COUNTERFACTUAL PREDICTIONS IN SCIENTIFIC EMULATORS WITHOUT CONTROLLED EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many scientific questions require reasoning about what was never observed: *What if the conditions, interventions, or history had been different?* Models can predict accurately on observed data yet fail on such what-if queries when correlated inputs are varied independently. A common remedy is to add controlled simulation data in which these factors are explicitly disentangled, but this requires access to a simulator, can be computationally expensive, and inherits the simulator's modeling assumptions. We introduce *ReRoute*, a framework for targeted scientific what-if prediction that combines factual data with partial mechanistic knowledge, without requiring controlled intervention data for adaptation. *ReRoute* fixes the queried input of a pretrained backbone to a reference value, reintroduces its variation through a known mechanistic pathway, and fine-tunes on the original factual data, while leaving downstream effects to the learned dynamics. We provide a causal identification result for this construction under explicit structural assumptions, with the core argument machine-checked in Lean. After showing that *ReRoute* achieves highly accurate counterfactual predictions in a controlled advection–diffusion system where exact responses are available, we turn to state-of-the-art climate emulation. On held-out coupled-climate interventions, *ReRoute* reduces aggregate climate error by 18.2–31.8% under severe CO_2 distribution shifts while preserving skill under standard conditions, at a small fraction of the cost of retraining on additional controlled simulations, without even accounting for the substantial expense of generating such data. Finally, on an emulator trained from historical ERA5 reanalysis, where no counterfactual reference exists, *ReRoute* preserves substantially more of the surface warming implied by the observed boundary conditions under a fixed-CO_2 counterfactual.

1 INTRODUCTION

Data-driven surrogate models, or emulators, are becoming a standard tool for approximating complex scientific systems. Given the current state of a system together with external conditions, they predict its target at a fraction of the cost of traditional simulation. But scientific questions often requires reasoning beyond predictions: We want to know how a system would respond under a change we have not directly observed: how would the atmosphere respond if CO_2 level had never risen (Watt-Meyer et al., 2025; Clark et al., 2026), how would a biological system respond to a different treatment or dose (Feuerriegel et al., 2024), or how would a physical system evolve under a modified forcing (Rigas et al., 2017)? Answering such *what-if questions* requires models to understand the mechanism of the underlying system, which is one step further than merely matching the observational distribution.

A central challenge arises when the training data contain only correlated rather than independently varied *drivers* (i.e. variables whose values influence the outcome of interest, such as environmental conditions, treatments, or physical forcings). A flexible model can use one input as a proxy for another and still achieve excellent predictive accuracy, while learning the wrong underlying mechanism. Standard validation does not expose this shortcut when the same dependencies persist at test time; it becomes visible only when a what-if query breaks them. A direct solution is to augment training with controlled experiments in which the relevant drivers are varied independently. Such experiments provide exactly the missing variation, but require access to a simulator, can be costly to generate at scale, and inherit the simulator's own modeling assumptions and biases. For many real scientific systems, the corresponding controlled perturbation data are simply unavailable.

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.

This challenge is particularly relevant in weather and climate science, where data-driven systems are increasingly prominent, as evidenced by recent advances such as FourCastNet (Pathak et al., 2022; Bonev et al., 2025) and GenCast (Price et al., 2025), as well as climate emulators like ACE2 series (Watt-Meyer et al., 2025; Clark et al., 2025). Historical climate trajectories illustrate the challenge of co-varying drivers: atmospheric CO_2 rises as sea surface temperature (SST) warms and sea ice declines. An emulator trained on these trajectories can therefore reproduce historical warming while relying too strongly on CO_2 as a proxy for the atmospheric response (Watt-Meyer et al., 2025; Zhang & Merlis, 2026; Clark et al., 2026). If CO_2 is held fixed while SST and sea ice are varied, this learned dependence can yield unrealistic responses. The model is thus not necessarily extrapolating beyond the range of any single input, but rather to an *unseen combination of inputs*. The controlled-experiment approach provides a direct remedy in this setting: additional physics-based simulations can be generated in which SST and CO_2 are varied more independently. Recent climate work uses precisely this strategy to decorrelate these drivers during training, substantially improving the separation of their effects (Clark et al., 2026).

Scientific models, however, often come with partial mechanistic knowledge. We may know how a targeted driver first enters the system, for example, through a radiative, chemical, mechanical, or physiological pathway, but without knowing enough to model all of its downstream interactions. This suggests a different source of information for answering what-if questions:

> *Can partial mechanistic knowledge enable reliable what-if predictions when controlled training experiments are unavailable?*

Proposed approach. We introduce *ReRoute*, a framework that answers targeted what-if questions by combining observational data with known mechanistic pathways, without requiring controlled or interventional training data. Let $\mathbf{s}$ denote the input variable targeted by the what-if query. *ReRoute* fixes the value seen by the pretrained backbone to a reference value $\mathbf{s}_0$ and reintroduces its factual value through a supplied mechanistic pathway $\Gamma(\mathbf{s}_0, \mathbf{s})$, which describes the first level response of intervening the system from $\mathbf{s}_0$ to $\mathbf{s}$ while leaving remaining responses to be learned from data.

We demonstrate that *ReRoute* admits a causal identification guarantee: under explicit structural assumptions, the target interventional response is identifiable from observational trajectories together with the known mechanism. The formal conditions and Lean-verified proofs are provided in § A.

> **When does *ReRoute* apply?** *ReRoute* is a general scientific-ML framework. It applies when three ingredients are available: (1) a pretrained emulator, (2) a targeted input whose variation is entangled with other inputs in the training data, and (3) partial knowledge coming from physics, chemistry, biology, or other domain sciences which only needs to characterize how this targeted input enters the system, not the full chain of subsequent effects. *ReRoute* uses this pathway to route the targeted perturbation while leaving the remaining dynamics learned from data.

This setting arises across scientific domains when a pretrained emulator takes several variables as inputs, although they were not varied independently in its training data. In historical climate data such as ERA5 (Hersbach et al., 2020), CO_2 forcing and SST co-evolve, so an emulator may entangle their effects. Radiative physics specifies how a CO_2 intervention modifies the atmospheric forcing while preserving the SST evolution, leaving the emulator to predict the downstream atmospheric response. In chemical reaction datasets, catalyst concentration is often set as a fixed fraction of reactant concentration, so experiments with more reactant also contain more catalyst. Chemistry specifies whether an increase should preserve the catalyst concentration or the catalyst-to-reactant ratio, while the emulator predicts reaction rate, yield, and selectivity (Clayden et al., 2012; Atkins et al., 2022). Exercise datasets often associate greater intensity with shorter duration because intense workloads cannot be sustained for as long. Physiology specifies whether an increase in intensity should preserve duration, total work, or energy expenditure, while the emulator predicts cardiovascular and metabolic responses (Kenney et al., 2025). Nutrition datasets may associate higher caloric intake with greater protein intake because increasing food intake often changes both. Nutritional knowledge specifies the macronutrient composition of the additional calories, allowing the emulator to predict effects on metabolism and body composition (Gropper et al., 2022). In all these examples, domain knowledge need only translate the targeted intervention into a scientifically coherent change to the model inputs, leaving the pretrained emulator to predict the downstream response.

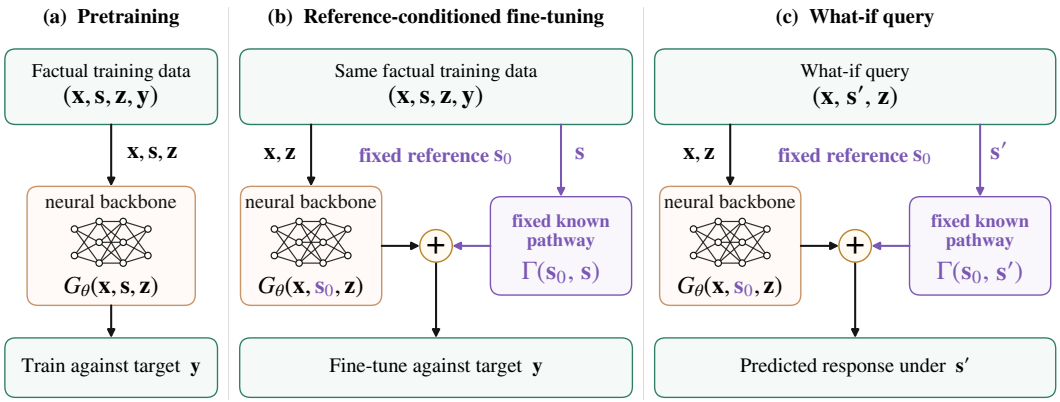

Figure 1: Overview of the *ReRoute* framework. **(a)** Standard training: the emulator G_θ is trained on factual samples $(\mathbf{x}, \mathbf{s}, \mathbf{z}, \mathbf{y})$. **(b)** *ReRoute* initializes from the pretrained operator G_θ, replaces the targeted input $\mathbf{s}$ seen by the backbone with the fixed reference $\mathbf{s}_0$, and reintroduces the factual value $\mathbf{s}$ only through the fixed known pathway $\Gamma(\mathbf{s}_0, \mathbf{s})$. The model is fine-tuned using the same factual data $(\mathbf{x}, \mathbf{s}, \mathbf{z}, \mathbf{y})$. **(c)** For a what-if query $\mathbf{s}'$, the backbone remains conditioned on $\mathbf{s}_0$, while $\mathbf{s}'$ enters through the pathway $\Gamma(\mathbf{s}_0, \mathbf{s}')$, producing the predicted response under the requested value.

Summary of Results. *ReRoute* consistently improves counterfactual prediction across controlled and climate CO_2 experiments. In advection–diffusion with exact counterfactual responses, it reduces response error by up to 89.6%, including over rollouts $10\times$ longer than the fine-tuning horizon. Applied to pretrained climate emulators, it reduces aggregate error by 18.2–31.8% under severe SST–CO_2 shifts without controlled-CO_2 supervision, while maintaining comparable skill under standard and mildly shifted conditions. Finally, in an emulator trained directly on historical ERA5 reanalysis (Hersbach et al., 2020), *ReRoute* preserves substantially more near-surface warming when CO_2 is held fixed while observed SST and sea-ice evolution are retained, and yields a tropical-cyclone climatology closer to ERA5. These gains require only light fine-tuning and no additional interventional data, at a fraction of the cost of retraining on controlled simulations.

2 Related Work

Scientific generalization beyond observed regimes. A central challenge in scientific machine learning is generalization to conditions that are absent or weakly represented in the training distribution, including unseen interventions, forcing regimes, and system configurations. Existing approaches seek such generalization by learning representations or structural invariances that remain stable across environments (Schölkopf et al., 2021; Yao et al., 2025; Yu et al., 2025), or by relying on sufficient observational support for the target regime (Imbens & Rubin, 2015; Hernán & Robins, 2020). In scientific applications, however, the most interesting queries often deliberately move beyond this support, while collecting new controlled experiments or simulations may be expensive or impossible. *ReRoute* addresses this setting through targeted mechanistic extrapolation: rather than learning a general-purpose representation that must transfer to arbitrary distribution shifts, it uses partial scientific knowledge about how a specific queried variable enters the system and leaves the remaining response to the pretrained model. This replaces broad assumptions about transfer across arbitrary shifts with a narrower, query-specific assumption about how the targeted variable enters the system.

Hybrid scientific modeling. Hybrid scientific models incorporate physical structure through physics-informed objectives (Raissi et al., 2019; Karniadakis et al., 2021; Li et al., 2024), dynamical or spectral regularization (McCabe et al., 2023; Chattopadhyay et al., 2023), hard constraints such as conservation laws (Beucler et al., 2019; 2021; Harder et al., 2023), and formulations that combine learned components with known mechanistic dynamics or closures (Pan & Duraisamy, 2018; Rackauckas et al., 2020). These approaches typically encode governing equations, conservation relations, or substantial dynamical structure of the system. *ReRoute* uses a more local form of scientific knowledge: it requires only the pathway through which the queried variable first enters the system, while complex downstream interactions and feedbacks remain learned by the pretrained emulator. This makes *ReRoute* a lightweight adaptation strategy for targeted what-if prediction without requiring a complete mechanistic model.

Climate emulators and forcing disentanglement. Our experiments use the ACE2 family of machine-learned climate emulators as a concrete instance of this problem. ACE2 is conditioned on prescribed SST, sea ice, and atmospheric CO_2, whose strong co-variation in historical training data can lead to unrealistic responses when these forcings are varied independently (Watt-Meyer et al., 2025). ACE2-SOM extends the training distribution to substantially different equilibrium CO_2 climates using controlled simulations (Clark et al., 2025), while Clark et al. (2026) directly address SST–CO_2 entanglement by generating random-CO_2 simulations in which the two forcings vary independently. Mahesh et al. (2026) further show that historically trained weather emulators can reproduce fast radiatively driven responses to unprecedented CO_2 perturbations when the relevant response is already represented in historical variability. *ReRoute* targets the complementary setting in which direct manipulation of the learned forcing input is insufficient and additional controlled simulations are unavailable: it supplies the known radiative entry pathway for CO_2 while retaining the emulator's learned atmospheric and coupled dynamics, enabling targeted forcing interventions using only the original training distribution.

3 *ReRoute*: Mechanism-Guided Adaptation

This section formalizes *ReRoute*. We first define the targeted what-if query, then introduce the known pathway and the reference-conditioned fine-tuning procedure. The key idea is simple: *remove the queried driver from the flexible backbone and reintroduce it through a deterministic pathway whose form is known*. The remaining system response is still learned from data.

> ***ReRoute* in one sentence**: Keep the input variable that we targeted to intervene fixed inside the pretrained backbone, route its variation through the known mechanism, and fine-tune the complete model on the original data.

3.1 Problem Setup: Targeted What-If Queries

Let $\mathbf{x} \in \mathcal{X}$ denote the current system state. We partition the externally specified inputs as

$$\mathbf{u} = (\mathbf{s}, \mathbf{z}), \qquad \mathbf{s} \in \mathcal{S}, \quad \mathbf{z} \in \mathcal{Z},$$

where $\mathbf{s}$ is the *targeted input variable* that we want to vary in a what-if query, and $\mathbf{z}$ contains the remaining conditions. The system response is

$$\mathbf{y} = f^*(\mathbf{x}, \mathbf{s}, \mathbf{z}), \qquad f^* : \mathcal{X} \times \mathcal{S} \times \mathcal{Z} \to \mathbb{R}^d. \tag{3.1}$$

Training samples come from the distribution available for model training $(\mathbf{X}, \mathbf{S}, \mathbf{Z}, \mathbf{Y}) \sim P_{\text{train}}$, where $\mathbf{Y} = f^*(\mathbf{X}, \mathbf{S}, \mathbf{Z})$. Our goal is to answer a targeted query $f^*(\mathbf{x}, \mathbf{s}', \mathbf{z})$, where $\mathbf{s}' \in \mathcal{S}$ is a requested value of the targeted driver. Throughout the paper, we use *counterfactual prediction* to mean predicting the evolution of a realized system under an alternative forcing trajectory while holding the remaining observed conditions fixed. Our causal analysis formalizes the corresponding intervention on the targeted driver at the distributional level.

