget
Joint Training	0.13	0.00%	4.48	0.00%	5.20	0.00%	8.31	0.00%
Finetuning	1.18	197.19%	9.84	418.34%	15.03	255.26%	0.15	119.50%
EWC	1.18	284.08%	9.81	822.30%	17.03	536.10%	0.15	158.08%
L2P	1.14	85.77%	8.14	122.69%	16.16	119.06%	0.23	71.49%
DualPrompt	1.17	31.27%	8.80	22.63%	17.25	27.45%	0.17	26.82%
CODA-Prompt	1.16	179.98%	9.31	175.03%	11.45	179.59%	0.15	109.95%
SimpleCIL	0.68	**0.00%**	6.05	**0.00%**	20.25	**0.00%**	0.21	**0.00%**
CPrompt	0.82	16.72%	8.41	-10.59%	23.44	0.35%	0.23	13.11%
SdLoRA	1.20	282.99%	10.36	502.21%	30.94	427.12%	0.18	160.03%
InfoLoRA	1.14	20.16%	8.59	26.28%	25.84	21.90%	0.28	14.78%
Ours	**0.57**	48.13%	**7.23**	174.93%	**15.31**	75.77%	**0.17**	39.57%

Table 3: **Ablation study in variable CL.** Performance is reported after learning all five variables. Lower wRMSE and higher ACC are better.

Mechanisms				z500		u10m		v10m		t850		t2m	
Reuse	Geometry	Routing	2-hop	wRMSE	ACC	wRMSE	ACC	wRMSE	ACC	wRMSE	ACC	wRMSE	ACC
×	×	×	×	368.77	0.93	2.50	**0.83**	2.59	**0.83**	1.87	**0.94**	1.64	0.94
✓	×	×	×	510.84	0.88	2.86	0.74	2.82	0.75	2.33	0.89	1.53	**0.96**
✓	✓	×	×	396.40	0.92	2.60	0.78	2.67	0.77	1.91	0.92	1.95	0.93
✓	✓	✓	×	328.79	**0.95**	2.46	0.81	2.50	0.80	1.80	0.93	1.71	0.95
✓	×	✓	✓	334.11	**0.95**	2.43	0.81	**2.43**	0.81	**1.66**	**0.94**	**1.46**	**0.96**
✓	✓	✓	✓	**315.64**	**0.95**	**2.41**	0.82	2.44	0.81	1.75	**0.94**	1.55	0.95

to 510.84 and decreases its ACC from 0.93 to 0.88, confirming that indiscriminate composition can introduce substantial interference. Adding task geometry partially recovers this degradation, reducing the Z_{500} wRMSE to 396.40, while transfer-aware routing further improves wRMSE across all five variables. Two-hop propagation provides another consistent gain over direct routing, reducing wRMSE from 328.79, 2.46, 2.50, 1.80, and 1.71 to 315.64, 2.41, 2.44, 1.75, and 1.55, respectively. Comparing the full model with its no-geometry variant further shows that task geometry primarily benefits the earliest variables, especially Z_{500}, although small trade-offs remain on some later variables. Overall, the results identify adaptive routing and graph propagation as the main mechanisms that convert stored adapters into useful transferable knowledge.

6 Conclusion

In this work, we studied continual learning for weather and climate foundation models across evolving atmospheric variables, geographical regions, and climate-response targets. Our empirical investigation shows that forgetting is governed by how the task evolves. Direct sequential adaptation can remain competitive under compatible task shifts, but changes in prediction targets can severely disrupt acquired capabilities. Moreover, retention alone is insufficient because cross-task transfer is uneven and directional. To address these challenges, we introduced a parameter-efficient and replay-free method that preserves task-specific updates in frozen low-rank adapters and selectively composes them through a causal task graph. Across all three settings, our method achieves the strongest overall balance between acquiring new capabilities and retaining previous ones, while revealing structured patterns of knowledge transfer. More broadly, our results suggest that weather and climate foundation models should be viewed as evolving systems instead of static checkpoints. Future work can extend CL to changing temporal regimes, resolutions, forecast horizons, and observing systems, while scaling to longer task sequences through adapter consolidation and bounded memory growth. Incorporating stronger physical and causal knowledge into transfer modeling may further improve robustness and interpretability. These directions could enable Earth-system foundation models to accumulate knowledge reliably throughout long-term deployment.

ETHICS STATEMENT

We acknowledge that all co-authors of this work have read and commit toadhering to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have included the source code with clear instructions, and will release them upon acceptance.

AI USE STATEMENT

AI-assisted tools were used only for improving writing clarity, grammar checking, and minor editing during manuscript preparation. The authors are solely responsible for the content, including the research methodology, experimental results, and conclusions.

REFERENCES

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.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Shengchao Chen, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. Foundation models for weather and climate data understanding: A comprehensive survey. *arXiv preprint arXiv:2312.03014*, 2023.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28463–28473. IEEE, 2024.

Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems*, 36:21757–21792, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

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.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023.

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.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647. IEEE, 2024.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 25904–25938. PMLR, 2023.

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. Gencast: Diffusion-based ensemble forecasting for medium-range weather. In *105th Annual AMS Meeting 2025*, volume 105, pp. 449275, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542, 2017.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, et al. Prithvi wxc: Foundation model for weather and climate. *arXiv preprint arXiv:2409.13598*, 2024.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149, 2022b.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer J Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 15: e2021MS002954, 2022.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. In *International Conference on Learning Representations*, volume 2025, pp. 58337–58353, 2025.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23554–23564. IEEE, 2024.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

A Task Relation and Descriptor Construction

GUST uses the same memory, routing, and composition mechanisms across all three CL scenarios. The scenario-specific components are the relation prior R_{ij} and the fixed descriptor $e(\mathcal{T}_i)$ used to initialize the task key in Eq. 2. For a source task $\mathcal{T}_j$ and recipient task $\mathcal{T}_i$, $R_{ij} \in [0, 1]$ measures their prior relevance before online transfer feedback is observed. All task statistics are computed when the corresponding task is encountered and retained only as lightweight metadata, without storing historical examples.

For two descriptor vectors $\mathbf{a}$ and $\mathbf{b}$, we define

$$\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\left(1 + \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2 + \epsilon}\right), \tag{9}$$

where $\epsilon > 0$ ensures numerical stability. Continuous descriptor components are rescaled using fixed, task-independent normalization before computing this similarity.

Variable CL. For $\mathcal{T}_i = \mathcal{T}^{\mathrm{WF}}(\{v_i\}, \Omega)$, the relation prior combines physical coupling with statistical association. The physical relation from source variable v_j to target variable v_i is

$$R_{ij}^{\mathrm{phys}} = c_{\mathrm{type}}(v_j, v_i) c_{\mathrm{level}}(v_j, v_i), \tag{10}$$

where $c_{\mathrm{type}} \in [0, 1]$ is a fixed directed relation between variable categories and $c_{\mathrm{level}} \in [0, 1]$ reflects compatibility between their vertical levels. This term serves as a coarse physical prior without explicitly modeling atmospheric dynamics. We complement it with an input-to-target association computed from the current dataset:

$$R_{ij}^{\mathrm{stat}} = \left|\mathrm{Corr}_{\mathcal{D}_i}\left(\langle \mathcal{X}_{v_j} \rangle_\Omega, \langle \mathcal{Y}_{v_i} \rangle_\Omega\right)\right|, \tag{11}$$

where $\langle \cdot \rangle_\Omega$ denotes spatial averaging. The resulting prior is

$$R_{ij} = \lambda_{\mathrm{phys}} R_{ij}^{\mathrm{phys}} + (1 - \lambda_{\mathrm{phys}}) R_{ij}^{\mathrm{stat}}, \tag{12}$$

where $\lambda_{\mathrm{phys}} \in [0, 1]$. If v_j is unavailable as an input channel for $\mathcal{T}_i$, only the physical relation is used. The fixed task descriptor $e(\mathcal{T}_i)$ is the frozen pretrained embedding of variable v_i.

