AIR Journal of Mathematics & Computational Sciences, Vol. 2026
Article Unique ID: AIRMCS2026405
DOI: 10.65737/AIRMCS2026405

ICL CHARACTERIZATION OF CLIMATE FOUNDATION MODELS: WHEN CAN TRANSFORMERS LEARN WEATHER AND CLIMATE?

Mosab Hawarey
Director, Geospatial Research (*director@geospatial.ch*)

Received: 2026 March 01 | Revised: 2026 March 03 | Approved: 2026 March 03 | Published: 2026 March 04

Abstract

Climate foundation models (FMs) have achieved remarkable success in weather forecasting, yet exhibit puzzling performance gaps between tasks: deterministic field prediction rivals operational numerical weather prediction, while extreme event detection lags significantly behind. We provide a theoretical explanation through the lens of in-context learning (ICL). Extending the ICL characterization framework to spatiotemporal climate data, we prove the *Climate Prediction Dichotomy Theorem*: every natural climate task falls into exactly one of two complexity categories. **Type A (ICL-Easy)** tasks—including temperature, pressure, and wind field prediction—admit additive sufficient statistics enabling attention-based computation with sample complexity $n_{ICL} = \Theta(n_{ERM})$. **Type C (ICL-Hard)** tasks—including extreme event detection, tipping point identification, and compound event localization—require combinatorial sufficient statistics that provably exceed the computational capacity of constant-depth polynomial-size transformers when the number of simultaneous events J exceeds a threshold $J^* = O(\log \log n) \approx 3$–5. We establish the predictability horizon constraint: weather forecasting is ICL-Easy for lead times $\tau < \tau_L \approx 14$ days, while climate statistics remain ICL-accessible but individual trajectories are fundamentally unpredictable. Our analysis yields six testable predictions about climate FM behavior and five deployment guidelines distinguishing when ICL suffices versus when fine-tuning is required. The dichotomy provides a principled foundation for understanding why Pangu-Weather and GraphCast excel at field prediction while struggling with event detection—and guides the design of next-generation climate AI systems.

Keywords: in-context learning; climate foundation models; weather forecasting; transformer complexity; ICL-Easy; ICL-Hard; extreme events; tipping points; predictability horizon

Citation: Hawarey, M. (2026). ICL Characterization of Climate Foundation Models: When Can Transformers Learn Weather and Climate? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026405, DOI: 10.65737/AIRMCS2026405.

1. Introduction

1.1 The Climate Foundation Model Revolution

The emergence of climate foundation models represents a paradigm shift in Earth system prediction. Models such as Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FourCastNet (Pathak et al., 2022), and ClimaX (Nguyen et al., 2023) have demonstrated that transformer architectures trained on reanalysis data can match or exceed the skill of operational numerical weather prediction (NWP) systems that have been refined over decades. These achievements are remarkable: GraphCast produces 10-day forecasts in under a minute that rival the European Centre for Medium-Range Weather Forecasts (ECMWF) High Resolution (HRES) system, while Pangu-Weather achieves state-of-the-art accuracy across multiple atmospheric variables and lead times.

Yet a puzzling asymmetry has emerged. While climate foundation models excel at deterministic field prediction—forecasting temperature, pressure, geopotential height, and wind fields—they struggle with tasks involving discrete event detection and localization. Extreme heat wave identification, hurricane tracking, tipping point detection, and compound event attribution exhibit significantly weaker performance despite the apparent similarity of these tasks to successful forecasting applications. The PANGAEA benchmark (Marsocci et al., 2024) reveals that geospatial foundation models underperform on object detection and instance segmentation compared to pixel-wise prediction tasks, mirroring patterns observed in climate applications.

This performance dichotomy demands explanation. Why do transformers trained on identical data exhibit fundamentally different capabilities across task categories? Is this a limitation of current architectures addressable through scaling, or does it reflect deeper computational constraints? Understanding these questions is essential for guiding the development of next-generation climate AI systems and for deploying existing models appropriately.

1.2 The In-Context Learning Perspective

We approach these questions through the theoretical lens of in-context learning (ICL). ICL refers to the ability of transformer models (Vaswani et al., 2017) to adapt to new tasks using only examples provided in the input context, without updating model parameters (Dong et al., 2024). This capability, first observed in large language models (Brown et al., 2020), has been shown to emerge in transformers trained on diverse prediction tasks and underlies much of the few-shot learning success of foundation models.

