Local Transitions and Global Admissibility: Composing Learned Atmospheric Update Operators

Flyxion
Independent Researcher

November 2026

Abstract

Hu et al. recently reported STRATA, a transformer emulator of a kilometer-scale global atmospheric model that is trained on small local tiles and deployed over the whole planet by repeated, overlapping application. The result is a convincing demonstration that a learned local transition rule can be composed into globally extended behavior at substantially lower reported energy cost per simulated day than the reference model. It is a much weaker demonstration that the composed rule preserves the constraints required of a physically admissible atmospheric history. This essay separates three properties that are easily conflated in evaluations of learned simulators, namely boundedness, constraint satisfaction, and physical fidelity, and argues that the validity of a global continuation cannot be inferred from the accuracy of its local operators alone. A short calculation shows that partition-of-unity blending of locally conservative flux-form updates introduces an explicit residual associated with spatial variations in the blending weights and disagreements among the predicted fluxes. The essay then states a trajectory-level admissibility criterion, relates it to constraint-first accounts of dynamics, and proposes a small conservative-advection experiment in which the claim can be tested directly.

1 What the Result Establishes

The central architectural move in STRATA [1] is to train locally and deploy globally. Rather than learning a transformation over all of the roughly twenty-five million horizontal cells of the reference simulation at once, the model learns a map from the state of a tile and its supplied forcing to the state of that tile one step later, with a step of ten minutes. The reference data come from a storm-resolving model with an effective dynamical grid spacing near 3.3 km, archived on a 4.9 km grid with 24 vertical levels, and the training set consists of seventeen days drawn from simulations representing different seasons. In schematic form the learned operator is

$$\widehat{X}^{(i)}_{t+\Delta t} = F_\theta\big(X^{(i)}_t, B^{(i)}_t\big), \tag{1}$$

where $X^{(i)}_t$ is the state of tile i and $B^{(i)}_t$ its forcing and context. A global step is assembled from overlapping tile predictions by a blending operator $\mathcal{B}$, and a trajectory is the iterate

$$X_{t+n\Delta t} = \widehat{F}^{\,n}(X_t), \qquad \widehat{F}(X) = \mathcal{B}\big(F_\theta(X|_{U_1}), \ldots, F_\theta(X|_{U_m})\big). \tag{2}$$

Tiles of about 576 km are defended by the observation that acoustic signals travel only about 200 km in ten minutes, which makes the locality of a single update a reasonable heuristic. The

distinction that matters is between the locality of an increment and the locality of a trajectory. Each step depends on a bounded neighborhood, yet the accumulated sequence of steps can carry a disturbance across continents and couple it to planetary circulation. This is the sense in which the system is more than super-resolution: its own output is the input to the next step, so fine-scale structure can feed back on later evolution. The broader problem of learning reusable local evolution rules has also been investigated in neural PDE solvers, including message-passing architectures that accommodate different discretizations and address autoregressive stability [11].

The reported engineering is careful. The network keeps height as a spatial axis through neighborhood attention in the horizontal and attention along the vertical column, in the lineage of rotary position embeddings [3] and neighborhood attention [2]. Its stereographic variant of rotary embedding builds coordinates centered on each tile and rotates wind components into a shared tangent frame, so that physically equivalent patterns at different latitudes are not presented to the network as geometrically distinct. The authors also analyze how patch embedding can introduce spatially periodic artifacts and how repeated application can amplify them, using the spectral radius $\rho(A)$ of a linearized update to ask whether small perturbations grow. A decoder built from bilinear upsampling and depthwise convolution is adopted to suppress the resulting checkerboard modes. The reported efficiency gains are large relative to the reference model: Table 1 of the paper gives about 48.1 simulated days per megawatt-hour on 16 H100 GPUs against 0.78 for the reference model, a ratio near 62, and a maximum throughput of 741 simulated days per wall-clock day at 512 GPUs. These figures concern the marginal cost of producing a trajectory. They do not include the cost of generating training data, training, validation, and correction.

2 What Remains Conditional

Three qualifications bound the claims that the evidence supports.

The first is that the model is an emulator of a numerical reference and not of the atmosphere. If F denotes the evolution operator of the reference model, the learning problem is $F_\theta \approx F$, and the scientific question splits into two validation problems: whether the reference model represents the physical atmosphere adequately, and whether the learned operator reproduces the relevant behavior of the reference. Perfect agreement at the second level would still be silent about the first, and for climate applications the first is dominated by uncertainties in cloud microphysics, aerosols, surface exchange, and unresolved turbulence. The prospect of century-scale simulation is a research aspiration of the paper, not one of its results.