> **Climate example**: *What if CO_2 had stayed at its 1940 level while other boundary conditions evolved as they did historically?* Here, $\mathbf{s}$ is atmospheric CO_2, $\mathbf{z}$ contains the remaining control inputs such as SST and sea ice, and f^* is the unknown full climate response. Given only the historical trajectory, our goal is to learn an emulator that can answer this unseen fixed-CO_2 query.

3.2 A Known Pathway for the Targeted Driver

Choose a reference value $\mathbf{s}_0 \in \mathcal{S}$. We assume access to a fixed pathway

$$\Gamma(\mathbf{s}_0, \cdot) : \mathcal{S} \to \mathbb{R}^d, \qquad \Gamma(\mathbf{s}_0, \mathbf{s}_0) = 0, \tag{3.2}$$

which describes how changing $\mathbf{s}$ away from $\mathbf{s}_0$ directly enters the model response.

Notably, the pathway does *not* need to describe the full system dynamics. It only needs to encode the part of the response that is scientifically known for the targeted driver. For example, in our climate application, Γ describes the immediate atmospheric effect of changing CO_2, while the subsequent adjustment of the climate state remains learned by the emulator.

What does *ReRoute* assume? *ReRoute* does *not* assume that the complete response to $\mathbf{s}'$ is known. It assumes only that we know how the targeted driver first enters the system. *ReRoute* then lets the learned model propagate that effect through the remaining dynamics.

3.3 Reference-Conditioned Fine-Tuning

Let G_{θ_0} denote a pretrained emulator. *ReRoute* initializes $\theta \leftarrow \theta_0$, replaces the targeted input seen by the backbone with the fixed reference $\mathbf{s}_0$, and reintroduces the factual value $\mathbf{s}$ only through the known pathway Γ:

$$f_\theta(\mathbf{x}, \mathbf{s}, \mathbf{z}) = G_\theta(\mathbf{x}, \mathbf{s}_0, \mathbf{z}) + \Gamma(\mathbf{s}_0, \mathbf{s}). \tag{3.3}$$

The *ReRoute*d model is fine-tuned against the original training targets:

$$\mathcal{L}_{\text{train}}(\theta) = \mathbb{E}_{P_{\text{train}}}\left[\left\|G_\theta(\mathbf{X}, \mathbf{s}_0, \mathbf{Z}) + \Gamma(\mathbf{s}_0, \mathbf{S}) - \mathbf{Y}\right\|^2\right]. \tag{3.4}$$

What is the intuition behind *ReRoute*? Intuitively, the pretrained backbone can use correlations between $\mathbf{s}$ and other inputs to explain the response. After *ReRoute*, the backbone always sees the same $\mathbf{s}_0$. Variation in the targeted driver can therefore enter the prediction only through Γ. Fine-tuning lets the backbone adjust to this new division of responsibility.

When answering a what-if query, the backbone remains fixed at the same reference condition $\mathbf{s}_0$, while the pathway receives the requested value $\mathbf{s}'$, as shown in Fig. 1.

3.4 Identifiability of Targeted Interventions

The reference-conditioned construction above also admits a causal interpretation. Under the structural causal model developed in § A, *ReRoute identifies the interventional response of the targeted variable from observational trajectories and a known first-level mechanism*, without requiring controlled intervention data. The result relies on the targeted variable entering the downstream system through the supplied mechanistic pathway, observational variation being informative about the downstream response, and this response being sufficiently constrained to determine the requested pathway value.

Theorem 3.1 (Informal)**.** *Let S be the variable targeted by a what-if query, and suppose that its direct effect on the downstream system is fully mediated by a known first-level pathway $\Gamma(\mathbf{s}_0, \mathbf{S})$. If observational variation in this pathway identifies the response of the same downstream system, and the responses observed over factual pathway values uniquely determine the response at the target value $\Gamma(\mathbf{s}_0, \mathbf{s}')$, then $P(X \mid \mathrm{do}(\mathbf{S} = \mathbf{s}'))$ is identifiable from the observational distribution together with Γ.*

Here, $\mathrm{do}(\mathbf{S} = \mathbf{s}')$ means considering a hypothetical system in which $\mathbf{S}$ is directly set to $\mathbf{s}'$, independently of the factors that normally determine it. This differs from conditioning on $\mathbf{S} = \mathbf{s}'$, which only selects cases where $\mathbf{S}$ naturally takes the value $\mathbf{s}'$ and thus preserves the relationships between $\mathbf{S}$ and the factors that influence it. Formal assumptions, statements, and Lean-verified proofs are given in § A.

Why observational data can be sufficient? Intuitively, the known mechanism changes the identification problem. Without additional structure, correlated observational drivers may be interchangeable to a flexible predictor: the model can fit the factual distribution while assigning the response to the wrong input. *ReRoute* removes this ambiguity at the point where the targeted input $\mathbf{S}$ enters the model. The backbone always receives the same reference value $\mathbf{s}_0$, so variation in $\mathbf{S}$ can affect the prediction only through $\Gamma(\mathbf{s}_0, \mathbf{S})$.

Observational trajectories then provide examples of how the unknown downstream system responds to different values of this pathway. *ReRoute* does not need to know those downstream mechanisms explicitly: it learns their response from the factual data. A what-if query simply evaluates the same learned downstream response at the pathway value corresponding to the requested intervention $\Gamma(\mathbf{s}_0, \mathbf{s}')$. Thus, the mechanistic pathway supplies the missing attribution, i.e., how the targeted variable enters the system, while observational data supply the remaining downstream dynamics.

3.5 *ReRoute* FOR AUTOREGRESSIVE SYSTEMS

Many scientific emulators, evolve the system autoregressively. Let $\mathbf{x}_{t+1} = f^*(\mathbf{x}_t, \mathbf{s}_t, \mathbf{z}_t)$ denote the true one-step transition, where $\mathbf{s}_t$ is the targeted forcing and $\mathbf{z}_t$ collects the remaining inputs. During fine-tuning, *ReRoute* predicts $\widehat{\mathbf{x}}_{t+1} = G_\theta(\mathbf{x}_t, \mathbf{s}_0, \mathbf{z}_t) + \Gamma(\mathbf{s}_0, \mathbf{s}_t)$, where the backbone is conditioned on the fixed reference value $\mathbf{s}_0$, while the observed change in the targeted forcing enters through the mechanistic pathway. For a what-if query with intervention $\mathbf{s}'_t$, only the pathway input is changed: $\widehat{\mathbf{x}}_{t+1} = G_\theta(\widehat{\mathbf{x}}_t, \mathbf{s}_0, \mathbf{z}_t) + \Gamma(\mathbf{s}_0, \mathbf{s}'_t)$. The predicted state is fed back at the next timestep. Thus, the pathway specifies only the first-level response to the intervention; its subsequent effects propagate through the learned autoregressive dynamics.

4 EXPERIMENTS AND RESULTS

4.1 ADVECTION–DIFFUSION TRANSPORT

We first evaluate *ReRoute* in a fully controlled dynamical system, where both the governing dynamics and the counterfactual response are known exactly. Advection–diffusion is a canonical model for the transport and mixing of quantities such as heat, pollutants, and atmospheric tracers, and appears broadly across physics, engineering, and environmental science. This provides a clean test of whether *ReRoute* can disentangle strongly correlated controls in a fundamental transport problem before moving to more complex settings. We consider the two-dimensional advection–diffusion equation on a periodic domain,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = D\nabla^2 u, \tag{4.1}$$

where c is the advection velocity and D is the diffusion coefficient. Training trajectories contain strongly correlated (c, D) pairs, with Pearson correlation $r = 0.989$, allowing the model to achieve low factual error within this correlated training regime without necessarily learning the effects of c and D independently. We generate 512 training trajectories from 32 initial conditions, each containing 100 transitions on a 64×64 grid, with independent validation and test initial conditions.

Our baseline is a Fourier neural operator (FNO) (Li et al., 2021) trained from scratch on these correlated trajectories. *ReRoute* is initialized from this pretrained FNO and fine-tuned only on the original factual data. Because the dynamics are known analytically, we can construct the exact one-step pathway

$$\Gamma(u_t; c, D) = \Phi^h_{c,D}(u_t) - \Phi^h_{c_0,D_0}(u_t), \tag{4.2}$$

with reference values $c_0 = 1$ and $D_0 = 0.010$. Here, $\Phi^h_{c,D}(u_t)$ denotes the field given by advancing u_t by one interval h with parameters c and D. We evaluate both rerouting D alone and jointly rerouting (c, D). The latter clamps both controls in the backbone and introduces their joint effect through Γ. We evaluate 100-step autoregressive rollouts from held-out initial conditions over a 41×41 (c, D) grid spanning $[0.5, 1.5] \times [0.002, 0.018]$, including combinations far from the correlated training support. Our main metric is the RMSE of the predicted counterfactual response relative to the exact solution. Dataset construction and FNO pretraining are detailed in § B.1, the exact pathway is derived in § B.2, and the fine-tuning procedure and metric definitions are given in § B.3.

***ReRoute* separates strongly correlated transport controls.** Fig. 6 shows counterfactual-response error across the full (c, D) parameter grid, extending beyond the correlated parameter combinations seen during training. The pretrained FNO performs well near the narrow training support but its error grows substantially as the two controls are varied independently. Rerouting D alone reduces this error even though advection speed c continues to enter through the learned backbone. Jointly rerouting (c, D) performs best across most of the grid, showing that *ReRoute* can recover the effects of multiple correlated controls when their joint entry mechanism is known.

The improvement reflects substantially smaller field-level errors. For the representative out-of-distribution intervention in Fig. 7, the baseline produces a large and spatially coherent response error after 100 rollout steps. Its RMSE is 0.1182. Rerouting D reduces the error to 0.0108, while jointly rerouting (c, D) reduces it to 0.0079, corresponding to reductions of 90.8% and 93.3%, respectively. Thus, the aggregate gains are not produced by cancellation across space: *ReRoute* substantially reduces the local field-level error itself.

The gains persist far beyond the fine-tuning horizon. Fig. 2 tracks counterfactual response error over the full 100-step autoregressive rollout, $10\times$ longer than the 10-step horizon used during optimization. Both *ReRoute* variants maintain substantially lower error throughout the rollout. At step 100, the baseline reaches an average error of 0.0784, compared with 0.0130 ($\downarrow$ 83.4%) and 0.0082 ($\downarrow$ 89.6%) when rerouting D and (c, D), respectively.

This experiment also tests two extensions beyond the formal setting of § A: the pathway depends on the evolving state and the targeted input contains multiple controls jointly. Although the current identification result does not directly cover these cases, the controlled system shows that both extensions work effectively in practice.

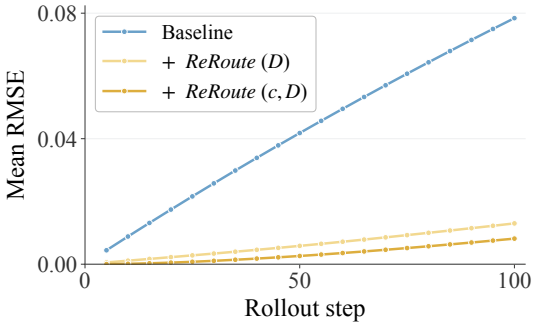

Figure 2: ***ReRoute* remains accurate over long autoregressive rollouts.** Counterfactual-response RMSE over 380 held-out parameter pairs far from the training diagonal, averaged across held-out initial conditions. Fine-tuning uses 10-step rollouts; evaluation extends to 100 steps.

4.2 Climate Emulation

ACE2 (Watt-Meyer et al., 2025) is a family of state-of-the-art six-hour autoregressive climate emulators conditioned on boundary forcings including sea-surface temperature (SST), sea ice, solar forcing, and global-mean CO_2. Despite strong predictive skill, the pretrained ACE2 model exhibits unrealistic sensitivity under unseen CO_2 perturbations: because historical SST and CO_2 co-vary, the model can reproduce the observed climate while attributing changes to the wrong forcing. For example, fixing CO_2 removes too much near-surface warming even when historical SST and sea ice are retained (Watt-Meyer et al., 2025). Recent work mitigates this by adding controlled random-CO_2 simulations (Clark et al., 2026), however, these controlled experiments are never available in real-world, limiting their application in actual climate re-analysis data (Hersbach et al., 2020). *ReRoute* instead uses a partial radiative mechanism and requires no controlled perturbation trajectories for training. We take $\mathbf{s} = CO_2$ as the intervention target and *ReRoute* its effect through a fixed radiative pathway, while conditioning the pretrained backbone on a reference concentration $\mathbf{s}_0$. We validated *ReRoute* in both large scale coupled atmosphere-ocean dynamics and real-world reanalysis data, showing substantial out-of-distribution performance improvements without additional controlled data and very light fine-tuning. Full experimental details are given in § F.

4.2.1 Coupled Climate With Held-Out Counterfactual Evaluation

Experiment Setup. We use the released ACE2S-SHiELD+$_{\text{no-RC}}$ checkpoint as our baseline throughout this experiment. This model is trained on a mixture of AMIP data (Atmospheric Model Intercomparison Project) and equilibrium-climate simulations. AMIP simulations are essentially atmosphere-only climate simulations in which the ocean surface, including sea-surface temperatures, is prescribed rather than dynamically simulated. Crucially, the model receives no random-CO_2 (no-RC) training data in which SST and CO_2 are varied independently. *ReRoute* is applied to the same pretrained checkpoint and fine-tuned with the objective function in Equation (3.4). We report the mean and standard deviation of the per-realization channel-averaged normalized spatial RMSE following Clark et al. (2026), together with the complete training, validation, and evaluation splits (see § F).

Standard climate conditions. *ReRoute preserves, and slightly improves factual climate skill under standard forcing conditions.* We first evaluate the held-out 2012–2020 AMIP factual setting, where SST and CO_2 retain the historical relationship represented in training. In this regime, both models achieve low aggregate error, with *ReRoute* slightly improving the mean per-realization α from 0.0359 to 0.0327. Thus, rerouting the CO_2 input does not require sacrificing the emulator's skill under standard forcing conditions; the adapted model continues to reproduce the factual climate over multi-year autoregressive rollouts.