Regional CL. For $\mathcal{T}_i = \mathcal{T}^{\mathrm{WF}}(\mathcal{V}, \Omega_i)$, each region is represented by a geographical descriptor $\mathbf{g}_i$ and a climatological descriptor $\mathbf{c}_i$. The former encodes the center and spatial extent of Ω_i, while the latter contains the means and standard deviations of the variables in $\mathcal{V}$ computed from $\mathcal{D}_i$. We define

$$R_{ij} = \lambda_{\mathrm{geo}} \,\mathrm{sim}(\mathbf{g}_i, \mathbf{g}_j) + (1 - \lambda_{\mathrm{geo}}) \,\mathrm{sim}(\mathbf{c}_i, \mathbf{c}_j), \tag{13}$$

where $\lambda_{\mathrm{geo}} \in [0, 1]$ balances geographical and climatological similarity. The fixed task descriptor is their concatenation, $e(\mathcal{T}_i) = [\mathbf{g}_i; \mathbf{c}_i]$.

Climate-projection CL. For $\mathcal{T}_i = \mathcal{T}^{\mathrm{CP}}(r_i, \Omega)$, tasks share the same forcing channels but differ in their climate-response targets. We characterize each target by a forcing-response signature

$$\mathbf{a}_i = \left[\mathrm{Corr}_{\mathcal{D}_i}(\langle \mathcal{X}_c \rangle_\Omega, \langle \mathcal{Y}_{r_i} \rangle_\Omega)\right]_{c \in \mathcal{C}}, \tag{14}$$

where $\mathcal{C}$ denotes the fixed set of forcing channels. The relation prior and task descriptor are

$$R_{ij} = \mathrm{sim}(\mathbf{a}_i, \mathbf{a}_j), \qquad e(\mathcal{T}_i) = [\mathbf{e}_{r_i}; \mathbf{a}_i], \tag{15}$$

where $\mathbf{e}_{r_i}$ is a fixed identity embedding of response target r_i. The forcing-response signature captures similarity in how different targets respond to the shared forcing trajectories.

For all scenarios, we set $R_{ii} = 1$ and $R_{ij} = 0$ for $j > i$. Although the regional and climate similarities can be symmetric, they are instantiated only on causal edges $j \to i$. Online utility and interference subsequently make the learned routing directional. Once $\mathcal{T}_i$ is acquired, its relation metadata and descriptor are frozen together with its adapter. The residual $\mathbf{d}_i$, task key $\mathbf{z}_i$, routing weights, and graph propagation follow Secs. 4.2 and 4.3.

B Continual Optimization Algorithm

Algorithm 1 summarizes the continual adaptation procedure.

Algorithm 1 Graph-Guided Continual Adaptation

Require: Pretrained backbone f_{θ_0}, task stream $\{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$

for $t = 1, \ldots, N$ **do**
  Initialize new adapter ΔW_t
  Construct task node $(\mathcal{T}_t, k_t, \Delta W_t)$
  Update task relation graph R
  Compute routing scores s_{ij}
  Obtain routing matrix M
  Optimize current adapter, geometry residual, and routing parameters
  Freeze ΔW_t
end for

C Detailed Results for Variable Continual Learning

Stages S_0–S_4 successively introduce Z500, U10, V10, T850, and T2M. Figure 4 reports each variable from its acquisition stage onward. The five tables give all supplied absolute wRMSE and ACC scores, preserving two decimal places. The constant joint-training scores are offline references, not a sequential training trajectory. "Ours" is labeled GUST throughout.

Performance across stages. GUST's advantage on older variables is most evident after the final update: its final wRMSE is lower than all other reported continual methods on Z500, U10, V10, and T850. Finetuning and EWC achieve lower wRMSE on the newly introduced T2M. These results exhibit a trade-off between retention and new-task prediction. Since initial scores differ across methods, absolute trajectories should not be interpreted as a comparison of forgetting alone. ACC ties are based on the supplied two-decimal precision; repeated-run uncertainty is not available in these results.

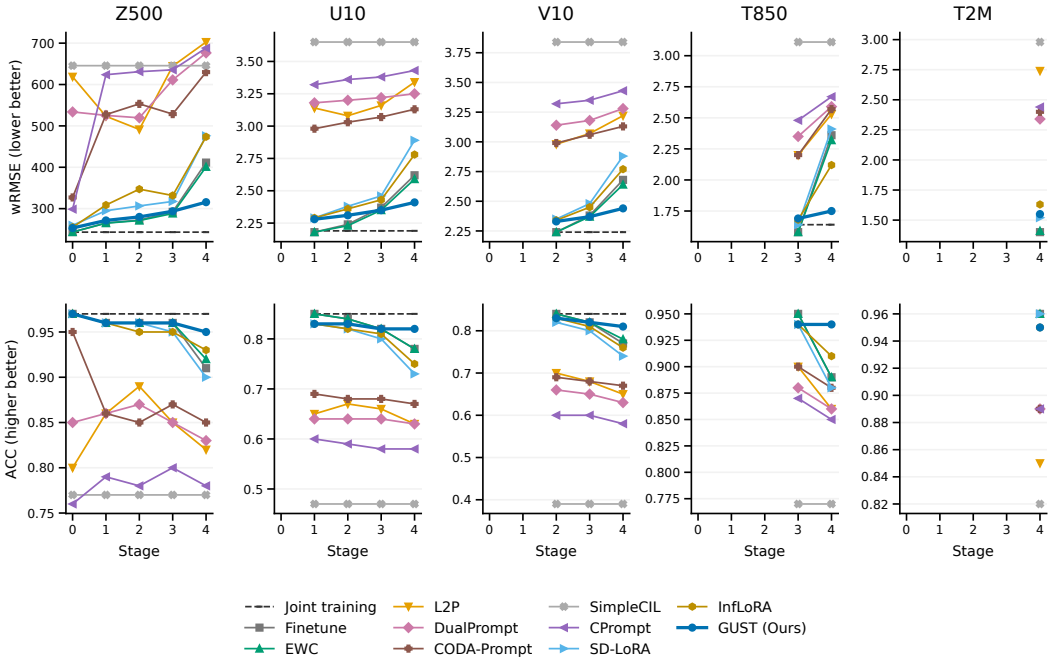


Figure 4: Variable-wise performance across the five learning stages. Top: wRMSE (lower is better); bottom: ACC (higher is better). Each column uses its own vertical scale. Curves begin at acquisition; T2M has only one observed stage. Dashed joint-training references are shown only over observed stages.

Table 4: Z500 performance after each observed stage. Bold marks the best continual result, including ties at the supplied precision; joint training is a separate reference.

	wRMSE ↓					ACC ↑				
Method	S_0	S_1	S_2	S_3	S_4	S_0	S_1	S_2	S_3	S_4
Joint training	243.19	243.19	243.19	243.19	243.19	0.97	0.97	0.97	0.97	0.97
Finetune	243.38	265.34	272.33	292.24	411.31	**0.97**	**0.96**	**0.96**	**0.96**	0.91
EWC	**243.34**	**264.95**	**271.22**	**288.25**	400.96	**0.97**	**0.96**	**0.96**	**0.96**	0.92
L2P	618.78	523.30	491.41	644.59	702.66	0.80	0.86	0.89	0.85	0.82
DualPrompt	533.78	525.37	519.54	611.14	676.39	0.85	0.86	0.87	0.85	0.83
CODA-Prompt	327.00	527.32	553.22	528.52	629.97	0.95	0.86	0.85	0.87	0.85
SimpleCIL	645.57	645.57	645.57	645.57	645.57	0.77	0.77	0.77	0.77	0.77
CPrompt	298.66	623.70	631.03	635.48	688.13	0.76	0.79	0.78	0.80	0.78
SD-LoRA	259.83	294.90	306.52	317.41	475.99	**0.97**	**0.96**	**0.96**	0.95	0.90
InfLoRA	255.05	308.66	347.14	331.72	473.63	**0.97**	**0.96**	0.95	0.95	0.93
GUST (Ours)	253.11	271.49	280.00	293.72	**315.64**	**0.97**	**0.96**	**0.96**	**0.96**	**0.95**