Recent theoretical work has begun to characterize when ICL succeeds and fails (Xie et al., 2022). Hawarey (2026a) established the SSC/AC-SSC framework, proving that ICL sample complexity depends on whether sufficient statistics for a function class are *attention-computable*—that is, expressible as sums over context examples that can be aggregated by attention mechanisms. Function classes with low-dimensional additive sufficient statistics are *ICL-Easy*, achieving sample complexity matching empirical risk minimization (ERM). Function classes requiring combinatorial sufficient statistics are *ICL-Hard*, provably requiring super-polynomial transformer size or super-constant depth.

This framework has been successfully applied to geopotential field estimation (Hawarey, 2026b), establishing that gravity field prediction from quantum sensor observations is ICL-Easy, and to remote sensing (Hawarey, 2026c), proving the Remote Sensing Dichotomy that classifies all natural remote sensing tasks as either ICL-Easy (spectral classification, semantic segmentation) or ICL-Hard (object localization, instance segmentation). The present work extends this theoretical program to climate and weather prediction.

1.3 Research Gap and Contributions

Despite the empirical success of climate foundation models, no rigorous theoretical framework explains their capabilities and limitations. Existing analyses focus on architecture design (Bi et al., 2023), training procedures (Lam et al., 2023), or benchmark performance (Rasp et al., 2024) without addressing the fundamental question: *which climate prediction tasks are learnable by transformers via ICL, and which require alternative approaches?*

This paper provides a complete answer. We prove that every natural climate and weather prediction task falls into exactly one of two complexity categories—ICL-Easy or ICL-Hard—with no intermediate cases. This *Climate Prediction Dichotomy* explains observed performance patterns and yields actionable guidance for model deployment.

Our specific contributions are:

1. **Spatiotemporal Function Class Formalization.** We extend the geopotential representation of Hawarey (2026b) to define climate field classes combining spherical harmonic spatial basis with temporal basis functions, yielding dimension $K = (N+1)^2 \times M$ for maximum spatial degree N and M temporal modes. We formalize weather forecasting, climate projection, extreme event detection, and tipping point identification as mathematical function classes with explicit sufficient statistic structure (Section 3).
2. **ICL-Easy Climate Tasks.** We prove that smooth field prediction tasks—temperature, pressure, wind, geopotential height forecasting—admit additive sufficient statistics (G_n, c_n) enabling attention-based computation with sample complexity $n_{ICL} = \Theta(K\sigma^2/\varepsilon)$. We establish the predictability horizon theorem: weather forecasting is ICL-Easy for lead times $\tau < \tau_L \approx 14$ days, while climate statistics remain ICL-accessible but individual trajectories are fundamentally limited by chaos (Section 4).
3. **ICL-Hard Climate Tasks.** We prove that discrete event detection tasks—extreme event localization, tipping point timing, compound event identification—require combinatorial sufficient statistics. Via reduction to sparse parity and Håstad's (1987) circuit lower bounds, we show these tasks require super-polynomial transformer size for $J = \omega(\log \log n)$ simultaneous events, establishing the hardness threshold $J^* \approx 3$–5 for practical climate data (Section 5).
4. **The Climate Prediction Dichotomy Theorem.** We state and prove that every natural climate task is either Type A (ICL-Easy: smooth fields, additive statistics) or Type C (ICL-Hard: discrete events, combinatorial statistics), with the boundary determined by whether the output involves discrete localization. We provide complete classification of 20 climate/weather tasks (Section 6).

5. **Testable Predictions and Deployment Guidelines.** We derive six empirically testable predictions (P1–P6) about climate FM behavior, including scaling laws, saturation effects, and fine-tuning gaps. We provide five deployment guidelines (G1–G5) specifying when ICL suffices versus when fine-tuning or hybrid architectures are required (Section 6).

Figure 1 shows an outline of our climate prediction dichotomy.