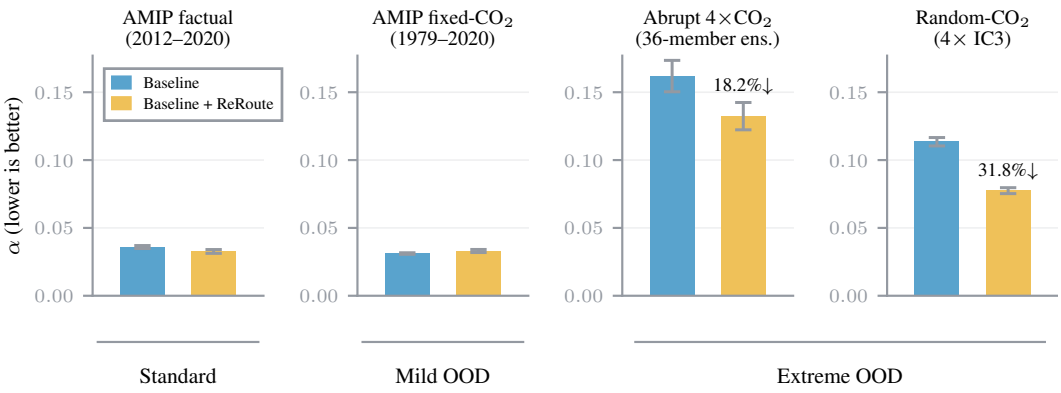

Figure 3: ***ReRoute* becomes increasingly beneficial under severe SST–CO_2 distribution shift.** We compare ACE2S-SHiELD+$_{\mathrm{no-RC}}$ (Baseline) with the same pretrained checkpoint augmented with *ReRoute*. Bars show the mean per-realization aggregate climate error α, with error bars denoting one standard deviation across independent ensembles. Performance remains comparable under standard AMIP and the milder fixed-CO_2 intervention, while *ReRoute* reduces error by 18.2% under abrupt-$4\times CO_2$ and by 31.8% under the held-out random-CO_2 forcing. The abrupt-$4\times CO_2$ experiment uses 36 initial-condition members; the remaining experiments use five stochastic realizations.

Mild distribution shift. *ReRoute remains comparable to the baseline under a mild forcing shift.* We hold CO_2 fixed over the 1979–2020 AMIP rollout while retaining the historical SST evolution, introducing a moderate mismatch between forcings that usually co-vary in training. Here, *ReRoute* obtains 0.0330 across stochastic realizations, close to the baseline's $\alpha = 0.0311$. The small difference shows that the mechanistic pathway leaves performance largely unchanged when the intervention remains relatively close to the range of climate states represented by the original training data. Together with the factual AMIP experiment, this establishes that *ReRoute* retains the pretrained emulator's climate skill under standard and mildly shifted forcing conditions.

Extreme distribution shifts. *The mechanistic pathway becomes increasingly valuable as the forcing relationship moves further out of distribution.* Under abrupt-$4\times CO_2$, CO_2 changes instantaneously while the atmosphere and slab ocean initially remain in the $1\times CO_2$ climate state, creating a forcing–state combination absent from training. Across the 36-member ensemble, *ReRoute* reduces mean per-member aggregate error by 18.2%. The improvement is even larger in the held-out random-CO_2 experiment, where CO_2 varies independently of SST over a wide forcing range: *ReRoute* reduces error by 31.8%, from 0.1135 to 0.0775. This exposes the failure mode targeted by *ReRoute*: when CO_2 departs strongly from the correlations encoded by the observational training distribution, directly conditioning the flexible backbone on CO_2 becomes unreliable, whereas routing the perturbation through the known radiative pathway substantially improves the resulting climate response.

4.2.2 Real-World Climate Reanalysis

Experiment Setup. We next apply *ReRoute* to real-world historical reanalysis, where controlled counterfactual simulations are unavailable. We use ACE2-ERA5 (Watt-Meyer et al., 2025), which is trained directly on ERA5 reanalysis, a reconstruction of historical atmospheric conditions obtained by combining weather observations with numerical weather prediction. In these data, SST and CO_2 co-evolve rather than varying independently, making their individual effects difficult to disentangle. Methods that rely on randomized or otherwise controlled CO_2 trajectories therefore cannot be applied directly. For evaluation, following the fixed-CO_2 diagnostic used in (Watt-Meyer et al., 2025), we rerun the historical experiment with CO_2 fixed at its 1940 value while retaining the other historical boundary conditions. We visualize the change in time-mean near-surface temperature relative to the corresponding historical-CO_2 rollout,

$$\Delta T_{2m} = T_{2m}^{\text{fixed } CO_2} - T_{2m}^{\text{historical } CO_2}. \tag{4.3}$$

Because the corresponding real-world counterfactual is unobserved, we use this experiment as a *qualitative test* of whether the emulator preserves the surface warming encoded by the prescribed boundary conditions.

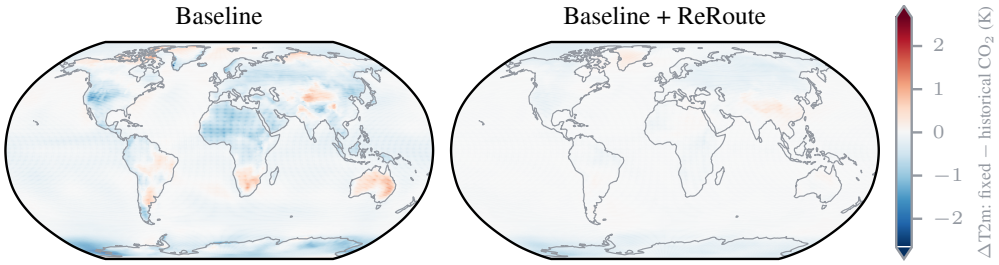

Figure 4: ***ReRoute* preserves surface warming under fixed CO_2 in ACE2-ERA5, i.e. near-zero values in the heatmap.** Shown is the change in time-mean near-surface temperature when CO_2 is held fixed at its 1940 value instead of following the historical trajectory, while retaining the other historical boundary conditions such as SST and sea-ice. The ACE2-ERA5 baseline exhibits widespread cooling, indicating that much of its historical warming is spuriously tied to the CO_2 input. With *ReRoute*, the spurious cooling is substantially corrected over most of the globe, showing that the model retains the surface warming driven by the prescribed boundary conditions.

Results. *ReRoute preserves substantially more surface warming under fixed CO_2 using reanalysis data alone.* As shown in Fig. 4, the ACE2-ERA5 baseline exhibits widespread negative ΔT_{2m} when CO_2 is fixed, indicating that much of the historical surface warming disappears despite retaining the same evolving SST and sea-ice boundary conditions, with the strongest response over land and high latitudes. With *ReRoute*, the fixed- versus historical-CO_2 difference becomes much smaller over most of the globe, so the emulator retains most of the surface warming even when the historical CO_2 trajectory is removed. This setting is particularly important because ERA5 provides no randomized controlled trajectories in which SST and CO_2 are varied independently. *ReRoute* instead uses the known CO_2 pathway to separate the targeted forcing response while learning the remaining climate dynamics entirely from historical reanalysis. *ReRoute* also improves the generated tropical-cyclone climatology: over the held-out 2001–2010 period, the same TempestExtremes tracker detects 49.8 cyclones yr^{-1} in the baseline and 43.6 yr^{-1} with ReRoute, much closer to ERA5's 40.3 yr^{-1} (Fig. 8).

5 Discussion

Predicting how a scientific system responds to an intervention usually requires either intervention data or a complete mechanistic model, both of which are often unavailable in practice. We introduced *ReRoute*, a lightweight framework for targeted scientific what-if queries that combines *observations + mechanism*. *ReRoute* uses partial mechanistic knowledge of how the queried variable enters the system, while observational data identify the downstream interactions, feedbacks, and long-term response. Under explicit assumptions on this first-level pathway, observational variation, and the downstream response class, we show that the targeted interventional distribution is identifiable. By requiring knowledge only of this targeted interface rather than the full governing system, *ReRoute* provides a practical route to mechanism-guided what-if prediction in scientific domains where controlled experiments are costly or impossible and mechanistic knowledge is incomplete.

Across controlled and realistic settings, *ReRoute* consistently improves what-if prediction under distribution shift, with the largest gains under the strongest shifts. In controlled advection–diffusion, it cuts counterfactual response error by up to 89.6% far beyond training support, over rollouts $10\times$ longer than the fine-tuning horizon. Under severe SST–CO_2 shifts in ACE2, it reduces aggregate error by 18.2–31.8% without random-CO_2 supervision, while preserving skill under standard and mild shifts. In an ERA5-trained emulator, it substantially restores near-surface warming under fixed CO_2 and yields a tropical cyclone climatology closer to ERA5 using historical reanalysis alone. Together, these results show that targeted mechanistic knowledge can deliver robust gains across systems and shifts with light fine-tuning, no counterfactual data, and a fraction of counterfactual retraining cost.

Limitations. The current theory is limited to a single targeted variable with a state-independent pathway, although our experiments indicate that *ReRoute* can remain effective in practice with state-dependent pathways and multiple jointly-rerouted variables. The framework also assumes near-exact knowledge of how the targeted variable enters the system, which can be limiting in applications where the mechanism is only approximately known. While *ReRoute* may remain beneficial under modest pathway misspecification, imperfect mechanistic knowledge could limit its effectiveness.

AI USE STATEMENT

Large language models (LLMs) were used to assist with manuscript preparation and experimental presentation, including refining the writing, generating figures, and creating tables. The core research ideas, methodology, experimental design, and scientific conclusions were developed by the authors.

ETHICS STATEMENT

This work uses existing climate reanalysis data, physics-based simulations, and pretrained scientific models; it involves no human subjects or sensitive personal data. *ReRoute* is intended for scientific research and should not be treated as a substitute for established physical climate models or as direct operational or policy guidance. Its predictions may inherit limitations from the underlying data, pretrained emulator, and supplied mechanistic pathway, especially under extreme distribution shifts. We therefore report assumptions, controlled evaluations where available, and known failure modes.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, training and evaluation splits, model checkpoints, pathway construction, fine-tuning procedure, hyperparameters, rollout protocols, ensemble settings, and evaluation metrics in § F. The appendix also specifies the exact experimental configurations used for each climate benchmark and the computation of all reported metrics. Upon acceptance, we will release the curated implementation of *ReRoute*, together with the training and evaluation configurations and scripts required to reproduce the main results and figures.

REFERENCES

Peter Atkins, Julio de Paula, and James Keeler. *Atkins' Physical Chemistry*. Oxford University Press, Oxford, 12th edition, 2022. ISBN 9780198847816. doi:10.1093/hesc/9780198847816.001.0001.

Julius Berner, Miguel Liu-Schiaffini, Jean Kossaifi, Valentin Duruisseaux, Boris Bonev, Kamyar Azizzadenesheli, and Anima Anandkumar. Principled approaches for extending neural architectures to function spaces for operator learning. *Nature Machine Intelligence*, 8:1173–1181, 2026. doi:10.1038/s42256-026-01267-z.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021. doi:10.1103/PhysRevLett.126.098302.

Tom G. Beucler, Stephan Rasp, Michael Pritchard, and Pierre Gentine. Achieving conservation of energy in neural network emulators for climate modeling. In *ICML 2019 Workshop on Climate Change: How Can AI Help?*, 2019. URL https://www.climatechange.ai/papers/icml2019/16.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2806–2823. PMLR, 2023. URL https://proceedings.mlr.press/v202/bonev23a.html.

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale, 2025. URL https://arxiv.org/abs/2507.12144.

Ashesh Chattopadhyay, Y. Qiang Sun, and Pedram Hassanzadeh. Challenges of learning multi-scale dynamics with AI weather models: Implications for stability and one solution, 2023. URL https://arxiv.org/abs/2304.07029.

Spencer K. Clark, Oliver Watt-Meyer, Anna Kwa, Jeremy McGibbon, Brian Henn, W. Andre Perkins, Elynn Wu, Lucas M. Harris, and Christopher S. Bretherton. ACE2-SOM: Coupling an ML atmospheric emulator to a slab ocean and learning the sensitivity of climate to changed CO_2.

Journal of Geophysical Research: Machine Learning and Computation, 2(4):e2024JH000575, 2025. doi:10.1029/2024JH000575.

Spencer K. Clark, Troy Arcomano, James P. C. Duncan, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, Elynn Wu, Lucas M. Harris, Oliver Watt-Meyer, and Christopher S. Bretherton. Disentangling the effects of sea surface temperature and CO_2 in global machine learned weather-climate emulators, 2026. URL https://arxiv.org/abs/2606.07928.

Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic Chemistry*. Oxford University Press, Oxford, 2nd edition, 2012. ISBN 9780199270293. doi:10.1093/hesc/9780199270293.001.0001.

Valentin Duruisseaux, Jean Kossaifi, and Anima Anandkumar. Fourier Neural Operators explained: A practical perspective, 2025. URL https://arxiv.org/abs/2512.01421.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Sareen S. Gropper, Jack L. Smith, and Timothy P. Carr. *Advanced Nutrition and Human Metabolism*. Cengage, 8th edition, 2022. ISBN 9780357449813.

Paula Harder, Alex Hernandez-Garcia, Venkatesh Ramesh, Qidong Yang, Prasanna Sattegeri, Daniela Szwarcman, Campbell Watson, and David Rolnick. Hard-constrained deep learning for climate downscaling. *Journal of Machine Learning Research*, 24(365):1–40, 2023. URL http://jmlr.org/papers/v24/23-0158.html.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL https://miguelhernan.org/whatifbook.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi:https://doi.org/10.1002/qj.3803.

Michael J. Iacono, Jennifer S. Delamere, Eli J. Mlawer, Mark W. Shephard, Shepard A. Clough, and William D. Collins. Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research: Atmospheres*, 113(D13):D13103, 2008. doi:10.1029/2008JD009944.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. ISBN 9781139025751. doi:10.1017/CBO9781139025751.

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. doi:10.1038/s42254-021-00314-5.

W. Larry Kenney, Jack H. Wilmore, and David L. Costill. *Physiology of Sport and Exercise*. Human Kinetics, Champaign, IL, 9th edition, 2025. ISBN 9781718228429.

Jean Kossaifi, Nikola Kovachki, Zongyi Li, David Pitt, Miguel Liu-Schiaffini, Robert J. George, Boris Bonev, Kamyar Azizzadenesheli, Julius Berner, Valentin Duruisseaux, and Anima Anandkumar. A library for learning neural operators. *Journal of Machine Learning Research*, 27(205):1–6, 2026. URL http://jmlr.org/papers/v27/26-0434.html.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL http://jmlr.org/papers/v24/21-1524.html.

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. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM / IMS J. Data Sci.*, 1(3), May 2024. doi:10.1145/3648506. URL https://doi.org/10.1145/3648506.

Ankur Mahesh, William D. Collins, Travis A. O'Brien, Paul B. Goddard, Sinclaire Zebaze, Shashank Subramanian, James P. C. Duncan, Oliver Watt-Meyer, Boris Bonev, Thorsten Kurth, Karthik Kashinath, Michael S. Pritchard, and Da Yang. Examining fast radiatively driven responses using machine-learning weather emulators, 2026. URL https://arxiv.org/abs/2602.16090.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=RFfUUtKYOG.