Table 5: U10 performance after each observed stage. Bold marks the best continual result, including ties at the supplied precision; joint training is a separate reference.

	wRMSE ↓				ACC ↑			
Method	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
Joint training	2.19	2.19	2.19	2.19	0.85	0.85	0.85	0.85
Finetune	**2.18**	2.24	2.37	2.62	**0.85**	**0.84**	**0.82**	0.78
EWC	**2.18**	**2.23**	**2.35**	2.59	**0.85**	**0.84**	**0.82**	0.78
L2P	3.14	3.08	3.16	3.34	0.65	0.67	0.66	0.63
DualPrompt	3.18	3.20	3.22	3.25	0.64	0.64	0.64	0.63
CODA-Prompt	2.98	3.03	3.07	3.13	0.69	0.68	0.68	0.67
SimpleCIL	3.65	3.65	3.65	3.65	0.47	0.47	0.47	0.47
CPrompt	3.32	3.36	3.38	3.43	0.60	0.59	0.58	0.58
SD-LoRA	2.29	2.38	2.46	2.89	0.83	0.82	0.80	0.73
InfLoRA	2.29	2.36	2.43	2.78	0.83	0.82	0.81	0.75
GUST (Ours)	2.28	2.31	**2.35**	**2.41**	0.83	0.83	**0.82**	**0.82**

Table 6: V10 performance after each observed stage. Bold marks the best continual result, including ties at the supplied precision; joint training is a separate reference.

	wRMSE ↓			ACC ↑		
Method	S_2	S_3	S_4	S_2	S_3	S_4
Joint training	2.24	2.24	2.24	0.84	0.84	0.84
Finetune	**2.24**	2.38	2.68	**0.84**	**0.82**	0.77
EWC	**2.24**	**2.37**	2.64	**0.84**	**0.82**	0.78
L2P	2.98	3.07	3.22	0.70	0.68	0.65
DualPrompt	3.14	3.18	3.28	0.66	0.65	0.63
CODA-Prompt	2.99	3.06	3.13	0.69	0.68	0.67
SimpleCIL	3.84	3.84	3.84	0.39	0.39	0.39
CPrompt	3.32	3.35	3.43	0.60	0.60	0.58
SD-LoRA	2.35	2.48	2.88	0.82	0.80	0.74
InfLoRA	2.34	2.45	2.77	0.83	0.81	0.76
GUST (Ours)	2.33	**2.37**	**2.44**	0.83	**0.82**	**0.81**

Table 7: T850 performance after each observed stage. Bold marks the best continual result, including ties at the supplied precision; joint training is a separate reference.

	wRMSE ↓		ACC ↑	
Method	S_3	S_4	S_3	S_4
Joint training	1.64	1.64	0.94	0.94
Finetune	**1.58**	2.36	**0.95**	0.89
EWC	**1.58**	2.32	**0.95**	0.89
L2P	2.20	2.53	0.90	0.86
DualPrompt	2.35	2.59	0.88	0.86
CODA-Prompt	2.20	2.57	0.90	0.88
SimpleCIL	3.11	3.11	0.77	0.77
CPrompt	2.48	2.67	0.87	0.85
SD-LoRA	1.64	2.41	0.94	0.88
InfLoRA	1.68	2.12	0.94	0.91
GUST (Ours)	1.69	**1.75**	0.94	**0.94**

Table 8: T2M performance after each observed stage. Bold marks the best continual result, including ties at the supplied precision; joint training is a separate reference.

	wRMSE ↓	ACC ↑
Method	S_4	S_4
Joint training	1.62	0.95
Finetune	**1.40**	**0.96**
EWC	1.41	**0.96**
L2P	2.74	0.85
DualPrompt	2.34	0.89
CODA-Prompt	2.40	0.89
SimpleCIL	2.98	0.82
CPrompt	2.44	0.89
SD-LoRA	1.52	**0.96**
InfLoRA	1.63	0.95
GUST (Ours)	1.55	0.95

D Detailed Results for Regional Continual Learning

The regional sequence in this experiment is Europe (EU), Australia (AU), North America (NA), South Asia (SA), South America (SAM), and East Asia (EA), corresponding to stages 1–6. All five variables are evaluated at a three-day forecast horizon. GUST is denoted by "Ours" in the source results. Scores are reported to the supplied precision of two decimal places.

Retention since acquisition. We compare each region with its own score immediately after acquisition. For method m, variable v, region r acquired at stage a_r, and evaluation stage $s \geq a_r$, we report

$$\Delta^{\mathrm{RMSE}}_{m,s,r} = \frac{100}{5} \sum_{v} \left(\frac{e_{m,s,r,v}}{e_{m,a_r,r,v}} - 1 \right), \tag{16}$$

$$\Delta^{\mathrm{ACC}}_{m,s,r} = \frac{100}{5} \sum_{v} \left(c_{m,a_r,r,v} - c_{m,s,r,v} \right). \tag{17}$$

Here e and c denote wRMSE and ACC. Positive values indicate deterioration; negative values indicate improvement. The wRMSE change is computed per variable before averaging, so quantities with different physical scales are not averaged directly. ACC changes are absolute differences expressed in percentage points, avoiding division by small or negative ACC values. These quantities describe retention relative to acquisition, not gaps to joint training or losses from the best historical score. Initial performance differs across methods; small changes alone do not establish strong prediction performance.

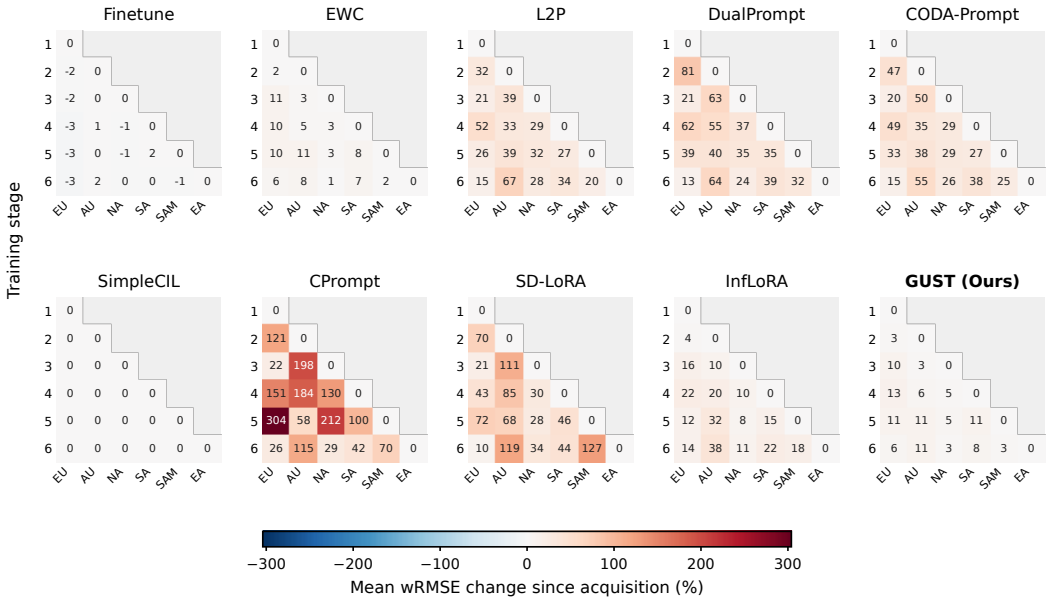


Figure 5: Regional wRMSE changes since acquisition, averaged over the five variables. Rows are training stages; columns are regions in acquisition order. All ten continual methods share one color scale. Gray cells denote regions not yet observed. Diagonal zeros follow from the definition; annotations are rounded to whole percentages.

ACC retention and absolute scores. Figure 6 provides the corresponding ACC changes. Table 9 lists the offline reference. The following method-specific tables report every available absolute score at every stage, including the acquisition stage; future-region scores are omitted rather than treated as zero. The source does not provide repeated-run uncertainty estimates.