The second is that seventeen days of global data are both impressive and easy to misread. Many tiles are extracted from each snapshot, which explains how a common local mechanism can be learned from little temporal coverage, but tiles from neighboring regions and successive times are strongly correlated, so the effective sample size is far below the raw count. Learning a shared local mechanism is a different task from learning the distribution of global states, which requires exposure to seasonal variability, rare events, and persistent circulation regimes. The evaluation reported consists of six held-out initializations at twelve-hour spacing in October 2020 with a rollout horizon of twenty-four hours, which supports a claim of useful short-range behavior and no more.

The third is the large-scale circulation. The paper's precipitation fields show organized, propagating convection, while the temperature and wind diagnostics show growing large-scale

error, global mean precipitation biased low by as much as twenty percent, and suppressed spectral power at length scales from roughly 100 to 2000 km. These are not peripheral defects. Writing the state as $X = X_L + X_S$ for large- and small-scale components, a faithful multiscale model must reproduce both $\dot{X}_L = F_L(X_L, X_S)$ and $\dot{X}_S = F_S(X_L, X_S)$, and the feedback between them is precisely what determines whether realistic-looking storms occur in a correct planetary state. A model that gets X_S plausibly while X_L drifts produces convincing weather in the wrong world. The authors suggest coupling the emulator to a coarse-resolution global emulator, which is sensible, and which raises its own problem of combining two operators without suppressing or double-counting their interaction.

3 Boundedness, Constraint Satisfaction, and Fidelity

The appendices describe the interventions that make the production model usable: multistep fine-tuning, the modified decoder, spectral filtering of large-scale vertical velocity, and corrections for negative humidity and precipitation. Each is a reasonable response to an observed failure, and together they show that a learned update operator does not inherit the constraints of the system it imitates. The three properties at stake can be stated separately.

Definition 1. A learned operator is *uniformly bounded* on a class K of initial states over horizon T if $\sup_{X_0 \in K} \sup_{0 \le t \le T} \|X_t\| \le M$ for a constant M fixed independently of the trajectory. It *satisfies a constraint* $g(X) \ge 0$ or $g(X) = 0$ if the constraint holds along its rollouts. It is *physically faithful* if the distribution of its trajectories, and the budgets accumulated along them, agree with those of the system being modeled to within a stated tolerance.

The horizon and the uniformity carry the content of the first property. For any single finite trajectory of finite states, boundedness is automatic, so the meaningful statements are a bound uniform over a class of initial conditions and a finite T, and the far stronger requirement of a bound uniform in T. Surviving a twenty-four-hour rollout is evidence for the former and none for the latter. Even finite-horizon uniform boundedness is mathematically weak, since for a continuous finite-dimensional update operator and a bounded set of initial states it often follows automatically, so the definition carries less weight than the properties it is meant to separate from the others. What matters in practice is stability under perturbation, persistence over increasing horizons, and the absence of physically inadmissible attractors, none of which is established by a bound alone. The properties are related and not equivalent. Clipping negative humidity enforces $q_v \ge 0$ and says nothing about whether the water budget closes. A spectral filter can remove an unstable large-scale mode and still leave the mechanism that excited it. A rollout can stay bounded for twenty-four hours while drifting toward the wrong climate over a year. Evaluations that report forecast error at a fixed lead time measure none of these directly, and the paper, appropriately, claims useful twenty-four-hour stability and nothing more comprehensive. The distinction belongs to the wider literature on structure-preserving numerical integration, where the long-time behavior of an approximation is studied in relation to the geometric and dynamical properties of the equations being approximated [9].