J. M. Monteiro, J. McGibbon, and R. Caballero. sympl (v. 0.4.0) and climt (v. 0.15.3) – towards a flexible framework for building model hierarchies in python. *Geoscientific Model Development*, 11(9):3781–3794, 2018. doi:10.5194/gmd-11-3781-2018. URL https://gmd.copernicus.org/articles/11/3781/2018/.

Shaowu Pan and Karthik Duraisamy. Data-driven discovery of closure models. *SIAM Journal on Applied Dynamical Systems*, 17(4):2381–2413, 2018. doi:10.1137/18M1177263.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators, 2022. URL https://arxiv.org/abs/2202.11214.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025. doi:10.1038/s41586-024-08252-9.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning, 2020. URL https://arxiv.org/abs/2001.04385.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi:10.1016/j.jcp.2018.10.045.

Georgios Rigas, Aimee S Morgans, and Jonathan F Morrison. Weakly nonlinear modelling of a forced turbulent axisymmetric wake. *Journal of Fluid Mechanics*, 814:570–591, 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi:10.1109/JPROC.2021.3058954.

Jiyuan Tan and Vasilis Syrgkanis. CausalSmith: A formally grounded, self-improving agentic framework for automated research in causal inference, 2026. URL https://arxiv.org/abs/2607.22511.

P. A. Ullrich, C. M. Zarzycki, E. E. McClenny, M. C. Pinheiro, A. M. Stansfield, and K. A. Reed. Tempestextremes v2.1: a community framework for feature detection, tracking, and analysis in large datasets. *Geoscientific Model Development*, 14(8):5023–5048, 2021. doi:10.5194/gmd-14-5023-2021. URL https://gmd.copernicus.org/articles/14/5023/2021/.

Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S. Bretherton. ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8(1): 205, 2025. doi:10.1038/s41612-025-01090-0.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lk2Qk5xjeu.

Jialin Yu, Yuxiang Zhou, Yulan He, Nevin L. Zhang, Junchi Yu, Philip Torr, and Ricardo Silva. Causal fine-tuning of pre-trained language models for robust test time adaptation. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025. URL https://openreview.net/forum?id=RqqC7wwq3X.

Bosong Zhang and Timothy M. Merlis. The equilibrium response of atmospheric machine-learning models to uniform sea surface temperature warming. *Journal of Advances in Modeling Earth Systems*, 18(9):e2026MS005815, 2026. doi:10.1029/2026MS005815.

Appendix

Table of Contents

A A Causal Perspective of ReRoute **14**
- A.1 Preliminaries . . . 14
- A.2 Source-Interface Reduction . . . 16
- A.3 Observational Identification of the Downstream Response . . . 19
- A.4 Admissible Models, Target Determination, and Identification . . . 20
- A.5 Lean Verification . . . 22

B Advection–Diffusion Experiment **23**
- B.1 Dataset and Pretrained FNO Backbone . . . 23
- B.2 State-Dependent Pathway . . . 23
- B.3 Evaluation and Additional Results . . . 24

C Climate Data and Pretrained Models **26**
- C.1 ACE2S-SHiELD+$_{\text{no-RC}}$: Controlled Climate Training Data . . . 26
- C.2 ACE2-ERA5: Historical Reanalysis Training Data . . . 26

D Computational Details **28**

E Climate-Specific Methodology **28**
- E.1 RRTMG CO_2 Pathway Construction . . . 28
- E.2 Tropical Cyclone Tracking Details . . . 29

F Climate Experiment Details and Additional Results **29**
- F.1 Coupled Climate Simulation . . . 29
- F.2 Historical Reanalysis with ACE2-ERA5 . . . 31

A A Causal Perspective of ReRoute

This section gives a causal identification analysis of *ReRoute*. We formalize when a target intervention can be recovered from observational data together with partial mechanistic knowledge, and make explicit the assumptions required for this recovery.

A.1 Preliminaries

We consider an acyclic structural causal model (SCM) with a distinguished *source variable* S and downstream variables $X := (X_1, \dots, X_N)$. Throughout this section, interventions act on S.

Definition A.1 (Structural causal model)**.** An SCM is a tuple $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{F}, P_U)$, where $\mathcal{U} = U_s, U_1, \dots, U_N$, and $\mathcal{V} = S, X_1, \dots, X_N$ are the exogenous and endogenous variables, P_U is the joint distribution of the exogenous variables, and $\mathcal{F} = f_s, f_1, \dots, f_N$ specifies the structural mechanisms. We assume that S is a source node, so

$$S := f_s(U_s),\ X_i \ := f_i\big(X_{\mathbf{pa}_{\mathcal{M}}(i)}, U_i\big), \qquad i = 1, \dots, N. \tag{A.1}$$

The associated acyclic graph $G_{\mathcal{M}}$ contains an edge $V_k \to V_i$ whenever the structural assignment for V_i depends on V_k. We write $\mathbf{pa}_{\mathcal{M}}(i)$ for the parents of X_i.

Because the model is acyclic, its structural equations can be solved recursively. The following parameterized lemma gives the existence, uniqueness, and measurability needed both for the original SCM and for the modified system introduced in Defn. A.3.

Assumption A.1 (Measurability and regularity conventions)**.** Throughout this section we use the following conventions.

1. The source takes values in a measurable space $\mathcal{S}$, and each downstream variable is real valued, so that X takes values in $\mathbb{R}^N$. Thus the downstream state space is standard Borel, and regular conditional distributions of X exist and are unique up to null sets.
2. P_U is a probability measure on the exogenous space. It may be an arbitrary joint law: the coordinates of $U_X := (U_1, \dots, U_N)$ need not be independent, so latent confounding among the downstream variables is allowed.
3. Every structural mechanism is *jointly* measurable in all of its arguments: f_s is measurable, and each f_i is measurable on the product of its parents' value spaces with $\mathcal{U}_i$.
4. Every mapping declared *known* below, including the source-interface response $\Gamma(s_0, \cdot)$ of Asm. A.2, is measurable.

Lemma A.1 (Recursive well-posedness)**.** *Let G be a finite acyclic graph on $1, \dots, N$, let Θ and $\mathcal{V}_1, \dots, \mathcal{V}_N$ be measurable spaces, and for each i let*

$$g_i : \mathbb{R}^{|\mathbf{pa}_G(i)|} \times \mathcal{V}_i \times \Theta \longrightarrow \mathbb{R}$$

be jointly measurable. Then for every $\theta \in \Theta$ and every $v = (v_1, \dots, v_N)$ the system

$$x_i = g_i\big(x_{\mathbf{pa}_G(i)}, v_i, \theta\big), \qquad i = 1, \dots, N,$$

has a unique solution $x = \Psi(\theta, v) \in \mathbb{R}^N$, and the resulting map $\Psi : \Theta \times \prod_i \mathcal{V}_i \to \mathbb{R}^N$ is jointly measurable.

Proof. Choose a topological order of G. At each vertex, the right-hand side depends only on variables earlier in the order, so its value is uniquely determined by values already computed. Recursing through the order gives existence and uniqueness. The same recursion proves joint measurability, since at each step the jointly measurable map g_i is composed with previously constructed jointly measurable maps of (θ, v). □

Apply Lem. A.1 to (A.1), taking G to be the restriction of $G_{\mathcal{M}}$ to the downstream variables and $\Theta = \mathcal{S}$ to carry the source value. Then X is a measurable function of (S, U_X), and hence of (U_s, U_X). We will apply the same lemma in Defn. A.3 with $\Theta = \mathcal{G}$.

Observational Distribution. Sampling $U \sim P_U$ and evaluating the structural equations induces the observational distribution $P^{\mathcal{M}}_{S,X}$ of (S, X). We write $P^{\mathcal{M}}_X$ for its X-marginal.

Definition A.2 (Source intervention)**.** Let S be a source node in the SCM (Defn. A.1). For an intervention value $\bar{s} \in \mathcal{S}$, the intervention $do(S = \bar{s})$ replaces the structural equation f_s by $S := \bar{s}$, while leaving all other structural equations and the exogenous distribution P_U unchanged. Let $\mathcal{M}_{do(S=\bar{s})}$ denote the resulting intervened SCM. The corresponding *interventional distribution* of the downstream variables X is $P^{\mathcal{M};do(S=\bar{s})}_X$.

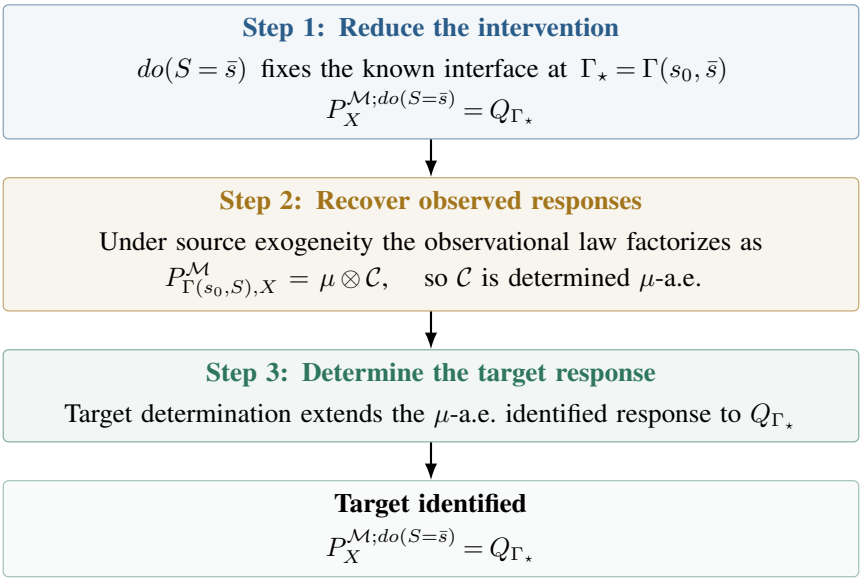

Figure 5: Roadmap of the identification argument.

Overview. Our goal is to identify $P_X^{\mathcal{M};do(S=\bar{s})}$ from the observational distribution $P^{\mathcal{M}}_{S,X}$ and a known *source-interface response* $\Gamma(s_0, s)$, without specifying the other downstream causal mechanisms. The argument has three steps. First, $do(S = \bar{s})$ fixes the known pathway at $\Gamma_\star := \Gamma(s_0, \bar{s})$, so the causal target becomes the downstream response at $\Gamma_\star$. Second, under source exogeneity, observational variation in $\Gamma(s_0, S)$ identifies the downstream response map for μ-almost every observed pathway value, where μ is the law of $\Gamma(s_0, S)$. Third, a target-determination condition converts this μ-almost-everywhere identification into identification at the particular point $\Gamma_\star$, including when the target lies outside the observational distribution. These three steps give identification in the two-model sense of Defn. A.7.

A.2 Source-Interface Reduction

We first reduce an intervention on S to a change at a lower-dimensional interface whose response is known.

Assumption A.2 (Known source interface)**.** Let

$$D := \mathrm{ch}_{G_{\mathcal{M}}}(S) \subseteq \{1, \dots, N\}$$

be the complete set of direct children of S, and fix a reference source value $s_0 \in \mathcal{S}$. For every $j \in D$, assume that the *source-interface response*

$$\Gamma_j(s_0, s),$$

is a known measurable function of $s \in \mathcal{S}$. Collect these responses as

$$\Gamma(s_0, s) := \big(\Gamma_j(s_0, s)\big)_{j \in D} \in \mathcal{G} := \mathbb{R}^{|D|} \tag{A.2}$$

with $\Gamma(s_0, s_0) = 0$.

The reference value s_0 enters through the notation $\Gamma(s_0, s)$ and through the decomposition of the child mechanisms in Asm. A.3.

Assumption A.3 (Source-additive child mechanism)**.** For every direct child $j \in D$, there exists an unknown jointly measurable mechanism r_j such that

$$X_j = r_j\left(X_{\mathbf{pa}_{\mathcal{M}}(j)\setminus S}, U_j\right) + \Gamma_j(s_0, S). \tag{A.3}$$

Equivalently, (A.3) says that for all $s, s' \in \mathcal{S}$, all parent configurations x, and all disturbances u,

$$f_j(x, s, u) - f_j(x, s', u) = \Gamma_j(s_0, s) - \Gamma_j(s_0, s'). \tag{A.4}$$

Thus the absolute contribution of S at the reference value s_0 need not be known; it is absorbed into r_j. No additive structure is imposed on the remaining downstream mechanisms or on how the other parents and U_j enter r_j.

The assumption does require the source contribution to be homogeneous at the interface. By (A.4), changing S changes X_j by the same amount for every configuration of the other parents and every value of U_j. Thus the direct source contribution cannot be modified by those variables, although effect modification can still occur downstream.

Remark A.1 (Intuition on Γ). By (A.4), $\Gamma_j(s_0, s)$ is the *direct source contribution* at X_j: the unit-level change in X_j caused by moving S from s_0 to s while holding the other parents of X_j and the disturbance U_j fixed. It is therefore a structural contrast local to the edge $S \to X_j$. Asm. A.3 requires this contrast to be constant across units and parent configurations.

This direct contribution is not generally the marginal interventional contrast at X_j. From (A.3), assuming the expectations below exist,

$$\begin{aligned}\mathbb{E}\big[X_j \mid do(S = s)\big] - \mathbb{E}\big[X_j \mid do(S = s_0)\big] = \big(\Gamma_j(s_0, s) - \Gamma_j(s_0, s_0)\big) \\ + \mathbb{E}\big[r_j\big(X^{do(S=s)}_{\mathbf{pa}(j)\setminus\{S\}}, U_j\big)\big] \\ - \mathbb{E}\big[r_j\big(X^{do(S=s_0)}_{\mathbf{pa}(j)\setminus\{S\}}, U_j\big)\big],\end{aligned} \tag{A.5}$$

whose first bracket is $\Gamma_j(s_0, s)$ under the normalization of Asm. A.2. The final two terms need not cancel. For example, another parent of X_j may itself be a descendant of S, such as a second child $k \in D$ with an edge $k \to j$. The marginal contrast at X_j then contains both the direct contribution and an indirect contribution through k. Only the direct contribution is assumed known.

The direct and marginal contrasts coincide under the following additional condition:

$$\textit{(ND)} \qquad \text{for every } j \in D, \text{ no parent of } X_j \text{ other than } S \text{ is a descendant of } S. \tag{A.6}$$

Under (A.6), $X_{\mathbf{pa}(j)\setminus S}$ is unaffected by the intervention, so the final two terms of (A.5) cancel and $\Gamma_j(s_0, s)$ can be obtained from a single-source experiment. Without (A.6), Γ must instead come from mechanistic knowledge of the edge $S \to X_j$. None of the results below require (A.6).