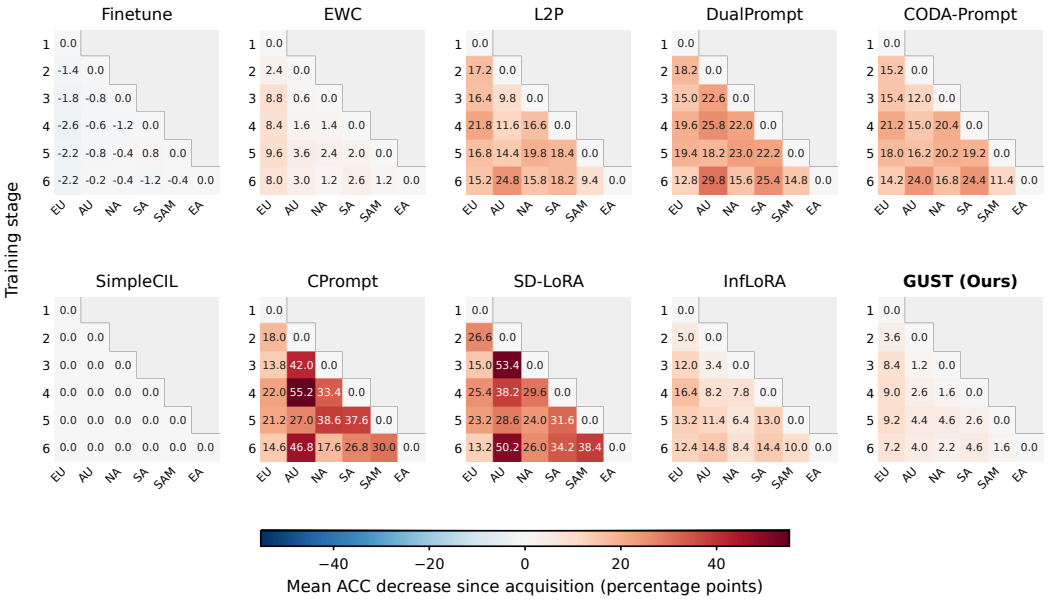


Figure 6: Mean regional ACC decrease since acquisition in percentage points. Positive values indicate lower ACC than at acquisition. All methods share one color scale; annotations use one decimal place. Gray cells indicate unseen regions.

Table 9: Offline joint-training reference for the regional experiment (three-day forecast horizon).

	wRMSE ↓					ACC ↑				
Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
EU	1092.49	5.51	5.82	4.24	2.45	0.33	0.30	0.13	0.28	0.53
AU	282.30	2.65	2.66	1.73	1.58	0.86	0.70	0.58	0.86	0.90
NA	849.10	4.37	5.12	2.99	1.69	0.57	0.50	0.36	0.55	0.68
SA	394.97	2.85	3.34	2.05	1.19	0.60	0.68	0.51	0.61	0.78
SAM	445.08	2.57	2.80	2.48	2.08	0.74	0.52	0.45	0.72	0.87
EA	791.85	4.21	4.64	2.99	1.89	0.54	0.51	0.31	0.56	0.80

Table 10: Complete regional results for Finetune at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1192.29	6.19	6.52	4.56	2.53	0.25	0.20	0.07	0.20	0.51
2	EU	1176.98	6.12	6.41	4.48	2.48	0.26	0.21	0.09	0.22	0.52
	AU	311.60	2.93	2.92	1.90	1.64	0.84	0.64	0.51	0.84	0.90
3	EU	1167.39	6.04	6.36	4.47	2.48	0.28	0.22	0.09	0.21	0.52
	AU	313.60	2.84	2.87	1.90	1.73	0.84	0.66	0.53	0.85	0.89
	NA	923.39	4.80	5.64	3.25	1.76	0.51	0.42	0.26	0.48	0.67
4	EU	1167.40	5.96	6.34	4.44	2.46	0.29	0.24	0.09	0.22	0.52
	AU	315.50	2.84	2.87	1.91	1.75	0.84	0.66	0.52	0.85	0.89
	NA	901.18	4.71	5.52	3.21	1.79	0.53	0.43	0.28	0.49	0.67
	SA	429.85	3.11	3.72	2.20	1.23	0.53	0.62	0.42	0.56	0.76
5	EU	1168.68	6.00	6.26	4.45	2.48	0.29	0.23	0.10	0.22	0.50
	AU	314.11	2.83	2.83	1.87	1.79	0.84	0.67	0.53	0.85	0.88
	NA	910.12	4.71	5.46	3.23	1.84	0.52	0.43	0.29	0.48	0.64
	SA	430.22	3.07	3.66	2.22	1.35	0.52	0.62	0.43	0.55	0.73
	SAM	487.07	2.82	3.09	2.67	2.21	0.68	0.42	0.34	0.68	0.85
6	EU	1166.94	5.88	6.20	4.40	2.57	0.30	0.24	0.09	0.22	0.49
	AU	329.89	2.86	2.85	1.90	1.79	0.84	0.66	0.52	0.84	0.88
	NA	892.74	4.63	5.35	3.22	1.96	0.53	0.44	0.30	0.48	0.61
	SA	424.39	3.02	3.54	2.17	1.33	0.55	0.64	0.46	0.57	0.73
	SAM	481.59	2.76	2.98	2.64	2.25	0.68	0.43	0.37	0.67	0.84
	EA	858.55	4.64	4.98	3.23	2.05	0.46	0.42	0.24	0.50	0.78

Table 11: Complete regional results for EWC at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1086.66	5.49	5.73	4.19	2.33	0.33	0.30	0.14	0.28	0.54
2	EU	1106.05	5.59	5.79	4.25	2.41	0.30	0.26	0.13	0.25	0.53
	AU	290.93	2.69	2.70	1.77	1.56	0.85	0.69	0.58	0.86	0.90
3	EU	1173.41	5.76	5.87	4.53	3.01	0.25	0.23	0.10	0.18	0.39
	AU	298.85	2.73	2.70	1.77	1.73	0.85	0.68	0.57	0.86	0.89
	NA	839.05	4.40	5.05	3.00	1.79	0.59	0.49	0.37	0.55	0.64
4	EU	1169.46	5.77	5.83	4.46	3.00	0.25	0.22	0.11	0.20	0.39
	AU	300.11	2.79	2.81	1.85	1.67	0.85	0.67	0.54	0.85	0.89
	NA	847.71	4.42	5.07	3.11	1.99	0.58	0.49	0.36	0.52	0.62
	SA	388.72	2.86	3.30	2.03	1.20	0.61	0.67	0.52	0.62	0.77
5	EU	1177.63	5.73	5.79	4.50	2.97	0.23	0.21	0.11	0.20	0.36
	AU	317.17	2.81	2.88	1.96	1.94	0.85	0.65	0.51	0.83	0.86
	NA	845.45	4.43	5.03	3.07	1.99	0.58	0.48	0.37	0.52	0.57
	SA	389.61	2.90	3.31	2.19	1.59	0.61	0.66	0.51	0.59	0.72
	SAM	449.88	2.58	2.81	2.51	2.19	0.72	0.51	0.44	0.70	0.85
6	EU	1148.10	5.66	5.82	4.46	2.62	0.24	0.25	0.09	0.19	0.42
	AU	315.12	2.79	2.84	1.90	1.82	0.84	0.66	0.52	0.84	0.87
	NA	840.25	4.46	5.05	3.04	1.86	0.58	0.48	0.36	0.53	0.63
	SA	394.07	2.93	3.37	2.19	1.48	0.60	0.65	0.49	0.58	0.74
	SAM	450.02	2.62	2.84	2.58	2.27	0.72	0.49	0.42	0.69	0.84
	EA	757.95	4.13	4.41	2.95	1.95	0.57	0.53	0.37	0.57	0.78