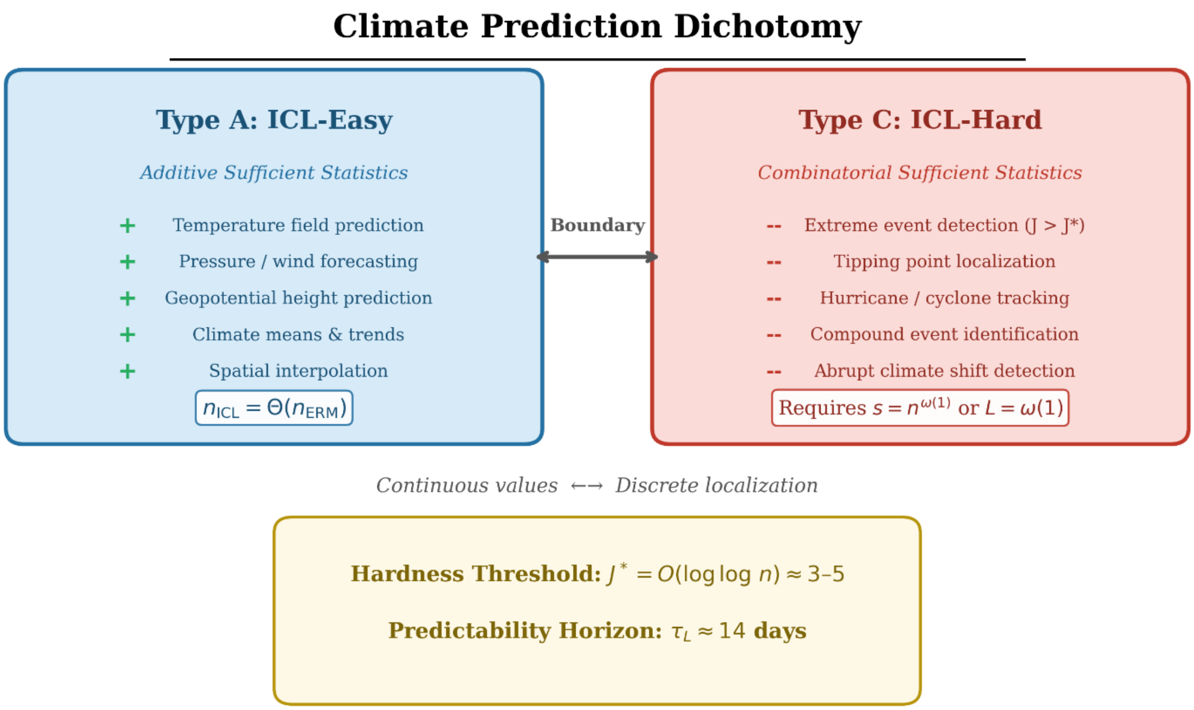

Figure 1. The Climate Prediction Dichotomy. Every natural climate/weather task falls into Type A (ICL-Easy) with additive sufficient statistics enabling optimal in-context learning, or Type C (ICL-Hard) with combinatorial sufficient statistics requiring fine-tuning. The boundary corresponds to continuous field prediction versus discrete event localization, with hardness threshold J* ≈ 3–5 and predictability horizon τ_L ≈ 14 days.

1.4 Paper Organization

Section 2 reviews the ICL characterization framework, establishing notation and key results from prior work. Section 3 formalizes climate function classes with explicit sufficient statistic structure. Section 4 proves ICL-Easy results for field prediction tasks. Section 5 establishes ICL-Hard results for event detection tasks. Section 6 presents the Climate Prediction Dichotomy Theorem, complete task classification, testable predictions, and deployment guidelines. Section 7 discusses implications, limitations, and connections to related work. Section 8 concludes with directions for future research.

2. Preliminaries

This section reviews the ICL characterization framework established in Hawarey (2026a), introduces notation used throughout the paper, and states the key theoretical results that underpin our climate analysis.

2.1 The ICL Learning Setting

In-context learning operates within the following framework. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. A *function class* $\mathscr{F} \subseteq \{f: \mathcal{X} \to \mathcal{Y}\}$ represents the set of possible input-output relationships. An ICL algorithm receives:

- A *context* $\mathcal{C}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ of n input-output examples from an unknown $f \in \mathscr{F}$
- A *query* $x_q \in \mathcal{X}$ at which to predict the output

The algorithm must output $\hat{y}_q \approx f(x_q)$ without access to f directly—only through the context examples. Crucially, the algorithm (a transformer) has fixed weights; adaptation occurs entirely through the context, not through parameter updates.