Definition A.3 (Forced downstream system)**.** Fix the reference value s_0 from Asm. A.2, and let

$$U_X := (U_1, \ldots, U_N),$$

denote the X-associated exogenous variables, distributed as P_{U_X}. For a pathway value $\eta = (\eta_j)_{j\in D} \in \mathcal{G}$, define the *forced downstream system* by replacing $\Gamma_j(s_0, S)$ with the constant η_j in every direct-child equation (A.3), while leaving all other structural equations unchanged:

$$x_j = r_j\big(x_{\mathbf{pa}_{\mathcal{M}}(j)\setminus\{S\}}, u_j\big) + \eta_j \quad (j \in D), \qquad x_i = f_i\big(x_{\mathbf{pa}_{\mathcal{M}}(i)}, u_i\big) \quad (i \notin D). \tag{A.7}$$

By Lem. A.1, applied with parameter space $\Theta = \mathcal{G}$, the system (A.7) has a unique solution for every $\eta \in \mathcal{G}$ and every u. The solution map

$$\Phi : \mathcal{G} \times \prod_i \mathcal{U}_i \longrightarrow \mathbb{R}^N, \qquad x = \Phi(\eta, u), \tag{A.8}$$

is jointly measurable. With the model's own exogenous variables we write $X = \Phi(\eta, U_X)$.

The value η may be any point in $\mathcal{G}$, not only a value of the form $\Gamma(s_0, s)$. This lets us define the response at the target $\Gamma_\star$ independently of which pathway values occur observationally.

Neither Φ nor the downstream causal graph is assumed known. Defn. A.3 only states that Φ exists and is measurable. For $i \notin D$, the equations in (A.7) do not involve S because $D = \mathrm{ch}_{G_\mathcal{M}}(S)$ contains all direct children of S.

Definition A.4 (Downstream response map)**.** For each pathway value $\eta \in \mathcal{G}$, let Q_η denote the distribution of

$$X = \Phi(\eta, U_X), \qquad U_X \sim P_{U_X}, \tag{A.9}$$

that is, $Q_\eta := \big(\Phi(\eta, \cdot)\big)_{\#} P_{U_X}$. The *downstream response map* is

$$\mathcal{C} : \eta \longmapsto Q_\eta. \tag{A.10}$$

Since Φ is jointly measurable and P_{U_X} is a probability measure, $\mathcal{C}$ is a Markov kernel from $\mathcal{G}$ to $\mathbb{R}^N$.

Thus $\mathcal{C}$ describes how the downstream system responds to changes in the known pathway, without requiring the downstream mechanisms themselves to be known.

We now reduce the intervention on S to evaluation of this response map. The next lemma gives the corresponding unit-level representation.

Lemma A.2 (Interface representation)**.** *Under Asms. A.2 and A.3, the following hold pointwise on the underlying probability space.*

1. *In the observational model,* $X = \Phi\big(\Gamma(s_0, S), U_X\big)$.
2. *In the intervened model* $\mathcal{M}_{do(S=\bar{s})}$, $X = \Phi\big(\Gamma(s_0, \bar{s}), U_X\big)$.

Proof. Let σ denote the value of S: observationally $\sigma = f_s(U_s)$, while under $do(S = \bar{s})$ it equals the constant $\bar{s}$.

Because $D = \mathrm{ch}_{G_\mathcal{M}}(S)$ contains every direct child of S, no equation indexed by $i \notin D$ involves S. These equations therefore remain $x_i = f_i(x_{\mathbf{pa}(i)}, u_i)$. For $j \in D$, Asm. A.3 gives $x_j = r_j(x_{\mathbf{pa}(j)\setminus S}, u_j) + \Gamma_j(s_0, \sigma)$. Hence X solves exactly (A.7) with $\eta = \Gamma(s_0, \sigma)$. By uniqueness in Lem. A.1, the solution is $\Phi(\Gamma(s_0, \sigma), U_X)$. □

Proposition A.1 (Intervention as pathway evaluation)**.** *For an intervention* $do(S = \bar{s})$*, define the target pathway response*

$$\Gamma_\star := \Gamma(s_0, \bar{s}). \tag{A.11}$$

Under Asms. A.2 and A.3,

$$P_X^{\mathcal{M};do(S=\bar{s})} = Q_{\Gamma_\star}. \tag{A.12}$$

Proof. The intervention $do(S = \bar{s})$ changes only the structural equation for S. It leaves the exogenous distribution unchanged, so U_X still has law P_{U_X}. By Lem. A.2(ii),

$$X = \Phi(\Gamma_\star, U_X).$$

Therefore

$$P_X^{\mathcal{M};do(S=\bar{s})} = \mathrm{Law}\big(\Phi(\Gamma_\star, U_X)\big) = \big(\Phi(\Gamma_\star, \cdot)\big)_{\#} P_{U_X} = Q_{\Gamma_\star},$$

where the final equality is Defn. A.4. □

Thus the original causal problem reduces to identifying the downstream response $Q_{\Gamma_\star}$ from observational data.

A.3 Observational Identification of the Downstream Response

We now ask what observational variation in S reveals about the response map $\eta \mapsto Q_\eta$. The key condition is that the source varies independently of the disturbances governing the downstream system.

Definition A.5 (Observed pathway distribution and support)**.** The observational pathway value is $\Gamma(s_0, S)$. Its distribution and support are

$$\mu := P^{\mathcal{M}}_{\Gamma(s_0,S)} = \big(\Gamma(s_0,\cdot)\big)_{\#} P^{\mathcal{M}}_S, \qquad \mathcal{G}_{\text{obs}} := \text{supp}(\mu), \tag{A.13}$$

where $P^{\mathcal{M}}_S$ is the S-marginal of the observational distribution $P^{\mathcal{M}}_{S,X}$. The map $\Gamma(s_0,\cdot)$ is measurable by Asm. A.2, so μ is well defined.

The set $\mathcal{G}_{\text{obs}}$ describes the pathway values generated by observed values of S. It is used only to describe the regimes in Remark A.2; no result below depends on it. The target $\Gamma_\star = \Gamma(s_0, \bar{s})$ need not belong to $\mathcal{G}_{\text{obs}}$.

Assumption A.4 (Source exogeneity)**.** The source disturbance is independent of all downstream disturbances:

$$U_s \perp U_X. \tag{A.14}$$

Since $S = f_s(U_s)$ and $\Gamma(s_0,\cdot)$ is measurable, Asm. A.4 implies

$$S \perp U_X, \qquad \Gamma(s_0, S) \perp U_X. \tag{A.15}$$

Thus changes in the observed pathway do not change the distribution of the downstream disturbances. Observational variation in the pathway therefore reveals the response of the same downstream system.

Lemma A.3 (Observational recovery of the downstream response)**.** *Under Asms. A.2 to A.4, the observational joint law of the pathway value and the downstream variables factorizes through the response map:*

$$P^{\mathcal{M}}_{\Gamma(s_0,S),,X} = \mu \otimes \mathcal{C}, \tag{A.16}$$

where $\mu \otimes \mathcal{C}$ denotes the law on $\mathcal{G} \times \mathbb{R}^N$ with first marginal μ and conditional kernel $\mathcal{C}$, i.e. $(\mu \otimes \mathcal{C})(A \times B) = \int_A Q_\eta(B), \mu(\mathrm{d}\eta)$. Equivalently, any regular conditional distribution of X given $\Gamma(s_0, S)$ satisfies

$$P^{\mathcal{M}}_{X|\Gamma(s_0,S)}(,\cdot \mid \eta) = Q_\eta \qquad \textit{for } \mu\textit{-almost every } \eta \in \mathcal{G}. \tag{A.17}$$

Hence the observational distribution $P^{\mathcal{M}}_{S,X}$ determines the downstream response map $\eta \mapsto Q_\eta$ up to μ-null sets, and only up to μ-null sets.

Proof. By Lem. A.2(i),

$$X = \Phi\big(\Gamma(s_0, S), U_X\big). \tag{A.18}$$

By (A.15), $\big(\Gamma(s_0,S), U_X\big)$ has product law $\mu \times P_{U_X}$. Push this law forward by $(\eta, u) \mapsto \big(\eta, \Phi(\eta, u)\big)$. Using (A.18),

$$P^{\mathcal{M}}_{\Gamma(s_0,S),\,X} = \big((\eta,u) \mapsto (\eta, \Phi(\eta,u))\big)_{\#}\big(\mu \times P_{U_X}\big) = \mu \otimes \mathcal{C},$$

because at each fixed η, $\big(\Phi(\eta,\cdot)\big)_{\#} P_{U_X} = Q_\eta$. This proves (A.16).

Since X takes values in the standard Borel space $\mathbb{R}^N$ (Asm. A.1), disintegration over the first marginal μ is unique up to μ-null sets. Hence (A.16) is equivalent to (A.17). □

Thus observational data determine $\eta \mapsto Q_\eta$ for μ-almost every pathway value, but not at arbitrary individual points.

The remaining question is whether this almost-everywhere information determines the response at the target $\Gamma_\star$. This issue arises even when $\Gamma_\star \in \mathcal{G}_{\text{obs}}$: if $\mu(\Gamma_\star) = 0$, an almost-everywhere statement alone does not fix the value at that point. We address this with a target-determination assumption on the admissible class of response maps.

A.4 Admissible Models, Target Determination, and Identification

Assumption A.5 (Target determination)**.** Let $\mathfrak{C}$ denote the admissible class of downstream response maps, let μ be the observed pathway distribution of Defn. A.5, and let $\Gamma_\star = \Gamma(s_0, \bar{s})$ be the target pathway value. For any two maps $\mathcal{C}, \widetilde{\mathcal{C}} \in \mathfrak{C}$,

$$Q_\eta = \widetilde{Q}_\eta \quad \text{for } \mu\text{-almost every } \eta \qquad \Longrightarrow \qquad Q_{\Gamma_\star} = \widetilde{Q}_{\Gamma_\star}. \tag{A.19}$$

Asm. A.5 is a condition on the triple $(\mathfrak{C}, \mu, \Gamma_\star)$, not on $\mathfrak{C}$ alone. In the identification theorem below, the two models being compared have the same pathway law μ, so the condition is evaluated at a common μ.

Definition A.6 (Admissible model)**.** Fix the interface data $\big(D, \Gamma(s_0, \cdot)\big)$ and a class $\mathfrak{C}$ of downstream response maps. An SCM $\mathcal{M}$ is *admissible* if it satisfies Defn. A.1 and Asm. A.1, satisfies Asms. A.2 to A.4 with that same D and that same $\Gamma(s_0, \cdot)$, and its downstream response map $\mathcal{C}^{\mathcal{M}}$ of Defn. A.4 belongs to $\mathfrak{C}$. We write $\mathfrak{A} = \mathfrak{A}\big(D, \Gamma(s_0, \cdot), \mathfrak{C}\big)$ for the set of admissible models.

Admissible models share only the specified interface $\big(D, \Gamma(s_0, \cdot)\big)$ and the requirement that their response maps belong to $\mathfrak{C}$. They may otherwise have different exogenous spaces, exogenous distributions, downstream graphs, and downstream mechanisms.

Definition A.7 (Identifiability of the target)**.** The target interventional distribution $P_X^{\cdot,;do(S=\bar{s})}$ is *identifiable from* $\big(P_{S,X}, \Gamma(s_0, \cdot)\big)$ *within* $\mathfrak{A}$ if for every pair $\mathcal{M}, \widetilde{\mathcal{M}} \in \mathfrak{A}$,

$$P_{S,X}^{\mathcal{M}} = P_{S,X}^{\widetilde{\mathcal{M}}} \qquad \Longrightarrow \qquad P_X^{\mathcal{M};do(S=\bar{s})} = P_X^{\widetilde{\mathcal{M}};do(S=\bar{s})}.$$

We can now combine the three steps.

Theorem A.1 (Identification of the target interventional distribution)**.** *Let $\mathfrak{C}$ be a class of downstream response maps, let $\bar{s} \in \mathcal{S}$ be an intervention value, and let $\mathfrak{A}$ be the set of models admissible for $\big(D, \Gamma(s_0, \cdot)\big)$ and $\mathfrak{C}$ (Defn. A.6). Let $P_{S,X}$ be an observational distribution, and let $\mu = \big(\Gamma(s_0, \cdot)\big)_{\#} P_S$ be the pathway law it induces. If Asm. A.5 holds for the triple $(\mathfrak{C}, \mu, \Gamma_\star)$, with $\Gamma_\star = \Gamma(s_0, \bar{s})$, then any two models in $\mathfrak{A}$ whose observational distribution equals $P_{S,X}$ induce the same target interventional distribution.*

If this holds for every observational distribution realizable within $\mathfrak{A}$, then the target is identifiable from $\big(P_{S,X}, \Gamma(s_0, \cdot)\big)$ within $\mathfrak{A}$ in the sense of Defn. A.7.

Proof. Let $\mathcal{M}, \widetilde{\mathcal{M}} \in \mathfrak{A}$ have the same observational distribution $P_{S,X}$, and let $\mathcal{C}, \widetilde{\mathcal{C}} \in \mathfrak{C}$ be their downstream response maps.

Step 1: the two models induce the same joint law of $\big(\Gamma(s_0, S), X\big)$. Because both models use the same measurable interface $\Gamma(s_0, \cdot)$, pushing their common observational distribution $P_{S,X}$ through $(s, x) \mapsto \big(\Gamma(s_0, s), x\big)$ gives

$$P^{\mathcal{M}}_{\Gamma(s_0,S),,X} = P^{\widetilde{\mathcal{M}}}_{\Gamma(s_0,S),,X} =: \nu. \tag{A.20}$$

Taking first marginals shows that both models also have the same pathway law, $\mu^{\mathcal{M}} = \mu^{\widetilde{\mathcal{M}}} =: \mu$. This is the μ used in Asm. A.5.

Step 2: the two response maps agree μ-almost everywhere. Apply Lem. A.3 to each model. Together with (A.20), this gives

$$\mu \otimes \mathcal{C} = \nu = \mu \otimes \widetilde{\mathcal{C}}.$$

Because X takes values in the standard Borel space $\mathbb{R}^N$, the disintegration of ν over μ is unique up to μ-null sets. Therefore

$$Q_\eta = \widetilde{Q}_\eta \qquad \text{for } \mu\text{-almost every } \eta. \tag{A.21}$$

Step 3: the target response is determined. Since $\mathcal{C}, \widetilde{\mathcal{C}} \in \mathfrak{C}$, Asm. A.5 and (A.21) imply

$$Q_{\Gamma_\star} = \widetilde{Q}_{\Gamma_\star}. \tag{A.22}$$

Conclusion. By Prop. A.1,

$$P_X^{\mathcal{M};do(S=\bar{s})} = Q_{\Gamma_\star}, \qquad P_X^{\widetilde{\mathcal{M}};do(S=\bar{s})} = \widetilde{Q}_{\Gamma_\star}. \tag{A.23}$$

Combining this with (A.22) shows that the two models have the same target interventional distribution.