4 Composition Is Not Free

Partition-of-unity constructions are well established in numerical approximation [7]. Their use as a means of assembling local approximations does not, however, imply preservation of

every invariant of the underlying evolution. The assembly step deserves more scrutiny than a statement that it is smooth. Consider tiles U_i with partition-of-unity weights $w_i \geq 0$, $\sum_i w_i = 1$, on a closed surface $\mathbb{S}^2$, and suppose that each tile produces an increment in flux form,

$$\Delta q_i = -\Delta t\, \nabla\cdot F_i + \Delta t\, S_i, \tag{3}$$

so that every tile is locally conservative in flux form, up to its source S_i. This is a statement about a local continuity equation and not about the total content of the tile: over U_i one has $\int_{U_i} \Delta q_i\, dA = -\Delta t \int_{\partial U_i} F_i \cdot n_i\, ds + \Delta t \int_{U_i} S_i\, dA$, so the tile exchanges quantity with its surroundings through its boundary. Write $\bar{F} = \sum_j w_j F_j$ for the blended flux.

Proposition 1. *The blended increment $\Delta q = \sum_i w_i \Delta q_i$ satisfies*

$$\Delta q = -\Delta t\, \nabla\cdot \bar{F} + \Delta t \sum_i w_i S_i + \Delta t\, R, \qquad R = \sum_i \nabla w_i \cdot (F_i - \bar{F}), \tag{4}$$

and consequently

$$\int_{\mathbb{S}^2} \Delta q\, dA = \Delta t \int_{\mathbb{S}^2} \sum_i w_i S_i\, dA + \Delta t \int_{\mathbb{S}^2} R\, dA. \tag{5}$$

Proof. The product rule gives $\nabla\cdot(w_i F_i) = w_i \nabla\cdot F_i + \nabla w_i \cdot F_i$, and summing over i yields $\sum_i w_i \nabla\cdot F_i = \nabla\cdot \bar{F} - \sum_i \nabla w_i \cdot F_i$. Therefore

$$\Delta q = -\Delta t \sum_i w_i \nabla\cdot F_i + \Delta t \sum_i w_i S_i = -\Delta t\, \nabla\cdot \bar{F} + \Delta t \sum_i \nabla w_i \cdot F_i + \Delta t \sum_i w_i S_i.$$

Since $\sum_i \nabla w_i = 0$, one has $\sum_i \nabla w_i \cdot F_i = \sum_i \nabla w_i \cdot (F_i - \bar{F}) = R$. The integral of $\nabla\cdot \bar{F}$ over a closed surface vanishes, which gives the second identity. □

The residual R is zero wherever the tiles agree on the flux or wherever the weights are constant, and it is generically nonzero in the overlap zones, where both conditions fail together. A nonzero pointwise residual does not by itself imply a nonzero global drift, since $\int R\, dA$ can vanish while $R \not\equiv 0$. The precise conclusion is that partition-of-unity blending does not guarantee conservation, and not that it must violate it; stated this way the result is immune to counterexamples built on cancellation. Smooth blending therefore guarantees continuity across tile boundaries and gives no guarantee of conservation. The proposition concerns a flux-form idealization of overlapping prediction assembly and is not a derivation of the internal conservation error of STRATA. Its significance for that model is diagnostic: overlapping state predictions do not acquire conservation guarantees merely because the assembly is smooth, and the actual budget residual must be measured for the complete implemented update.

A minimal analytic example shows both behaviors. On the periodic interval $[0, 1)$ take $w_1 = \frac{1}{2}(1 + \cos 2\pi x)$ and $w_2 = 1 - w_1$, each tile defined on the whole circle, which is the extreme case of overlap. If the two tiles predict constant fluxes $F_1 = a$ and $F_2 = b$ with no sources, then $\Delta q_1 = \Delta q_2 = 0$ and $\bar{F} = b + (a - b)w_1$, while $R = (a - b)w_1'$ is nonzero wherever $\sin 2\pi x \neq 0$, yet $\oint R\, dx = 0$ and no mass drifts. If instead $F_1 = \sin 2\pi x$ and $F_2 = 0$, each tile conserves total mass separately, since $\Delta q_1 = -2\pi \Delta t \cos 2\pi x$ and $\Delta q_2 = 0$ both integrate to zero, but the blended increment is $-\pi \Delta t \cos 2\pi x\, (1 + \cos 2\pi x)$ and

$$\int_0^1 \Delta q\, dx = \Delta t \int_0^1 w_1'\, (F_1 - F_2)\, dx = -\pi \Delta t \int_0^1 \sin^2 2\pi x\, dx = -\tfrac{\pi}{2}\, \Delta t \neq 0.$$

Two locally conservative operators therefore compose into a globally nonconservative one, with no neural network involved.