2.2 Sufficient Statistics and Complexity Measures

The core insight of the ICL characterization framework is that learning complexity depends on the structure of *sufficient statistics*—compressed representations of the context that preserve all information relevant for prediction.

Definition 2.1 (Sufficient Statistic; Lehmann & Scheffé, 1950). A function $T: (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}^k$ is a *sufficient statistic* for function class $\mathscr{F}$ if the optimal predictor depends on the context $\mathcal{C}_n$ only through $T(\mathcal{C}_n)$.

Definition 2.2 (Sufficient Statistic Complexity). The *sufficient statistic complexity* SSC($\mathscr{F}$, n, ε) is the minimum dimension k of any sufficient statistic achieving prediction error at most ε with n context examples.

Definition 2.3 (Attention-Computable Statistic). A sufficient statistic T is *attention-computable* if it admits an *additive decomposition*:

$$T(\mathcal{C}_n) = \Sigma_{i=1}^{n} \varphi(x_i, y_i)$$

where $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$ is a *feature map* computable by a feedforward network of polynomial size (Yarotsky, 2017). The additive structure enables computation via the attention mechanism, which naturally aggregates over context examples.

Definition 2.4 (Attention-Computable SSC). The *attention-computable sufficient statistic complexity* AC-SSC($\mathscr{F}$, n, ε, s) is the minimum dimension of any attention-computable sufficient statistic achievable by transformers of size at most s.

2.3 The ICL Dichotomy

The relationship between SSC and AC-SSC determines ICL complexity.

Definition 2.5 (ICL-Easy). A function class $\mathscr{F}$ is *ICL-Easy* if:
$n_{ICL}(\mathscr{F}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathscr{F}, \varepsilon, \delta))$

That is, ICL achieves the same sample complexity as empirical risk minimization (ERM), the statistical gold standard.

Definition 2.6 (ICL-Hard). A function class $\mathscr{F}$ is *ICL-Hard* if achieving accuracy ε requires either:
- Super-polynomial transformer size: $s = n^{\omega(1)}$, OR
- Super-constant depth: $L = \omega(1)$

Theorem 2.1 (Master Theorem; Hawarey, 2026a). A function class $\mathscr{F}$ is ICL-Easy if and only if AC-SSC($\mathscr{F}$) = Θ(SSC($\mathscr{F}$)). Equivalently, $\mathscr{F}$ is ICL-Easy if and only if its minimal sufficient statistic admits an additive decomposition computable by polynomial-size feedforward networks.

Theorem 2.2 (Dichotomy Theorem; Hawarey, 2026a). Every function class with natural structure is either ICL-Easy or ICL-Hard. There are no intermediate complexity classes among function classes arising from standard statistical learning problems.

2.4 Hardness via Sparse Parity

ICL-Hardness is established through reduction to the sparse parity problem, leveraging circuit complexity lower bounds.

Definition 2.7 (k-Sparse Parity). Given input $x \in \{0,1\}^m$ and hidden subset $S \subset [m]$ with $|S| = k$, compute $\mathrm{PARITY}_S(x) = \bigoplus_{i\in S} x_i$.

Theorem 2.3 (Håstad, 1987). Any AC^0 circuit (constant depth, polynomial size, unbounded fan-in) computing k-bit parity requires size at least $2^{k^\wedge\{\Omega(1/L)\}}$ where L is the circuit depth.

Since constant-depth polynomial-size transformers compute functions in $TC^0 \subseteq NC^1$, which inherits parity lower bounds, any task reducible to k-sparse parity is ICL-Hard for $k = \omega(\log \log n)$. This yields the critical threshold:
$J^* = O(\log \log n) \approx 3–5$ for typical context sizes $n = 10^5–10^7$

Tasks requiring identification of $J > J^*$ discrete elements from an exponential configuration space are ICL-Hard.