If the target-determination condition holds for the pathway law induced by every realizable $P_{S,X}$, then every pair of admissible models with the same observational distribution agrees on the target. This is exactly the identifiability condition in Defn. A.7. □

Where the extrapolation content sits. Steps 1 and 2 determine the response map only up to μ-null sets. The passage from this almost-everywhere information to the value at the single target point $\Gamma_\star$ comes entirely from Asm. A.5. There are three relevant regimes.

Proposition A.2 (Positive-mass targets)**.** *If $\mu(\Gamma_\star) > 0$, then Asm. A.5 holds at $\Gamma_\star$ for* every *class $\mathfrak{C}$.*

Proof. Suppose $Q_\eta = \widetilde{Q}_\eta$ for μ-almost every η. Then the disagreement set $A := \eta : Q_\eta \neq \widetilde{Q}_\eta$ is contained in a μ-null measurable set N. If $Q_{\Gamma_\star} \neq \widetilde{Q}_{\Gamma_\star}$, then $\Gamma_\star \in A \subseteq N$, so $\mu(\Gamma_\star) \leq \mu(N) = 0$, contradicting the assumption. □

Remark A.2 (Three regimes). We consider the three following regimes:

1. *Positive-mass target.* If $\mu(\Gamma_\star) > 0$, then Prop. A.2 makes Asm. A.5 automatic for every $\mathfrak{C}$. The target itself occurs with positive observational probability.

2. *In-support, μ-null target.* If $\Gamma_\star \in \mathcal{G}_{\text{obs}}$ but $\mu(\{\Gamma_\star\}) = 0$, as is typical when μ is nonatomic, then (A.21) does not determine $Q_{\Gamma_\star}$ by itself. A regularity restriction on $\mathfrak{C}$ is needed.

   Continuity is sufficient. Suppose every $\eta \mapsto Q_\eta$ in $\mathfrak{C}$ is continuous under any Hausdorff topology on laws over $\mathbb{R}^N$ that separates distinct laws, such as the topology of weak convergence. If Q and $\widetilde{Q}$ differed at $\Gamma_\star$, disjoint neighborhoods of their two values would pull back to an *open* neighborhood of $\Gamma_\star$ on which $Q \neq \widetilde{Q}$. Because $\Gamma_\star \in \text{supp}(\mu)$, this neighborhood has positive μ-measure, contradicting (A.21). Other restrictions on $\mathfrak{C}$ may also give target determination.

3. *Out-of-support target.* If $\Gamma_\star \notin \mathcal{G}_{\text{obs}}$, then Asm. A.5 is an extrapolation assumption. The observational distribution alone cannot verify it; it must follow from the model class $\mathfrak{C}$ or from additional mechanistic structure.

Connection to learning. The identification result gives a direct learning problem. Observational samples provide pairs

$$\big(\Gamma(s_0, S_r), X_r\big),$$

from which a model learns the downstream response $\eta \mapsto Q_\eta$. The counterfactual prediction is then obtained by evaluating this response at $\Gamma_\star = \Gamma(s_0, \bar{s})$. These observations are ordinary factual pairs, not unit-level counterfactual pairs. By Lem. A.3, they determine the response map only μ-almost everywhere. Generalization to $\Gamma_\star$ therefore comes from the structural restrictions encoded by the admissible class $\mathfrak{C}$, exactly as formalized by Asm. A.5 and the three regimes in Remark A.2.

A.5 Lean Verification

We formalized and machine-checked the core identification argument in Lean 4 using `mathlib` and *Causalean*, the causal-inference library distributed with the CausalSmith project (Tan & Syrgkanis, 2026). The development comprises roughly nine hundred lines of Lean across four modules. It covers the structural reduction from an intervention on the source variable to a forced pathway value, observational recovery of the downstream response under source exogeneity, and the final target-identification theorem.

Correspondence with the mathematical development. The Lean development follows the mathematical structure above, but the relevant objects are constructed explicitly rather than introduced only through notation. In particular, the recursive downstream solution, the forced system, and the induced response kernels are all defined as Lean objects and their required measurability properties are proved.

Mathematical object or result	Lean counterpart
SCM structure and realized probability model	`Mechanisms`, `Realized`
Recursive solution and Lem. A.1	`Mechanisms.solveOffset`, `solveOffset_spec`, `eq_solveOffset`, `measurable_solveOffset`
Known source interface, Asm. A.2	`Interface`
Source-additive structure, Asm. A.3	encoded in `Mechanisms.solve_spec`
Source intervention, Defn. A.2	`Realized.doSource`, `Realized.doLaw`
Forced downstream system, Defn. A.3	`Mechanisms.forced`
Downstream response map, Defn. A.4	`Realized.response`
Interface representation, Lem. A.2	`Mechanisms.solve_eq_forced_pathway`
Intervention reduction, Prop. A.1	`Realized.doLaw_eq_response_pathway`
Source exogeneity, Asm. A.4	`Realized.exog`
Observational recovery, Lem. A.3	`Realized.pathwayJoint_eq_compProd`, `Realized.condDistrib_eq_response`
Target determination, Asm. A.5	`TargetDetermines`
Identification theorem, Thm. A.1	`reroute_identification`
Positive-mass target, Prop. A.2	`targetDetermines_of_measure_singleton_ne_zero`
Continuous in-support target, Remark A.2(ii)	`targetDetermines_of_continuous_of_mem_support`

Role of the formalization. The structural part of the argument was implemented directly. In particular, the forced downstream system is defined recursively, with existence, uniqueness, and measurability established explicitly. This is important because replacing the source contribution by a pathway value does not by itself define a measurable downstream map. The resulting construction makes Prop. A.1 a theorem about the SCM rather than a notational substitution.

For the measure-theoretic steps, we reuse existing results from `mathlib` and Causalean. Under source exogeneity, the joint law of the observed pathway value and downstream state factorizes through the response kernel, yielding Lem. A.3. The main identification proof then uses almost-everywhere uniqueness of conditional kernels to show that two observationally equivalent admissible models have downstream response maps that agree μ-almost everywhere. The target-determination assumption upgrades this agreement to the particular target value $\Gamma_\star$, after which Prop. A.1 gives equality of the target interventional distributions.

B ADVECTION–DIFFUSION EXPERIMENT

This section includes additional details regarding the controlled Advection–Diffusion experiments.

B.1 DATASET AND PRETRAINED FNO BACKBONE

We use a two-dimensional advection–diffusion system to demonstrate that *ReRoute* extends beyond the climate setting. This isolates two generalizations: the pathway depends on the evolving state, and the targeted input can contain either two controls jointly. Unlike the climate application, both the exact dynamics and exact pathway are available here. So, this experiment tests the *ReRoute* construction directly against known ground truth.

Dataset. We consider a scalar field $u(x, y, t)$ on the periodic domain $[0, 2\pi)^2$ following

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = D\nabla^2 u,$$

where c is the advection speed in the x direction and D is the diffusivity.

We discretize the field on a 64×64 grid and evaluate the exact Fourier solution every $h = 0.01$ time units, retaining 101 states corresponding to 100 transitions. Initial conditions are smooth, zero-mean, unit-RMS Gaussian random fields containing modes through wavenumber six.

The training distribution contains strongly correlated combinations of c and D. For each initial condition, we draw 16 stratified values $q \in [-1, 1]$ and independent noise $\epsilon \sim \mathcal{U}(-0.15, 0.15)$. We then set $c = 1 + 0.5q$ and $D = 0.010 + 0.008\ \mathrm{clip}(q + \epsilon, -1, 1)$. Each initial condition receives a newly sampled set of parameter pairs, producing 512 unique training trajectories across 32 initial conditions, with parameter correlation $r = 0.989$. The validation and test sets each contain eight independent initial conditions. The sampled training parameters are visualized in Fig. 6.

Pretraining. We pretrain a Fourier neural operator (FNO) (Li et al., 2021) from scratch on these correlated trajectories and use it as the baseline throughout the experiment. The FNO is a prominent example of a neural operator, a class of models that learn mappings between function spaces and are therefore particularly well suited for approximating PDE solution operators (Kovachki et al., 2023; Berner et al., 2026). The FNO architecture composes integral operator layers parameterized in Fourier space (Duruisseaux et al., 2025), enabling efficient global interactions through fast Fourier transforms on uniform grids. We use the standard FNO implementation provided in the `neuraloperator` library (Kossaifi et al., 2026).

In our experiments, the FNO receives the current scalar field u_t, together with c and D, each repeated across the spatial grid as an additional input channel, and predicts a residual update around the current state. The model has four FNO blocks, 64 hidden channels, and 16 Fourier modes along each spatial dimension. We pretrain the backbone for 30 epochs using a 10-step autoregressive factual objective. We evaluate the models over 100-step rollouts, ten times the optimization horizon. On the 128 held-out in-distribution test trajectories, the pretrained FNO has mean factual RMSE 0.00184 after 10 steps and 0.01448 after 100 steps.

B.2 STATE-DEPENDENT PATHWAY

Exact state-dependent pathway. For this system, the exact one-step update is known. Advection translates the field by a distance ch in the x direction, while diffusion progressively smooths its smaller-scale structure. In Fourier space, these effects correspond to a phase shift and wavenumber-dependent damping:

$$\mathcal{F}[\Phi^h_{c,D}(u_t)](\mathbf{k}) = \exp\left[-hD|\mathbf{k}|^2 - ihck_x\right]\ \mathcal{F}[u_t](\mathbf{k}),$$

where $\Phi^h_{c,D}(u_t)$ denotes the field given by advancing u_t by one interval h with parameters c and D, and $\mathcal{F}$ denotes the Fourier transform.

We choose $c_0 = 1$ and $D_0 = 0.010$, the centers of their respective training ranges, as the reference values and define the pathway as

$$\Gamma(u_t; c, D) = \Phi^h_{c,D}(u_t) - \Phi^h_{c_0,D_0}(u_t).$$

The pathway therefore gives the exact change in the next state caused by moving from the reference parameters to the requested parameters. Note that this is state-dependent, as the change is recomputed from the current predicted field at every rollout step. When rerouting both controls, we clamp both c and D to their reference values in the backbone and introduce their joint effect through the pathway:

$$\widehat{u}_{t+1} = G_\theta(u_t, c_0, D_0) + \Gamma(u_t; c, D).$$

We also evaluate a one-variable version that fixes and reroutes only D, while the backbone continues to receive the requested value of c:

$$\widehat{u}_{t+1} = G_\theta(u_t, c, D_0) + \Phi^h_{c,D}(u_t) - \Phi^h_{c,D_0}(u_t).$$

Since the exact update is available, the pathway itself introduces no approximation error when rerouting either a single control or the joint target (c, D), allowing us to isolate whether the fine-tuned backbone uses this pathway effectively.

B.3 Evaluation and Additional Results

We evaluate the models on a 41×41 grid spanning $c \in [0.5, 1.5]$ and $D \in [0.002, 0.018]$, breaking the strong dependence present during training. For each of the eight held-out initial conditions, the exact solution gives trajectories under both the reference and intervention parameters. For the rollout curve, we average over the 380 grid points satisfying $|\tilde{c} - \tilde{D}|/\sqrt{2} \geq 0.75$, where $\tilde{c} = (c-1)/0.5$ and $\tilde{D} = (D - 0.010)/0.008$ are normalized coordinates. No trajectories from this grid are used for training or fine-tuning.

Our primary metric is the counterfactual-response RMSE,

$$\mathcal{E}_{\mathrm{cf}}(t; c, D) = \mathrm{RMSE}\left[\left(\widehat{u}_t^{c,D} - \widehat{u}_t^{c_0,D_0}\right) - \left(u_t^{c,D} - u_t^{c_0,D_0}\right)\right].$$

This removes errors shared by the factual and counterfactual trajectories to directly measures whether each model predicts the change caused by the intervention.

Results. Fig. 6 evaluates whether the models can separate c and D after observing them only in strongly correlated pairs. The baseline error increases away from the narrow training support. Rerouting D reduces this error even though c continues to enter through the backbone, while jointly rerouting (c, D) gives the lowest error across most of the grid.

Fig. 7 shows the spatial response residual for a representative intervention far from the training distribution. The baseline produces large, spatially coherent errors. Some residual structure remains under both *ReRoute* variants, but its magnitude is more than an order of magnitude smaller. So, the aggregate improvements comes from substantially weaker field-level errors rather than cancellation across space.

Finally, Fig. 2 follows the three methods over a 100-step autoregressive rollout. Both *ReRoute* variants maintain substantially lower error at every evaluated step, with error growing most slowly when (c, D) are rerouted jointly. The improvement therefore extends well beyond the ten-step horizon used during fine-tuning.

The state-dependent pathways and joint rerouting of multiple controls lie outside the current formalization of *ReRoute*, so the theoretical guarantees do not directly hold here. Nevertheless, both extensions work in practice here, and jointly rerouting (c, D) performs best because c translates spatial structure while D damps its smaller scales. Rerouting only D still leaves the baseline to extrapolate the advection response from strongly correlated training data.

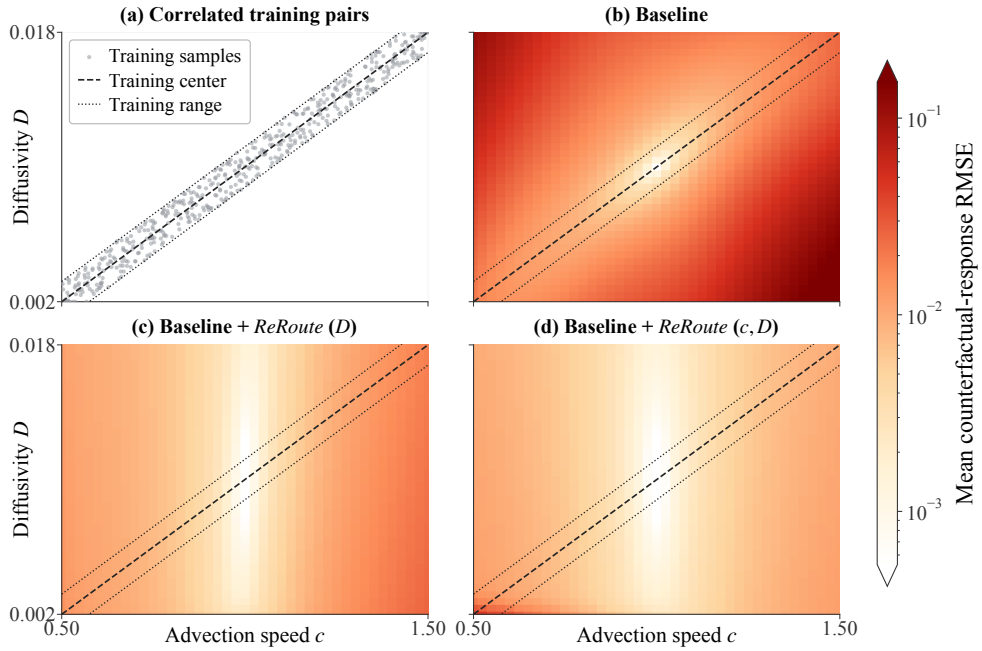

Figure 6: ***ReRoute* separates strongly correlated advection and diffusion controls across the full parameter grid.** The left panel shows the 512 parameter pairs used for training; dashed and dotted lines denote their center and range. The remaining panels show mean counterfactual-response RMSE after 100 rollout steps for the pretrained baseline, *ReRoute* applied only to diffusivity D, and *ReRoute* applied jointly to (c, D). Errors are averaged across eight held-out initial conditions. *ReRoute* reduces error away from the training support, with the joint pathway giving the lowest errors.