Table 12: Complete regional results for L2P at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1091.69	5.52	5.71	4.23	2.35	0.32	0.29	0.13	0.26	0.54
2	EU	1403.15	5.90	5.85	5.74	4.34	0.09	0.17	0.08	0.12	0.22
	AU	304.19	2.82	2.82	1.84	1.74	0.84	0.65	0.51	0.85	0.87
3	EU	1234.15	5.83	5.77	4.85	4.02	0.16	0.19	0.12	0.10	0.15
	AU	423.89	3.17	3.10	2.56	3.36	0.74	0.60	0.43	0.75	0.71
	NA	850.43	4.47	5.07	3.06	1.75	0.57	0.47	0.34	0.52	0.66
4	EU	1518.46	6.14	5.87	5.91	6.30	0.06	0.13	0.08	0.04	0.14
	AU	420.80	3.18	3.12	2.43	2.95	0.75	0.55	0.40	0.76	0.68
	NA	962.15	4.80	5.37	3.67	3.46	0.47	0.37	0.25	0.33	0.31
	SA	394.25	2.88	3.33	2.07	1.29	0.60	0.66	0.50	0.60	0.74
5	EU	1323.99	6.21	5.92	4.96	4.13	0.12	0.16	0.10	0.13	0.19
	AU	432.86	3.18	3.16	2.54	3.34	0.76	0.53	0.36	0.74	0.61
	NA	980.03	5.05	5.43	3.90	3.42	0.43	0.36	0.22	0.29	0.27
	SA	470.03	3.38	3.61	2.80	1.99	0.45	0.50	0.41	0.33	0.49
	SAM	459.18	2.66	2.87	2.56	2.26	0.70	0.45	0.39	0.69	0.84
6	EU	1165.71	5.86	5.92	4.59	3.52	0.22	0.16	0.10	0.14	0.16
	AU	602.79	3.37	3.18	3.12	4.07	0.60	0.48	0.35	0.60	0.45
	NA	934.53	4.94	5.32	3.53	3.48	0.51	0.33	0.25	0.39	0.29
	SA	438.27	3.50	3.64	2.56	2.61	0.50	0.49	0.37	0.42	0.41
	SAM	530.90	2.81	3.00	2.91	3.63	0.64	0.38	0.33	0.61	0.64
	EA	764.10	4.19	4.43	3.00	2.00	0.56	0.51	0.36	0.55	0.77

Table 13: Complete regional results for DualPrompt at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1125.22	5.62	5.74	4.33	2.52	0.27	0.25	0.11	0.22	0.46
2	EU	2176.85	6.35	5.98	8.14	7.78	0.03	0.08	0.04	0.06	0.19
	AU	311.34	2.90	2.90	1.88	1.88	0.84	0.63	0.47	0.84	0.85
3	EU	1270.86	5.92	5.86	5.08	4.27	0.12	0.16	0.06	0.07	0.15
	AU	614.21	3.16	3.07	3.35	4.24	0.55	0.56	0.40	0.56	0.43
	NA	864.41	4.54	5.09	3.13	1.88	0.56	0.45	0.33	0.50	0.60
4	EU	1908.13	6.04	5.93	6.50	7.03	0.02	0.11	0.04	0.03	0.13
	AU	559.80	3.36	3.26	3.12	3.74	0.54	0.47	0.33	0.59	0.41
	NA	1155.17	4.84	5.39	4.21	3.88	0.31	0.36	0.24	0.22	0.21
	SA	395.90	2.89	3.33	2.08	1.31	0.60	0.66	0.50	0.60	0.73
5	EU	1801.82	6.24	5.93	5.40	4.97	0.02	0.06	0.04	0.08	0.14
	AU	463.03	3.47	3.30	2.57	3.37	0.73	0.41	0.27	0.72	0.59
	NA	1106.93	5.04	5.46	3.96	3.84	0.30	0.29	0.20	0.30	0.20
	SA	489.73	3.58	3.64	2.65	2.47	0.41	0.42	0.38	0.39	0.38
	SAM	460.71	2.69	2.89	2.57	2.37	0.70	0.43	0.38	0.68	0.81
6	EU	1242.65	5.93	5.88	4.69	3.45	0.15	0.12	0.07	0.10	0.23
	AU	573.78	3.67	3.43	3.37	3.98	0.53	0.36	0.23	0.58	0.44
	NA	989.18	5.07	5.45	3.63	3.18	0.48	0.30	0.21	0.38	0.29
	SA	464.82	3.99	3.77	2.67	2.61	0.43	0.40	0.32	0.40	0.27
	SAM	642.01	2.94	3.04	3.27	4.28	0.56	0.32	0.30	0.56	0.52
	EA	760.80	4.20	4.43	2.98	2.01	0.57	0.51	0.35	0.55	0.76

Table 14: Complete regional results for CODA-Prompt at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1105.98	5.54	5.73	4.29	2.43	0.31	0.27	0.12	0.24	0.51
2	EU	1601.10	5.83	5.87	6.33	5.70	0.08	0.21	0.08	0.08	0.24
	AU	313.20	2.92	2.92	1.91	1.86	0.83	0.62	0.46	0.83	0.85
3	EU	1239.96	5.91	5.99	4.93	3.91	0.17	0.15	0.07	0.11	0.18
	AU	527.78	3.32	3.09	3.05	3.79	0.68	0.55	0.41	0.72	0.63
	NA	866.61	4.53	5.10	3.10	1.80	0.55	0.45	0.33	0.50	0.63
4	EU	1622.84	6.11	6.01	5.97	5.96	0.04	0.10	0.06	0.04	0.15
	AU	432.81	3.39	3.34	2.53	3.19	0.71	0.49	0.32	0.74	0.58
	NA	1065.73	4.90	5.39	3.77	3.38	0.34	0.35	0.24	0.28	0.23
	SA	397.36	2.91	3.35	2.10	1.31	0.59	0.66	0.49	0.59	0.73
5	EU	1564.44	6.23	6.20	5.67	4.17	0.06	0.09	0.06	0.12	0.22
	AU	447.58	3.38	3.29	2.65	3.35	0.74	0.44	0.28	0.73	0.59
	NA	997.76	5.00	5.57	3.92	3.35	0.38	0.32	0.21	0.30	0.24
	SA	475.19	3.39	3.60	2.62	2.16	0.42	0.49	0.40	0.36	0.43
	SAM	461.66	2.68	2.88	2.58	2.35	0.71	0.44	0.39	0.68	0.83
6	EU	1234.02	5.90	5.99	4.69	3.47	0.17	0.15	0.07	0.13	0.22
	AU	583.61	3.58	3.39	3.00	3.59	0.55	0.39	0.25	0.64	0.56
	NA	1040.92	5.22	5.49	3.75	3.02	0.46	0.28	0.20	0.35	0.33
	SA	492.47	3.65	3.70	2.68	2.66	0.36	0.47	0.34	0.37	0.30
	SAM	594.56	2.90	3.07	3.10	3.80	0.61	0.33	0.30	0.60	0.64
	EA	768.95	4.21	4.46	3.00	2.02	0.56	0.50	0.34	0.54	0.76