The proposition motivates a definition of the property an assembly should have. Let $\mathcal{A}_i$ be the set of locally admissible increments on U_i and $\mathcal{A}_G$ the set of globally admissible increments, and let $\mathcal{K} \subseteq \prod_i \mathcal{A}_i$ be a compatibility set of tuples of local increments that share a common initial state and compatible boundary data. The set $\mathcal{K}$ encodes compatibility of the inputs and not agreement of the predicted fluxes, since otherwise the preceding counterexample could be excluded by definition.

Definition 2. An assembly operator $\mathcal{B}$ is *admissibility-preserving* if $\mathcal{B}(\mathcal{K}) \subseteq \mathcal{A}_G$.

The example above shows that partition-of-unity blending of increments is not admissibility-preserving when the global condition is conservation of total mass, so local admissibility together with input compatibility does not imply a globally admissible composition. This gives the argument a plain structure: a counterexample to automatic admissibility preservation, followed by a definition of the property that a composition procedure ought to possess.

The proposition is conditional on flux-form predictions. A network that predicts the increment Δq_i directly has no F_i to inspect, and the corresponding defect is then simply whatever the integrated increment fails to match, which can only be measured and not controlled by construction. This suggests a direct diagnostic for the model in question, namely the conservation residuals of mass and water before and after stitching, reported separately at tile boundaries and at the face boundaries of the underlying cubed-sphere grid.

A second compatibility condition concerns the geometry. Predictions on overlapping tiles are expressed in different local frames and ought to agree after the transition map, $X_i'|_{U_i \cap U_j} = T_{ij}(X_j'|_{U_i \cap U_j})$. Stereographic coordinates and tangent-frame wind rotation are steps toward this condition, and averaging is the pragmatic means of resolving whatever disagreement remains. Averaging incompatible predictions does not demonstrate that they were compatible. Two notions of continuation can accordingly be distinguished. A *weak* continuation yields a globally defined state by reconciling local predictions numerically. A *strong* continuation yields local predictions that satisfy stated compatibility conditions before reconciliation, or comes with an argument that reconciliation preserves them. The system described is a practical instance of the first and makes partial progress toward the second.

Two geometric consistency tests should be kept distinct. Coordinate changes between overlapping local frames must satisfy the cocycle condition on triple overlaps, $T_{ij}T_{jk}T_{ki} = I$ with composition order adjusted to the convention in use, whereas parallel transport around a closed spatial path generally has nontrivial holonomy. On a sphere of radius a, the oriented rotation produced by parallel transport around a simple closed loop is the enclosed area divided by a^2, modulo 2π, a consequence of the Gauss–Bonnet theorem [6]. A geometric consistency test must therefore distinguish errors in the composition of coordinate transitions from discrepancies between learned transport and the holonomy prescribed by the underlying curvature.

5 Admissible Histories

A sequence of plausible states need not be an admissible history. Let $\gamma = (X_0, \ldots, X_T)$ be a trajectory and let $\mathcal{C}[\gamma]$ collect the violations of specified physical and geometric conditions, including budgets accumulated across the whole history.

Definition 3. A trajectory γ is *admissible at tolerance ε relative to a declared constraint set* if $\|\mathcal{C}[\gamma]\| \leq \varepsilon$, where $\mathcal{C}$ collects the violations of exactly those constraints. A learned operator is *admissibly composable* on a class of initial states if its iterates are admissible at tolerance ε for every initial state in the class and every horizon up to T.

Admissibility is relative to the declared constraints and does not establish fidelity unless those constraints suffice to characterize the intended dynamics. The identity operator $X_{t+1} = X_t$ preserves mass, energy, and positivity indefinitely and represents no atmospheric evolution at all. The point reinforces the earlier separation of constraint satisfaction from fidelity, and it keeps admissibility from becoming a synonym for physical correctness. A snapshot evaluation inspects properties of individual X_t. A trajectory evaluation inspects the relations between successive states, and a persistent moisture-budget error is invisible to the first and decisive for the second, since each frame may show realistic precipitation while the history as a whole is inadmissible. This motivates a training objective with an explicit history term,

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_c \, \mathcal{L}_{\text{constraint}} + \lambda_h \, \mathcal{L}_{\text{history}}, \tag{6}$$

in which the first term measures agreement with the reference, the second penalizes violations of local requirements, and the third penalizes accumulated deviations that no single snapshot reveals. This formulation is a proposal and is not part of the paper. Multistep fine-tuning already moves in this direction by penalizing error over several steps, though its target is agreement with the reference and not the budgets themselves.