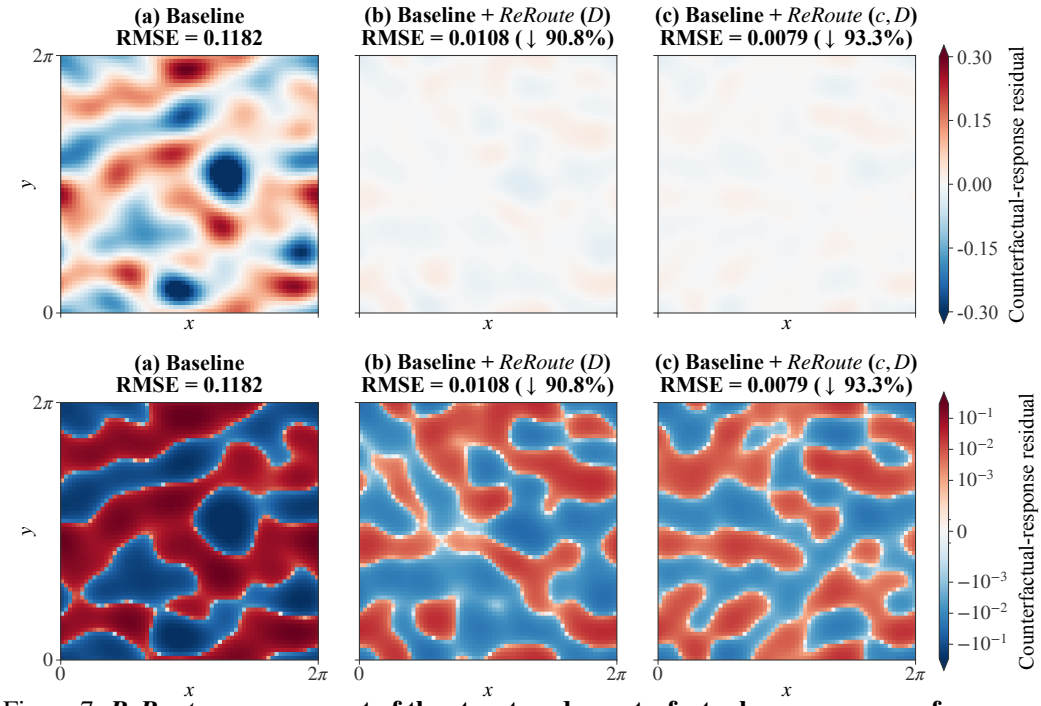

Figure 7: ***ReRoute* removes most of the structured counterfactual-response error for a representative held-out field.** Shown is the signed response residual (**in linear and log scales**) after 100 rollout steps for one held-out initial condition at $c = 0.5$ and $D = 0.018$, a parameter combination far from the correlated training support. The baseline has RMSE 0.1182. Rerouting D alone reduces this to 0.0108, while jointly rerouting (c, D) reduces it to 0.0079.

C Climate Data and Pretrained Models

We use two pretrained ACE2-family climate emulators in the same order in § 4. The controlled climate benchmark uses ACE2S-SHiELD+$_{\text{no-RC}}$ (Clark et al., 2026), trained on physics-based SHiELD simulations without random-CO_2 data. The reanalysis experiment uses ACE2-ERA5 (Watt-Meyer et al., 2025), trained directly on historical ERA5 reanalysis. Both models evolve the atmosphere autoregressively at six-hour resolution on a 1° Gaussian grid with eight atmospheric layers. This section summarizes the training data and variable interface of the two pretrained backbones; intervention and rollout protocols are given in § F.

C.1 ACE2S-SHiELD+$_{\text{no-RC}}$: Controlled Climate Training Data

For the controlled climate experiments, we use the released ACE2S-SHiELD+$_{\text{no-RC}}$ checkpoint from Clark et al. (2026). The underlying atmospheric simulations are generated with the GFDL SHiELD model at C96 resolution (approximately 100 km) with 79 finite-volume hybrid sigma–pressure levels. SHiELD output is regridded to a 1° Gaussian grid, spherical-harmonic round-trip filtered, and vertically coarsened to eight atmospheric layers before being used by ACE2S.

Crucially, the no-RC checkpoint is trained with *no random-CO_2 simulations*. Its 90-year training set contains 45 years of AMIP simulations and 45 years of equilibrium-climate simulations, as summarized in Tab. 1. The AMIP simulations use prescribed historical SST, sea ice, and CO_2. The equilibrium data span 1×, 2×, and 4× the reference CO_2 concentration. Although the equilibrium reference simulations originate from SHiELD coupled to a slab ocean model, SST is prescribed when these data are used for ACE2S training. Thus, the model is exposed to several equilibrium CO_2 climates but receives no training trajectories in which CO_2 is varied independently of the corresponding climate state.

Table 1: Training and validation data for ACE2S-SHiELD+$_{\text{no-RC}}$. Time bounds are inclusive; IC denotes initial-condition ensemble member.

Split	Simulation samples			Years
Training				
AMIP	IC1: 1979–2008; IC2: 1981-07–1986-06, 1991-07–1996-06, 2001-07–2006-06			45
Equilibrium	1×CO_2:	IC1: 2031–2040,	IC2: 2032–2036	45
	2×CO_2:	IC1: 2031–2040,	IC2: 2032–2036	
	4×CO_2:	IC1: 2031–2040,	IC2: 2032–2036	
Validation				
AMIP	IC1: 2009–2011			3
Equilibrium	IC2: 2038 at each of 1×, 2×, and 4×CO_2			3

ACE2S-SHiELD+ uses essentially the same prognostic, forcing, and diagnostic interface as ACE2-ERA5, summarized in Tab. 2. There are two relevant differences. First, ACE2S-SHiELD+ additionally predicts total frozen precipitation rate as an output-only diagnostic, which enters the atmospheric energy budget. Second, CO_2 is supplied as a next-timestep forcing, so the specified concentration corresponds to the six-hour interval over which the state update and associated boundary-energy fluxes are computed.

C.2 ACE2-ERA5: Historical Reanalysis Training Data

For the real-world reanalysis experiment, we use ACE2-ERA5 (Watt-Meyer et al., 2025), trained directly on ERA5 (Hersbach et al., 2020). The processed dataset spans 1940–2022. Training uses 1940–1995, 2011–2019, and 2021–2022; 1996–2000 is used for validation; and 2001–2010 and 2020 are held out for testing. These are the splits used by the released ACE2-ERA5 checkpoint.

Table 2: ACE2-ERA5 variable interface, which is also shared by ACE2S-SHiELD+ except for the differences described in § C.1. For vertically resolved variables, $k = 0, \ldots, 7$ indexes the eight atmospheric layers from the top of the atmosphere toward the surface. "Snapshot" denotes an instantaneous field, "Mean" a mean over the six-hour timestep, and "Invariant" a time-independent field. TOA denotes top of atmosphere.

Symbol	Description	Units	Time
Prognostic variables (input and output)			
T_k	Air temperature	K	Snapshot
q_k^T	Specific total water (vapor + condensates)	$\mathrm{kg\,kg^{-1}}$	Snapshot
u_k	Eastward wind	$\mathrm{m\,s^{-1}}$	Snapshot
v_k	Northward wind	$\mathrm{m\,s^{-1}}$	Snapshot
T_s	Skin temperature over land or sea ice	K	Snapshot
p_s	Surface pressure	Pa	Snapshot
$T_{2\mathrm{m}}$	2-m air temperature	K	Snapshot
$q_{2\mathrm{m}}$	2-m specific humidity	$\mathrm{kg\,kg^{-1}}$	Snapshot
$u_{10\mathrm{m}}$	10-m eastward wind	$\mathrm{m\,s^{-1}}$	Snapshot
$v_{10\mathrm{m}}$	10-m northward wind	$\mathrm{m\,s^{-1}}$	Snapshot
Forcing variables (input only)			
$\mathrm{DSWRF_{TOA}}$	Downward shortwave radiative flux at TOA	$\mathrm{W\,m^{-2}}$	Mean
T_s	Skin temperature over open ocean (SST)	K	Snapshot
z_s	Surface topographic height	m	Invariant
f_l	Land fraction	–	Invariant
f_o	Ocean fraction	–	Snapshot
f_{si}	Sea-ice fraction	–	Snapshot
$\mathrm{CO_2}$	Global-mean atmospheric $\mathrm{CO_2}$	ppm	Snapshot
Diagnostic variables (output only)			
$\mathrm{USWRF_{TOA}}$	Upward shortwave radiative flux at TOA	$\mathrm{W\,m^{-2}}$	Mean
$\mathrm{ULWRF_{TOA}}$	Upward longwave radiative flux at TOA	$\mathrm{W\,m^{-2}}$	Mean
$\mathrm{USWRF_{sfc}}$	Upward shortwave radiative flux at surface	$\mathrm{W\,m^{-2}}$	Mean
$\mathrm{ULWRF_{sfc}}$	Upward longwave radiative flux at surface	$\mathrm{W\,m^{-2}}$	Mean
$\mathrm{DSWRF_{sfc}}$	Downward shortwave radiative flux at surface	$\mathrm{W\,m^{-2}}$	Mean
$\mathrm{DLWRF_{sfc}}$	Downward longwave radiative flux at surface	$\mathrm{W\,m^{-2}}$	Mean
P	Surface precipitation rate, all phases	$\mathrm{kg\,m^{-2}\,s^{-1}}$	Mean
$(\partial \mathrm{TWP}/\partial t)_{\mathrm{adv}}$	Total-water-path tendency from advection	$\mathrm{kg\,m^{-2}\,s^{-1}}$	Mean
LHF	Surface latent heat flux	$\mathrm{W\,m^{-2}}$	Mean
SHF	Surface sensible heat flux	$\mathrm{W\,m^{-2}}$	Mean
Z_{500}	500-hPa geopotential height	m	Snapshot
T_{850}	850-hPa air temperature	K	Snapshot

ERA5 is natively represented on 137 model levels. Following Watt-Meyer et al. (2025), the fields are horizontally regridded to a 1° Gaussian grid and vertically coarsened to eight terrain-following hybrid sigma–pressure layers. Air temperature, specific total water, and horizontal winds are represented on each vertical layer; the remaining state, forcing, and diagnostic fields are two-dimensional. The temporal resolution is six hours. State variables are represented as instantaneous snapshots, while radiative, turbulent, and precipitation fluxes are means over the six-hour interval.

Tab. 2 lists the ACE2-ERA5 variable interface. Prognostic variables are supplied as inputs and predicted at the next timestep, forcing variables are input only, and diagnostic variables are predicted outputs that are not autoregressively fed back into the state. Surface skin temperature has different roles by surface type: over land and sea ice it is prognostic, whereas over open ocean it is prescribed and corresponds to sea-surface temperature (SST).

D COMPUTATIONAL DETAILS

***ReRoute* fine-tuning.** *ReRoute* is initialized from the released pretrained ACE2 checkpoints and does not retrain the underlying climate emulator from scratch. Fine-tuning for both the ACE2S-SHiELD+$_{\text{no-RC}}$ and ACE2-ERA5 experiments was performed on two NVIDIA A100 GPUs. Each run completed in less than 24 hours of wall-clock time and was continued until the training and validation metrics saturated. Experiment-specific optimization hyperparameters are reported in § F.

Seed sensitivity and checkpoint selection. We repeated fine-tuning with three random seeds and observed no material differences in convergence or downstream performance. We therefore selected a single high-performing checkpoint for the full long-horizon analysis and used it unchanged across the corresponding factual and counterfactual evaluations. The checkpoint evaluation criterion follows the those for the baseline models, as reported in (Watt-Meyer et al., 2025; Clark et al., 2026).

We report uncertainty across independent stochastic realizations and initial conditions, as specified in § F. For example, the abrupt-$4\times CO_2$ experiment reports variability across its 36-member initial-condition ensemble. We do not additionally report error bars across fine-tuning seeds because variation across the three adaptation runs was small relative to the predictive variability of the stochastic ACE2S emulator.

Compute comparison. The compute reported above refers only to *ReRoute* adaptation and excludes the cost of training the released ACE2 backbones and generating the pre-existing ERA5 and SHiELD datasets and evaluation simulations.

Our reported $18\times$ compute reduction applies specifically to the controlled SHiELD benchmark, where a directly comparable ACE2S-SHiELD+ model trained with additional random-CO_2 simulations is available. We compare the GPU-hours required to adapt the released no-RC checkpoint with *ReRoute* against the GPU-hours required to train that controlled-data counterpart. We do not make this compute comparison for ACE2-ERA5, for which no analogous model trained on counterfactual reanalysis trajectories exists.

E CLIMATE-SPECIFIC METHODOLOGY

This section provides the climate-specific methodological details used in our ACE2 experiments. We first describe the construction of the RRTMG-based CO_2 pathway and then the tropical-cyclone tracking procedure used for evaluation.

E.1 RRTMG CO_2 PATHWAY CONSTRUCTION

ReRoute requires only the first-level mechanistic response of the targeted variable. We compute the CO_2 radiative pathway offline using the RRTMG radiative-transfer scheme (Iacono et al., 2008), through its implementation in the Climate Modelling and Diagnostics Toolkit (`CliMT`) (Monteiro et al., 2018).

We first compute a climatological-mean atmospheric state from the training data. Holding this atmospheric state fixed, we run RRTMG at a reference CO_2 concentration $\mathbf{s}_0$ and at perturbed concentrations c. This isolates the instantaneous radiative response to changing CO_2, without including the subsequent dynamical adjustment of the atmosphere or ocean.

For each atmospheric level k, we compute the change in radiative heating relative to the reference state and normalize by the logarithmic CO_2 perturbation:

$$q_k^{\log} := \frac{q_k(\mathbf{s}) - q_k(\mathbf{s}_0)}{\log(\mathbf{s}/\mathbf{s}_0)}, \tag{E.1}$$

where $q_k(\mathbf{s})$ denotes the RRTMG radiative heating response at CO_2 concentration c. We similarly compute logarithmic response coefficients for the longwave radiative fluxes at the top of atmosphere and surface. These response coefficients are computed once and stored as fixed spatial fields.