Table 15: Complete regional results for SimpleCIL at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1254.74	5.80	5.83	4.58	3.00	0.11	0.16	0.04	0.11	0.32
2	EU	1254.74	5.80	5.83	4.58	3.00	0.11	0.16	0.04	0.11	0.32
	AU	373.50	3.39	3.22	2.26	2.38	0.75	0.44	0.23	0.75	0.74
3	EU	1254.74	5.80	5.83	4.58	3.00	0.11	0.16	0.04	0.11	0.32
	AU	373.50	3.39	3.22	2.26	2.38	0.75	0.44	0.23	0.75	0.74
	NA	974.04	4.91	5.34	3.46	2.20	0.39	0.30	0.19	0.33	0.42
4	EU	1254.74	5.80	5.83	4.58	3.00	0.11	0.16	0.04	0.11	0.32
	AU	373.50	3.39	3.22	2.26	2.38	0.75	0.44	0.23	0.75	0.74
	NA	974.04	4.91	5.34	3.46	2.20	0.39	0.30	0.19	0.33	0.42
	SA	435.67	3.28	3.59	2.31	1.58	0.49	0.53	0.37	0.48	0.54
5	EU	1254.74	5.80	5.83	4.58	3.00	0.11	0.16	0.04	0.11	0.32
	AU	373.50	3.39	3.22	2.26	2.38	0.75	0.44	0.23	0.75	0.74
	NA	974.04	4.91	5.34	3.46	2.20	0.39	0.30	0.19	0.33	0.42
	SA	435.67	3.28	3.59	2.31	1.58	0.49	0.53	0.37	0.48	0.54
	SAM	534.26	3.06	3.15	2.91	3.20	0.61	0.19	0.15	0.59	0.66
6	EU	1254.74	5.80	5.83	4.58	3.00	0.11	0.16	0.04	0.11	0.32
	AU	373.50	3.39	3.22	2.26	2.38	0.75	0.44	0.23	0.75	0.74
	NA	974.04	4.91	5.34	3.46	2.20	0.39	0.30	0.19	0.33	0.42
	SA	435.67	3.28	3.59	2.31	1.58	0.49	0.53	0.37	0.48	0.54
	SAM	534.26	3.06	3.15	2.91	3.20	0.61	0.19	0.15	0.59	0.66
	EA	934.78	4.79	4.74	3.61	3.03	0.31	0.27	0.16	0.30	0.43

Table 16: Complete regional results for CPrompt at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1185.81	5.89	6.15	4.54	2.60	0.20	0.19	0.03	0.18	0.48
2	EU	2423.19	10.30	10.52	9.52	8.91	0.03	0.03	0.01	0.03	0.08
	AU	329.29	3.10	3.07	1.97	1.86	0.82	0.58	0.42	0.83	0.87
3	EU	1346.28	6.35	6.39	5.28	4.35	0.09	0.11	0.03	0.04	0.12
	AU	1223.82	4.43	4.04	7.87	8.24	0.42	0.33	0.25	0.24	0.18
	NA	937.09	4.91	5.66	3.41	2.07	0.48	0.36	0.21	0.42	0.54
4	EU	2387.10	12.99	14.69	10.71	9.38	-0.03	0.00	0.01	-0.01	0.01
	AU	1111.82	7.39	7.78	5.75	5.58	0.25	0.10	0.01	0.24	0.16
	NA	1822.92	10.09	11.99	7.79	6.37	0.10	0.07	0.03	0.09	0.05
	SA	447.94	3.25	3.72	2.26	1.37	0.48	0.56	0.38	0.52	0.70
5	EU	3241.84	23.83	22.55	14.80	16.88	0.02	-0.03	-0.02	0.03	0.02
	AU	528.15	4.37	4.06	3.22	3.63	0.65	0.22	0.12	0.62	0.56
	NA	2304.35	13.17	13.74	10.52	10.26	0.03	0.01	0.01	0.01	0.02
	SA	899.90	5.26	5.04	4.28	4.28	0.15	0.14	0.09	0.19	0.19
	SAM	492.78	2.91	3.11	2.71	2.42	0.66	0.35	0.28	0.65	0.82
6	EU	1438.38	6.98	6.84	5.38	4.19	0.09	0.06	0.01	0.03	0.16
	AU	885.74	4.80	4.44	4.81	4.88	0.34	0.17	0.06	0.33	0.28
	NA	1057.52	5.78	6.22	4.10	3.82	0.40	0.20	0.10	0.25	0.18
	SA	570.68	4.40	4.37	3.02	2.66	0.32	0.23	0.18	0.28	0.29
	SAM	938.18	4.20	3.89	4.42	5.46	0.34	0.14	0.13	0.31	0.34
	EA	826.41	4.52	4.87	3.26	2.33	0.48	0.42	0.21	0.46	0.68

Table 17: Complete regional results for SD-LoRA at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1158.25	5.70	6.08	4.43	2.51	0.28	0.26	0.11	0.22	0.50
2	EU	2021.43	7.40	6.09	7.02	7.17	-0.03	0.01	0.00	0.01	0.05
	AU	298.06	2.85	2.83	1.86	1.61	0.85	0.65	0.54	0.85	0.90
3	EU	1292.51	6.21	5.94	4.92	4.42	0.17	0.14	0.05	0.09	0.17
	AU	679.83	4.36	3.63	4.13	5.20	0.41	0.11	0.01	0.33	0.26
	NA	940.21	4.80	5.61	3.25	1.84	0.50	0.43	0.27	0.48	0.64
4	EU	1834.05	7.11	6.18	6.02	4.85	-0.02	0.01	-0.01	0.00	0.12
	AU	484.71	4.24	3.58	3.64	4.71	0.61	0.26	0.11	0.57	0.33
	NA	1177.77	5.77	5.87	4.01	3.25	0.24	0.17	0.12	0.18	0.13
	SA	421.19	3.00	3.52	2.17	1.32	0.55	0.64	0.46	0.57	0.73
5	EU	2157.74	7.07	6.44	7.41	6.99	0.01	0.03	0.00	0.05	0.12
	AU	554.57	3.65	3.55	2.90	3.91	0.69	0.33	0.19	0.67	0.48
	NA	1137.50	5.16	5.89	4.04	3.35	0.31	0.26	0.09	0.26	0.20
	SA	546.64	4.08	4.06	2.87	2.87	0.41	0.25	0.16	0.25	0.30
	SAM	495.49	2.76	3.01	2.71	2.39	0.68	0.44	0.36	0.66	0.83
6	EU	1290.15	6.07	5.93	4.69	3.27	0.16	0.15	0.06	0.10	0.24
	AU	759.96	4.38	4.04	3.99	5.28	0.50	0.12	-0.01	0.42	0.25
	NA	1191.81	5.75	6.06	4.03	3.54	0.30	0.23	0.11	0.24	0.14
	SA	509.72	4.71	4.19	2.82	2.58	0.39	0.17	0.16	0.29	0.23
	SAM	1598.03	5.19	3.72	5.61	7.00	0.32	0.07	-0.01	0.34	0.33
	EA	845.11	4.54	4.96	3.19	2.07	0.48	0.44	0.25	0.51	0.76

Table 18: Complete regional results for InfLoRA at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1103.55	5.53	5.82	4.25	2.43	0.33	0.29	0.12	0.26	0.51
2	EU	1130.81	5.65	5.82	4.32	2.72	0.28	0.25	0.07	0.23	0.43
	AU	298.64	2.74	2.74	1.82	1.72	0.85	0.67	0.55	0.85	0.88
3	EU	1210.30	5.92	6.04	4.74	3.57	0.24	0.22	0.07	0.14	0.24
	AU	321.19	2.95	2.92	2.01	2.00	0.83	0.62	0.48	0.84	0.86
	NA	858.78	4.53	5.10	3.09	2.05	0.57	0.46	0.35	0.51	0.51
4	EU	1279.99	6.08	5.99	5.04	3.89	0.17	0.19	0.06	0.07	0.20
	AU	336.02	3.24	3.18	2.20	2.27	0.82	0.53	0.39	0.81	0.84
	NA	974.11	4.75	5.31	3.51	2.31	0.52	0.42	0.28	0.40	0.39
	SA	396.33	2.89	3.35	2.12	1.36	0.60	0.66	0.50	0.59	0.70
5	EU	1213.37	5.88	5.99	4.62	3.21	0.19	0.18	0.07	0.12	0.29
	AU	421.98	3.15	3.11	2.26	2.83	0.80	0.55	0.39	0.79	0.70
	NA	921.20	4.61	5.23	3.50	2.36	0.54	0.43	0.28	0.45	0.38
	SA	440.66	3.42	3.55	2.40	1.73	0.55	0.49	0.41	0.46	0.49
	SAM	458.35	2.64	2.84	2.56	2.33	0.72	0.47	0.42	0.70	0.83
6	EU	1227.14	5.82	5.88	4.61	3.44	0.21	0.19	0.06	0.12	0.31
	AU	424.85	3.19	3.29	2.67	2.86	0.76	0.53	0.31	0.75	0.71
	NA	931.34	4.76	5.40	3.41	2.61	0.53	0.38	0.25	0.44	0.38
	SA	458.38	3.40	3.60	2.39	2.11	0.51	0.54	0.39	0.47	0.42
	SAM	500.82	2.97	3.12	3.04	3.29	0.68	0.34	0.27	0.64	0.71
	EA	772.94	4.19	4.50	3.03	2.11	0.54	0.51	0.34	0.54	0.74