The same distinction clarifies the role of the corrections applied during rollout. The state accepted for continuation is the output of the network together with every filter and clipping that follows it, and an honest description of the simulation must attribute the trajectory to that whole procedure. A record of each transition that separates the raw prediction from the accepted state, along with the operator version, the forcing, and the constraint checks, would make the history auditable in the sense that deterministic replay of an event log is auditable in append-only systems. Logging every cell update is infeasible, and the practical form is periodic checkpoints with content-addressed state identifiers and summarized conservation diagnostics.

6 Constraint-First Readings

The preceding distinctions have a natural statement in constraint-first accounts of dynamics, in which a transition is a member of an admissible set determined by the current state and not a free choice of next state. In such a reading, a field theory with a scalar capacity Φ, a vector flow $\mathbf{v}$, and an entropy field S would require that an update $\Delta\mathcal{X}_i$ proposed on region U_i satisfy $\Delta\mathcal{X}_i \in \mathcal{A}(\mathcal{X}_t|_{U_i})$, and that the blended update satisfy a second condition at the global level,

$$\mathcal{B}(\{\Delta\mathcal{X}_i\}) \in \mathcal{A}_{\text{global}}(\mathcal{X}_t). \tag{7}$$

Local admissibility of each proposal and global admissibility of their composition are different requirements, and the proposition above is an instance of their difference: each tile can be conservative while the composition is not. These are conceptual correspondences and do not amount to an identification of meteorological variables with those of any particular framework, and the paper does not use such a framework. The correspondence earns its place only if $\mathcal{A}$ is made concrete enough to measure, for instance through dry-air mass, water and energy budgets,

positivity conditions, and consistent fluxes. Entropy requires particular care, since a coarse-grained atmospheric entropy budget includes transport, irreversible production, and exchange with surroundings, and cannot be reduced to a demand that entropy increase separately within every tile.

The information-theoretic analysis in the paper bears on the same point. The authors argue that kilometer-scale emulation needs about ten times more computation per grid point than extrapolation from coarse models would predict, and they support the claim with differential entropies of standardized increments, covariance structure, and compressibility, comparing the reference simulation with reanalysis data [4]. The scaling observation is empirical and meaningful. The proposed explanation is less secure, because differential entropy depends on variables, normalization, and sampling interval, and because the entropy of an increment is not the conditional entropy $H(X_{t+\Delta t} \mid X_t, B_t)$ that governs predictability. A field can have high marginal entropy and be highly predictable given its past, and a low-entropy variable can be hard to predict when its dynamics depend on omitted information. Three quantities are worth separating: the information needed to represent a state, the information needed to predict its successor, and the information needed to preserve the distinctions on which a physically valid continuation depends. Two configurations that are close under a coarse spatial metric may differ in a localized moisture gradient that decides whether convection develops within a few steps, and a representation that discards that gradient preserves appearance while destroying predictive relevance. A constrained rate-distortion formulation [5] expresses the question:

$$R(D, \varepsilon) = \inf I(X; Z) \quad \text{subject to} \quad \mathbb{E}\, d(X, \widehat{X}) \le D, \;\; \mathbb{E}\, \|\mathcal{C}[\widehat{\gamma}]\| \le \varepsilon. \tag{8}$$

Here Z is a compressed representation of the state, d measures reconstruction distortion between the state and its reconstruction $\widehat{X}$, and $\widehat{\gamma}$ is the trajectory generated from the decoded representation. The infimum is taken over a specified class of stochastic encoders and associated decoders and rollout operators, and the mutual information requires additional care for deterministic encoders of continuous variables. Unlike ordinary rate-distortion theory, the proposed constraint also restricts the admissibility residual accumulated along the reconstructed trajectory, so this is a proposed extension of the classical formulation and not an established result. The question it poses is how much information a model must retain to preserve admissible future evolution and not merely to reconstruct the present state.