For a requested CO_2 concentration $\mathbf{s}_t$, the pathway is obtained by scaling the stored coefficients by

$$\log(\mathbf{s}_t/\mathbf{s}_0)\,.$$

For atmospheric temperature, the corresponding six-hour increment at level k is

$$\Delta T_k = \frac{g\,\Delta t}{c_p\,\Delta p_k} q_k^{\log} \log\left(\frac{\mathbf{s}_t}{\mathbf{s}_0}\right), \tag{E.2}$$

where Δp_k is the pressure thickness of the layer, c_p is the specific heat capacity of air, and $\Delta t = 6\,\mathrm{h}$ is the ACE2-SOM timestep. The stored longwave-flux response coefficients are scaled in the same way for $\mathrm{ULWRF}_{\mathrm{toa}}$, $\mathrm{DLWRF}_{\mathrm{sfc}}$, and $\mathrm{ULWRF}_{\mathrm{sfc}}$.

This defines the mechanistic pathway

$$\Gamma(\mathbf{s}_0, \mathbf{s}_t),$$

which captures only the instantaneous first-level radiative effect of the CO_2 intervention. *Notably, it does not prescribe the full downstream climate response*. Subsequent changes in temperature, moisture, circulation, surface fluxes, and the slab ocean emerge through the autoregressive ACE2-SOM dynamics. In all experiments, the stored pathway coefficients are kept fixed.

E.2 Tropical Cyclone Tracking Details

As an additional evaluation of the weather statistics generated by ACE2-ERA5, we compare tropical-cyclone (TC) climatology between ERA5, the pretrained ACE2-ERA5 baseline, and ACE2-ERA5+*ReRoute* over the held-out 2001–2010 period using six-hourly 1° fields. TCs are detected and tracked with exactly the same **TempestExtremes v2.1** pipeline (Ullrich et al., 2021) for all three datasets, so differences in the resulting tracks arise from the generated atmospheric fields rather than from the detection procedure. Candidate storm centers are first identified with `DetectNodes` as local minima in mean sea-level pressure satisfying a closed-contour pressure drop of at least 200 Pa within a 5.5° radius. Following the ACE2-ERA5 evaluation protocol (Watt-Meyer et al., 2025), the warm-core criterion is expressed using the ACE2 T_3 atmospheric layer, whose pressure range is approximately 250–400 hPa, because ACE2 does not directly predict the upper-level geopotential thickness conventionally used by the tracker. A candidate must exhibit a decrease of at least 0.4 K in T_3 away from the storm center, approximately matching the hydrostatic thickness criterion used in the standard TempestExtremes configuration. Detected centers are then linked through time using `StitchNodes` to construct complete cyclone trajectories. We apply identical detection, stitching, duration, and displacement criteria to ERA5, Baseline, and Baseline+*ReRoute*, and summarize the resulting spatial track climatology over the full 10-year period. This controlled use of a common tracker permits a direct comparison of the model-generated TC statistics against ERA5.

F Climate Experiment Details and Additional Results

This section provides the full experimental configurations and additional results for our climate evaluations. We first describe the controlled coupled-climate experiments with ACE2S-SHiELD+, followed by the historical reanalysis experiments with ACE2-ERA5.

F.1 Coupled Climate Simulation

Experimental setup. We evaluate *ReRoute* on the released ACE2S-SHiELD+$_{\mathrm{no-RC}}$ checkpoint (Clark et al., 2026), a six-hour autoregressive global climate emulator with the same stochastic SFNO (Bonev et al., 2023) architecture and prognostic variables as ACE2S-SHiELD+, but trained without the random-CO_2 simulations used to explicitly decorrelate SST and CO_2. The no-RC checkpoint is trained on 90 years of data: 45 years of AMIP simulations and 45 years of equilibrium-climate simulations spanning $1\times$, $2\times$, and $4\times CO_2$. Validation uses held-out AMIP and equilibrium-climate years, and checkpoint selection uses 120 years of inline inference across the three equilibrium climates (details about the dataset see § C.1). Although the equilibrium reference simulations originate from ocean-coupled SHiELD, training, validation, and inline inference are performed with prescribed SST, following Clark et al. (2026). Thus, the model observes multiple equilibrium CO_2 climates but receives no training trajectories in which CO_2 is varied independently of the corresponding climate state. We use the released no-RC checkpoint as the pretrained baseline throughout this experiment.

Counterfactual query. We evaluate an abrupt-$4\times CO_2$ intervention using the SOM-coupled SHiELD benchmark of Clark et al. (2026). Each simulation is initialized from the $1\times CO_2$ climate and the atmospheric CO_2 concentration is instantaneously increased to $4\times CO_2$, while the slab ocean subsequently evolves interactively through the surface fluxes predicted by the atmospheric model. This intervention places the model in a forcing–state combination absent from its training distribution: although equilibrium climates at both $1\times$ and $4\times CO_2$ are represented during training, the model never observes a $1\times$ climate state subjected instantaneously to $4\times CO_2$ forcing. We follow the reference protocol and evaluate a 36-member ensemble of 90-day rollouts, with initial conditions spaced one month apart over three years to average weather variability and sample the seasonal cycle. The corresponding SHiELD simulations provide an independent reference trajectory for every ensemble member, enabling direct quantitative evaluation of the intervention response.

***ReRoute* configuration.** For this experiment, the CO_2 pathway is the fixed RRTMG radiative-response pathway described in § E.1, with the reference concentration set to $\mathbf{s}_0 = 363.43\,\mathrm{ppm}$ and pathway scaling fixed to one. The pretrained backbone is always conditioned on $\mathbf{s}_0$, while the abrupt change in CO_2 is supplied through $\Gamma(\mathbf{s}_0, \mathbf{s}_t)$. The baseline instead receives the requested CO_2 concentration through its native conditioning. The coupled evaluation is performed without random-CO_2 training trajectories or abrupt-$4\times CO_2$ supervision.

Evaluation Metrics Following Clark et al. (2026), our primary long-horizon metric is the channel-mean area-weighted RMSE of the normalized time- and ensemble-mean bias.

Let $\tilde{y}_{c,m,s}(\phi,\lambda)$ and $\tilde{\hat{y}}_{c,m,s}(\phi,\lambda)$ denote the normalized SHiELD reference and model prediction for channel c, ensemble member m, and six-hourly timestep s. Here, ϕ and λ denote latitude and longitude, respectively. We first form the time-mean bias map for each member,

$$\Delta_{c,m}(\phi,\lambda) = \frac{1}{T}\sum_{s=1}^{T}\left[\tilde{\hat{y}}_{c,m,s}(\phi,\lambda) - \tilde{y}_{c,m,s}(\phi,\lambda)\right],$$

and then average the bias maps across the $M = 36$ initial-condition members,

$$\overline{\Delta}_c(\phi,\lambda) = \frac{1}{M}\sum_{m=1}^{M}\Delta_{c,m}(\phi,\lambda).$$

For each output channel, the spatial error is

$$\mathrm{RMSE}_c = \left(\sum_{\phi,\lambda} w_{\phi,\lambda}\overline{\Delta}_c(\phi,\lambda)^2\right)^{1/2},$$

where $w_{\phi,\lambda}$ are normalized spherical-area weights. We then average across the $C = 51$ ACE2S-SHiELD+ output channels,

$$\alpha_{\mathrm{ens}} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{RMSE}_c.$$

Averaging over time and ensemble members before the nonlinear spatial RMS matches the evaluation convention of Clark et al. (2026).

Member-wise uncertainty. To quantify variation across the 36 independent initial conditions, we additionally compute the same aggregate error separately for each ensemble member,

$$\alpha_m = \frac{1}{C}\sum_{c=1}^{C}\left(\sum_{\phi,\lambda} w_{\phi,\lambda}\Delta_{c,m}(\phi,\lambda)^2\right)^{1/2},$$

and report the mean and population standard deviation of $\{\alpha_m\}_{m=1}^{M}$. This is the statistic shown with error bars in Fig. 3. The native no-RC model obtains 0.1619 ± 0.0116, while *ReRoute* obtains 0.1324 ± 0.0101, corresponding to an 18.2% reduction in mean member-wise error. The ensemble-mean-first metric α_{ens} and the member-wise statistic quantify complementary properties: the former measures error after weather noise has been averaged across the ensemble, while the latter exposes variability across independent initial conditions.

F.2 Historical Reanalysis with ACE2-ERA5

Experimental setup. We evaluate *ReRoute* on ACE2-ERA5 (Watt-Meyer et al., 2025), a six-hour autoregressive global atmospheric emulator trained directly on ERA5 reanalysis. ACE2-ERA5 operates at 1° horizontal resolution with eight atmospheric layers and is conditioned on prescribed sea-surface temperature (SST), sea ice, incoming solar radiation, global-mean atmospheric CO_2, and static surface fields. The released model is trained on ERA5 data from 1940–1995, 2011–2019, and 2021–2022, with 1996–2000 used for validation and 2001–2010 and 2020 reserved for testing (details about the dataset are provided in § C.2). Because SST, sea ice, and atmospheric CO_2 co-evolve throughout the historical record, ACE2-ERA5 never observes controlled variations that independently separate their effects. We use the released ACE2-ERA5 checkpoint as the pretrained baseline and initialize *ReRoute* from exactly the same model. The architecture, normalization statistics, conservation corrections, prognostic variables, prescribed boundary conditions, and six-hour autoregressive timestep are unchanged; *ReRoute* modifies only the pathway through which CO_2 enters the emulator.

Counterfactual query. We ask how the historical atmosphere would evolve if atmospheric CO_2 had remained fixed at its 1940 value while the remaining observed boundary conditions followed their historical trajectories. We therefore perform two otherwise identical 81-year rollouts from January 1940 through December 2020.

In the factual rollout, CO_2 follows its historical trajectory; in the intervention rollout, CO_2 is held fixed at its 1940 value, while SST, sea ice, incoming solar radiation, and all other prescribed forcings remain identical to the historical experiment. This intervention breaks the historical SST–CO_2 association present in the ERA5 training distribution. No observed atmosphere exists under the same realized 1940–2020 SST and sea-ice trajectory with CO_2 fixed, so this experiment is evaluated through the physical structure of the predicted response.

***ReRoute* adaptation.** We initialize *ReRoute* from the released ACE2-ERA5 checkpoint and adapt it using only factual historical ERA5 samples.

Training uses two-step autoregressive rollouts with an MSE objective applied at both rollout steps. We optimize with AdamW using learning rate 2×10^{-6} and zero weight decay, with a linear warmup from 0.1 to 1 times the base learning rate over the first 60 optimization steps. We use batch size four, no gradient accumulation or mixed precision, and exponential-moving-average model parameters for validation. The RRTMG pathway is constructed as described in § E.1, with reference concentration $\mathbf{s}_0$ set to the mean CO_2 concentration of the ACE2-ERA5 training data..

Fixed-CO_2 response. We evaluate the intervention using the spatial response of near-surface air temperature. Let $T_{2\mathrm{m}}^{\mathrm{hist}}(\phi, \lambda, t)$ and $T_{2\mathrm{m}}^{\mathrm{fix}}(\phi, \lambda, t)$ denote the predicted 2-m temperature under historical and fixed CO_2, respectively. We compute the time-mean response

$$\Delta T_{2\mathrm{m}}(\phi, \lambda) = \overline{T}_{2\mathrm{m}}^{\mathrm{fix}}(\phi, \lambda) - \overline{T}_{2\mathrm{m}}^{\mathrm{hist}}(\phi, \lambda), \tag{F.1}$$

where ϕ and λ denote latitude and longitude, respectively, and the overbar denotes the mean over the 1940–2020 rollout. Because SST and sea ice are identical between the two experiments, $\Delta T_{2\mathrm{m}}$ isolates the change produced by altering the emulator's CO_2 pathway while retaining the same prescribed surface evolution.

The native ACE2-ERA5 baseline produces widespread negative $\Delta T_{2\mathrm{m}}$, indicating that fixing CO_2 removes a substantial portion of the warming supported by the evolving boundary conditions. *ReRoute* produces a much weaker fixed-minus-historical response over most of the globe, preserving substantially more of the surface warming encoded by the prescribed SST and sea-ice evolution.

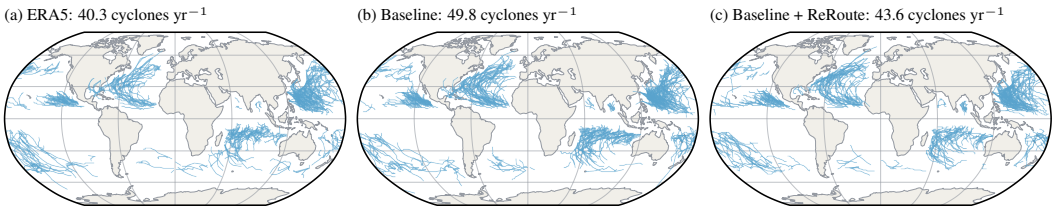

Figure 8: ***ReRoute* improves climatology of tropical cyclones in ACE2-ERA5.** Tropical cyclone tracks detected in ERA5, the pretrained ACE2-ERA5 baseline, and ACE2-ERA5 + *ReRoute* over the held-out evaluation period. The same detection and tracking procedure of tropical cyclones are applied to all three datasets. *ReRoute* produces a spatial distribution of cyclone activity that more closely matches the ERA5 reference than the pretrained baseline.

Tropical-cyclone evaluation. We additionally evaluate whether the factual climate generated by *ReRoute* reproduces the tropical-cyclone statistics of ERA5. Following the ACE2 evaluation protocol of Watt-Meyer et al. (2025), we analyze tropical-cyclone-like features over the held-out 2001–2010 period using the TempestExtremes tracking package. We apply the same detection and tracking parameters to ERA5, the pretrained ACE2-ERA5 baseline, and ACE2-ERA5+*ReRoute*. The tracker follows the default tropical-cyclone configuration used for ACE2, with the warm-core criterion adapted to the variables predicted by the emulator: instead of upper-level geopotential thickness, we use the T_3 atmospheric layer, corresponding approximately to 250–400 hPa, and require temperature to decrease by 0.4 K away from the cyclone center. We compare the resulting track climatology and global occurrence frequency across the three datasets. Because the identical tracker is applied to ERA5 and both emulator rollouts, differences directly reflect the generated atmospheric statistics rather than differences in event identification.

As shown in Fig. 8, the proposed *ReRoute* approach produces a distribution of tropical-cyclone tracks that more closely matches the ERA5 reference than the pretrained ACE2-ERA5 baseline. Details are summarized in § E.2.