Table 19: Complete regional results for GUST (Ours) at a three-day forecast horizon. Stages and region abbreviations follow Appendix D. Only observed regions are included.

		wRMSE ↓					ACC ↑				
Stage	Region	Z500	U10	V10	T850	T2M	Z500	U10	V10	T850	T2M
1	EU	1085.55	5.47	5.75	4.22	2.39	0.32	0.30	0.14	0.25	0.52
2	EU	1127.69	5.59	5.80	4.36	2.48	0.28	0.25	0.11	0.21	0.50
	AU	288.85	2.70	2.72	1.79	1.56	0.86	0.69	0.58	0.86	0.90
3	EU	1162.08	5.73	5.87	4.56	3.07	0.25	0.25	0.09	0.15	0.37
	AU	296.89	2.75	2.77	1.83	1.68	0.85	0.68	0.56	0.85	0.89
	NA	856.09	4.46	5.12	3.03	1.80	0.57	0.48	0.35	0.53	0.64
4	EU	1169.86	5.73	5.84	4.65	3.38	0.24	0.24	0.09	0.16	0.35
	AU	303.63	2.80	2.85	1.88	1.71	0.85	0.66	0.53	0.84	0.88
	NA	871.35	4.49	5.15	3.21	2.05	0.56	0.47	0.34	0.51	0.61
	SA	390.15	2.87	3.30	2.03	1.19	0.61	0.67	0.52	0.62	0.77
5	EU	1182.01	5.72	5.83	4.55	3.15	0.23	0.23	0.10	0.17	0.34
	AU	325.78	2.83	2.89	1.91	1.92	0.83	0.65	0.50	0.83	0.86
	NA	879.26	4.51	5.13	3.16	2.09	0.55	0.45	0.33	0.48	0.53
	SA	392.63	2.92	3.33	2.25	1.67	0.61	0.65	0.51	0.58	0.71
	SAM	449.90	2.59	2.84	2.50	2.15	0.73	0.50	0.44	0.71	0.85
6	EU	1161.19	5.73	5.88	4.44	2.64	0.25	0.23	0.09	0.18	0.42
	AU	329.51	2.85	2.91	1.95	1.84	0.83	0.65	0.50	0.84	0.87
	NA	866.33	4.55	5.16	3.07	1.93	0.56	0.45	0.33	0.51	0.61
	SA	404.85	3.00	3.42	2.23	1.42	0.58	0.63	0.47	0.56	0.72
	SAM	456.92	2.65	2.89	2.60	2.24	0.72	0.48	0.42	0.69	0.84
	EA	786.52	4.21	4.57	2.99	1.96	0.54	0.51	0.33	0.56	0.78

E Detailed Results for Climate-Projection Continual Learning

Stages 1–4 successively introduce tas, pr, pr90, and the diurnal temperature range (DTR, labeled "diurnal" in the source). Figure 7 shows all four metrics for each target. NRMSEs and NRMSEg denote spatial and global NRMSE; lower is better for every metric. Total is reproduced as supplied, rather than reconstructed from rounded component values. Source Total entries for unseen targets are sometimes zero-filled; these are treated as unavailable, shown as dashes in the tables, and excluded from plots and rankings. All observed values retain two decimals. The supplied results contain no repeated-run uncertainty estimates.

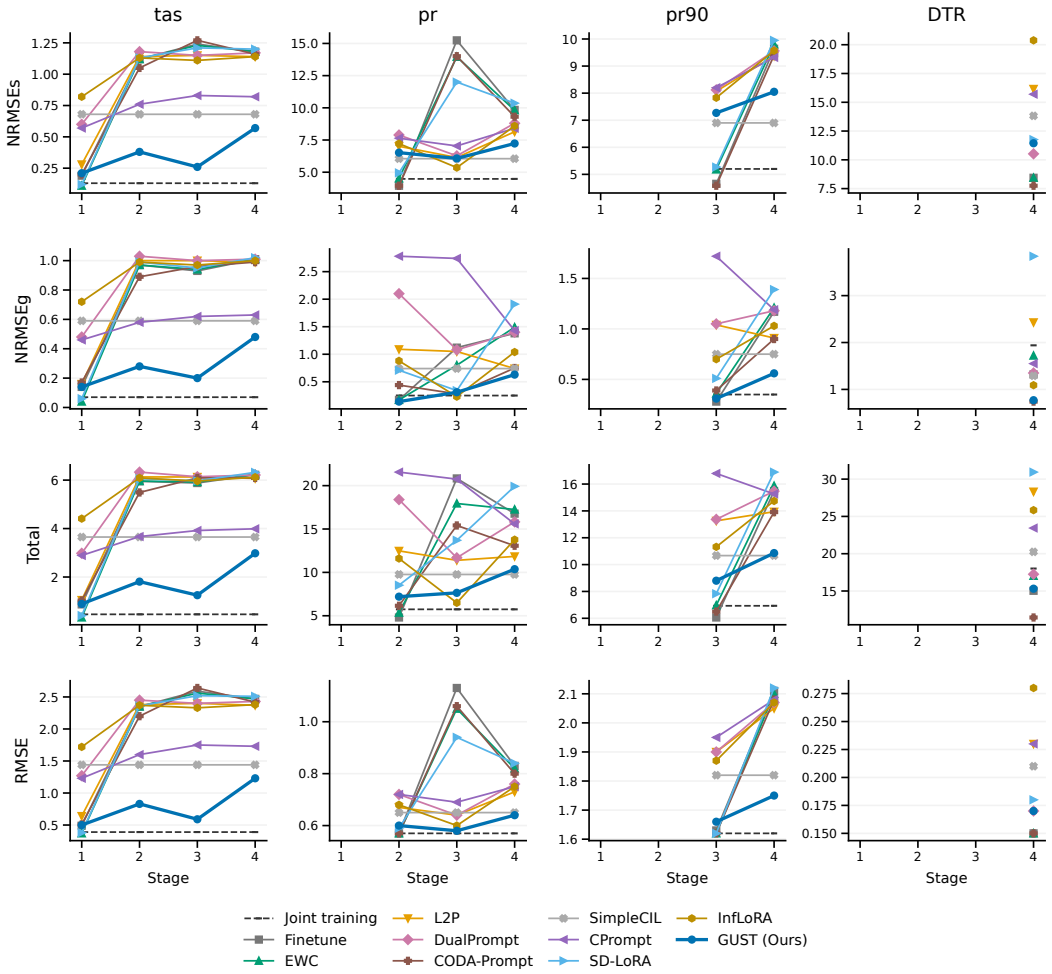


Figure 7: Climate-projection performance across learning stages. Columns correspond to prediction targets; rows correspond to metrics. Each panel has its own vertical scale. Trajectories start at acquisition, so DTR has a single observed stage. Dashed lines indicate fixed offline joint-training references over observed stages. Initial performance varies across methods; absolute scores alone do not isolate forgetting.

Table 20: Complete NRMSEs results for climate-projection continual learning (lower is better). Stages 1–4 introduce tas, pr, pr90, and DTR, respectively. Dashes indicate targets not yet observed. Bold marks the best continual result for the same stage and target, including ties at two-decimal precision; joint training is excluded from this ranking.