7 A Test Where the Answer Is Known

The claim that local accuracy does not settle global admissibility can be tested in a setting where the invariants are exact. Finite-volume methods provide the relevant numerical baseline because their discrete conservation properties follow from flux accounting across cell interfaces [8], and conservation-oriented discretization has a substantial history in atmospheric modeling [10]. Take a scalar density on a periodic square domain Ω obeying $\partial_t q + \nabla \cdot (q\mathbf{u}) = 0$ with a prescribed velocity field, so that $M(t) = \int_\Omega q\, dx$ is constant. Train a local neural update operator on tiles, assemble global predictions with overlapping partition-of-unity weights, and track the relative mass drift

$$\varepsilon_M(t) = \frac{|M(t) - M(0)|}{|M(0)|}. \tag{9}$$

Compare three assemblies, referred to below as increment blending (ordinary blending of predicted increments), flux blending (blending of fluxes followed by a divergence), and shared-

interface fluxes (a conservative assembly in which interface fluxes are shared by adjacent tiles so that they cancel exactly). Record in addition the overlap disagreement $\varepsilon_{ij} = \|X'_i - T_{ij}(X'_j)\|_{U_i \cap U_j}$ before blending and the long-horizon error against a trusted numerical solver. The Proposition predicts that increment blending permits mass drift driven by R, and that flux blending eliminates global drift under a conservative discretization of the divergence. The two flux-based assemblies are not automatically different, since a conservative discrete divergence of a single blended flux field already uses consistent interface fluxes in many finite-volume implementations, and the two may then coincide. To make the comparison meaningful the discretizations must be specified. Flux blending interpolates the tile-predicted fluxes onto a common grid before applying a conservative divergence. Shared-interface fluxes instead reconcile explicitly the numerical fluxes supplied independently by neighboring tiles, so that adjacent cells use the same flux with opposite signs. Global conservation and local conservative consistency are separate requirements, and the experiment tests whether the additional reconciliation improves long-horizon fidelity, without presuming that it is a strictly stronger conservation property. The informative outcome is whether enforcing compatibility improves the continuation of the system at long horizons, and not merely the smoothness of the output. The natural next step is shallow-water flow on the sphere, where curvature, wave propagation, and global coupling make the geometric and conservation conditions demanding in the way the atmospheric case is.

Remark 1 (Epistemic status). The blending-residual proposition and the periodic counterexample are exact results, with the explicit hypothesis that tile updates are in flux form. The notion of an admissibility-preserving assembly is a definition. The trajectory-level admissibility criterion, the history term in the training objective, the constrained rate-distortion formulation, and the conservative-advection comparison are proposals, and none of them has been tested here. Statements about STRATA rest on the reported results of its authors and not on an independent evaluation.

8 Conclusion

The work under discussion shows that a learned local transition rule can be composed over an entire planetary domain with substantially lower reported inference energy cost than the reference simulation, on the hardware and accounting the authors describe, while retaining substantial short-range dynamical structure. It also exposes the limits of that strategy. Local prediction accuracy does not give global dynamical fidelity, geometric consistency does not give conservation, numerical stability does not give physical correctness, and a sequence of realistic states is not thereby an admissible history. The general proposition suggested by the case is that the validity of a global continuation cannot be inferred solely from the accuracy of its local transition operators and must also be established through the compatibility of their composition and the preservation of constraints across the resulting history. The proposition is broad enough to apply well outside atmospheric science and specific enough to be tested with the conservative-advection experiment above.

References

[1] Z. Hu, A. Subramaniam, N. Keen, T. Ge, J. Pathak, M. S. Abbas, S. Ravuri, K. Kashinath, N. Mahfouz, P. Caldwell, M. Pritchard, and N. Brenowitz. Scaling Storm-Resolving

Atmospheric AI Simulation to the Entire Planet. arXiv:2606.31248, June 2026.

[2] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi. Neighborhood Attention Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[3] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv:2104.09864, 2021.

[4] H. Hersbach et al. The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, 2nd edition. Wiley, 2006.

[6] M. P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice-Hall, 1976.

[7] I. Babuška and J. M. Melenk. The Partition of Unity Method. *International Journal for Numerical Methods in Engineering*, 40(4):727–758, 1997.

[8] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[9] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd edition. Springer, 2006.

[10] A. Arakawa and V. R. Lamb. Computational Design of the Basic Dynamical Processes of the UCLA General Circulation Model. In *Methods in Computational Physics*, volume 17, pages 173–265. Academic Press, 1977.

[11] J. Brandstetter, D. E. Worrall, and M. Welling. Message Passing Neural PDE Solvers. In *International Conference on Learning Representations*, 2022.