Method	Stage	tas	pr	pr90	DTR
Joint training	Offline	0.13	4.48	5.20	8.31
Finetune	1	0.19	–	–	–
	2	1.12	**3.95**	–	–
	3	1.24	15.24	4.65	–
	4	1.18	9.84	9.64	8.44
EWC	1	**0.11**	–	–	–
	2	1.12	4.52	–	–
	3	1.23	13.96	5.18	–
	4	1.18	9.81	9.80	8.44
L2P	1	0.28	–	–	–
	2	1.14	7.01	–	–
	3	1.15	6.14	8.07	–
	4	1.14	8.14	9.37	16.16
DualPrompt	1	0.60	–	–	–
	2	1.18	7.89	–	–
	3	1.15	6.27	8.12	–
	4	1.17	8.80	9.56	10.52
CODA-Prompt	1	0.20	–	–	–
	2	1.05	**3.95**	–	–
	3	1.27	14.01	**4.58**	–
	4	1.16	9.31	9.40	**7.75**
SimpleCIL	1	0.68	–	–	–
	2	0.68	6.05	–	–
	3	0.68	6.05	6.90	–
	4	0.68	**6.05**	**6.90**	13.82
CPrompt	1	0.57	–	–	–
	2	0.76	7.64	–	–
	3	0.83	7.04	8.20	–
	4	0.82	8.41	9.32	15.70
SD-LoRA	1	0.12	–	–	–
	2	1.13	4.95	–	–
	3	1.21	12.00	5.27	–
	4	1.20	10.36	9.95	11.76
InfLoRA	1	0.82	–	–	–
	2	1.13	7.22	–	–
	3	1.11	**5.36**	7.83	–
	4	1.14	8.59	9.57	20.38
GUST (Ours)	1	0.21	–	–	–
	2	**0.38**	6.52	–	–
	3	**0.26**	6.07	7.27	–
	4	**0.57**	7.23	8.05	11.44

Table 21: Complete NRMSEg results for climate-projection continual learning (lower is better). Stages 1–4 introduce tas, pr, pr90, and DTR, respectively. Dashes indicate targets not yet observed. Bold marks the best continual result for the same stage and target, including ties at two-decimal precision; joint training is excluded from this ranking.

Method	Stage	tas	pr	pr90	DTR
Joint training	Offline	0.07	0.25	0.35	1.94
Finetune	1	0.14	–	–	–
	2	0.97	0.17	–	–
	3	0.93	1.12	**0.28**	–
	4	1.01	1.38	1.17	1.32
EWC	1	**0.04**	–	–	–
	2	0.97	0.17	–	–
	3	0.94	0.80	0.36	–
	4	1.01	1.49	1.21	1.72
L2P	1	0.16	–	–	–
	2	1.00	1.09	–	–
	3	1.00	1.05	1.04	–
	4	0.99	0.74	0.91	2.43
DualPrompt	1	0.48	–	–	–
	2	1.03	2.10	–	–
	3	1.00	1.08	1.05	–
	4	1.01	1.40	1.18	1.35
CODA-Prompt	1	0.17	–	–	–
	2	0.89	0.44	–	–
	3	0.96	0.28	0.39	–
	4	0.99	0.75	0.90	**0.74**
SimpleCIL	1	0.59	–	–	–
	2	0.59	0.74	–	–
	3	0.59	0.74	0.75	–
	4	0.59	0.74	0.75	1.29
CPrompt	1	0.46	–	–	–
	2	0.58	2.78	–	–
	3	0.62	2.74	1.72	–
	4	0.63	1.43	1.19	1.55
SD-LoRA	1	0.06	–	–	–
	2	0.99	0.71	–	–
	3	0.95	0.34	0.51	–
	4	1.02	1.91	1.39	3.84
InfLoRA	1	0.72	–	–	–
	2	0.99	0.88	–	–
	3	0.97	**0.23**	0.70	–
	4	1.00	1.04	1.03	1.09
GUST (Ours)	1	0.14	–	–	–
	2	**0.28**	**0.14**	–	–
	3	**0.20**	0.31	0.31	–
	4	**0.48**	**0.63**	**0.56**	0.77

Table 22: Complete Total results for climate-projection continual learning (lower is better). Stages 1–4 introduce tas, pr, pr90, and DTR, respectively. Dashes indicate targets not yet observed. Bold marks the best continual result for the same stage and target, including ties at two-decimal precision; joint training is excluded from this ranking.

Method	Stage	tas	pr	pr90	DTR
Joint training	Offline	0.46	5.74	6.93	18.01
Finetune	1	0.87	–	–	–
	2	5.95	**4.80**	–	–
	3	5.88	20.83	**6.06**	–
	4	6.22	16.77	15.48	15.03
EWC	1	**0.33**	–	–	–
	2	5.96	5.35	–	–
	3	5.92	17.94	6.97	–
	4	6.24	17.25	15.87	17.03
L2P	1	1.05	–	–	–
	2	6.13	12.48	–	–
	3	6.13	11.39	13.26	–
	4	6.07	11.84	13.94	28.32
DualPrompt	1	2.98	–	–	–
	2	6.33	18.39	–	–
	3	6.14	11.68	13.37	–
	4	6.21	15.82	15.47	17.25
CODA-Prompt	1	1.03	–	–	–
	2	5.49	6.15	–	–
	3	6.09	15.41	6.50	–
	4	6.09	13.07	13.91	**11.45**
SimpleCIL	1	3.65	–	–	–
	2	3.65	9.76	–	–
	3	3.65	9.76	10.67	–
	4	3.65	**9.76**	**10.67**	20.25
CPrompt	1	2.89	–	–	–
	2	3.67	21.56	–	–
	3	3.92	20.75	16.79	–
	4	3.99	15.59	15.26	23.44
SD-LoRA	1	0.41	–	–	–
	2	6.06	8.52	–	–
	3	5.97	13.70	7.83	–
	4	6.32	19.92	16.89	30.94
InfLoRA	1	4.41	–	–	–
	2	6.09	11.60	–	–
	3	5.96	**6.49**	11.32	–
	4	6.12	13.77	14.74	25.84
GUST (Ours)	1	0.89	–	–	–
	2	**1.81**	7.20	–	–
	3	**1.25**	7.64	8.80	–
	4	**2.98**	10.37	10.86	15.31

Table 23: Complete RMSE results for climate-projection continual learning (lower is better). Stages 1–4 introduce tas, pr, pr90, and DTR, respectively. Dashes indicate targets not yet observed. Bold marks the best continual result for the same stage and target, including ties at two-decimal precision; joint training is excluded from this ranking.

Method	Stage	tas	pr	pr90	DTR
Joint training	Offline	0.39	0.57	1.62	0.15
Finetune	1	0.49	–	–	–
	2	2.35	**0.57**	–	–
	3	2.59	1.13	1.63	–
	4	2.47	0.83	2.11	**0.15**
EWC	1	**0.37**	–	–	–
	2	2.35	**0.57**	–	–
	3	2.56	1.05	**1.62**	–
	4	2.47	0.82	2.10	**0.15**
L2P	1	0.64	–	–	–
	2	2.38	0.67	–	–
	3	2.40	0.64	1.90	–
	4	2.37	0.73	2.05	0.23
DualPrompt	1	1.27	–	–	–
	2	2.45	0.72	–	–
	3	2.40	0.64	1.90	–
	4	2.43	0.76	2.07	0.17
CODA-Prompt	1	0.51	–	–	–
	2	2.20	**0.57**	–	–
	3	2.64	1.06	**1.62**	–
	4	2.42	0.80	2.08	**0.15**
SimpleCIL	1	1.44	–	–	–
	2	1.44	0.65	–	–
	3	1.44	0.65	1.82	–
	4	1.44	0.65	1.82	0.21
CPrompt	1	1.23	–	–	–
	2	1.60	0.72	–	–
	3	1.75	0.69	1.95	–
	4	1.73	0.75	2.08	0.23
SD-LoRA	1	0.38	–	–	–
	2	2.36	0.59	–	–
	3	2.52	0.94	**1.62**	–
	4	2.51	0.84	2.12	0.18
InfLoRA	1	1.72	–	–	–
	2	2.37	0.68	–	–
	3	2.33	0.60	1.87	–
	4	2.38	0.75	2.07	0.28
GUST (Ours)	1	0.50	–	–	–
	2	**0.83**	0.60	–	–
	3	**0.59**	**0.58**	1.66	–
	4	**1.23**	**0.64**	**1.75**	0.17