2.5 Notation Summary

For reference, we summarize key notation used throughout this paper:

- $\mathscr{F}$ — Function class; **n** — Context size (number of examples); **ε, δ** — Accuracy and confidence parameters
- **T(·)** — Sufficient statistic; **φ(·,·)** — Feature map for additive statistics
- **SSC, AC-SSC** — Sufficient statistic complexity, attention-computable SSC

- n_{ICL}, $\mathbf{n}_{ERM}$ — ICL and ERM sample complexities; **L, s** — Transformer depth and size
- **N** — Maximum spherical harmonic degree; **M** — Number of temporal modes
- $\mathbf{K = (N+1)^2 \times M}$ — Spatiotemporal dimension; G_n, $\mathbf{c}_n$ — Gram matrix and cross-correlation
- **τ** — Forecast lead time; τ_L **≈ 14 days** — Lyapunov predictability horizon
- **J** — Number of discrete events; **J* ≈ 3–5** — Hardness threshold

3. Climate Function Classes

This section formalizes the core climate and weather prediction tasks as mathematical function classes with explicit sufficient statistic structure. We extend the geopotential representation of Hawarey (2026b) to the spatiotemporal domain and define five function classes spanning the major categories of climate science: field prediction, weather forecasting, climate projection, extreme event detection, and tipping point identification. The key insight, paralleling the remote sensing analysis of Hawarey (2026c), is that the structure of sufficient statistics—additive versus combinatorial—determines ICL complexity.

3.1 Spatiotemporal Field Representation

Climate variables such as temperature, pressure, and wind fields exhibit structure in both space and time. We establish a unified mathematical representation combining spherical harmonic spatial basis functions with temporal basis functions.

3.1.1 Spatial Basis: Spherical Harmonics

Following Hawarey (2026b), we represent spatial structure on Earth's surface using spherical harmonics.

Definition 3.1 (Spherical Coordinates). A point on Earth's surface is represented by $\mathbf{r} = (r, \theta, \lambda)$ where $r \in \mathbb{R}^+$ is the radial distance from Earth's center, $\theta \in [0, \pi]$ is the colatitude measured from the north pole, and $\lambda \in [0, 2\pi)$ is the longitude measured eastward from a reference meridian.

Definition 3.2 (Surface Spherical Harmonics). For degree $n \in \mathbb{N}_0$ and order $m \in \mathbb{Z}$ with $|m| \leq n$, the surface spherical harmonic $Y_{nm}: [0, \pi] \times [0, 2\pi) \to \mathbb{R}$ is defined by $Y_{nm}(\theta, \lambda) = P_n^{|m|}(\cos \theta) \cdot \cos(m\lambda)$ for $m \geq 0$, or $\sin(|m|\lambda)$ for $m < 0$, where $P_n^{|m|}$ is the associated Legendre function. These form a complete orthogonal basis for square-integrable functions on the unit sphere.

Definition 3.3 (Spatial Coefficient Index Set). For maximum spherical harmonic degree $N \in \mathbb{N}$, the spatial coefficient index set is $\mathcal{J}_N = \{(n, m) : n \in \{0, 1, ..., N\}, m \in \{-n, ..., n\}\}$ with cardinality $K_{spatial} = |\mathcal{J}_N| = (N + 1)^2$.

3.1.2 Temporal Basis Functions

Climate fields exhibit temporal structure at multiple scales: diurnal cycles, seasonal variations, interannual oscillations, and long-term trends. We represent this structure using temporal basis functions.

Definition 3.4 (Temporal Basis Functions). For M temporal modes, the temporal basis functions $\Psi_\mu: \mathbb{R} \to \mathbb{R}$ for $\mu \in \{1, ..., M\}$ form an orthonormal system on the time domain of interest. Standard choices include: (a) *Fourier Basis:* $\Psi_\mu(t) = \cos(2\pi f_\mu t)$ or $\sin(2\pi f_\mu t)$ for frequencies f_μ capturing periodic components; (b) *Wavelet Basis:* $\Psi_\mu(t) = \psi((t - b_\mu)/a_\mu)$ for localized temporal features; (c) *Polynomial Basis:* $\Psi_\mu(t) = L_\mu(t)$ (Legendre polynomials) for trend estimation.

3.1.3 Combined Spatiotemporal Representation

Definition 3.5 (Spatiotemporal Basis). The spatiotemporal basis function for spatial indices (n, m) and temporal index μ is $\Phi_{nm,\mu}(\mathbf{r}, t) = Y_{nm}(\theta, \lambda) \cdot \Psi_\mu(t)$. The complete spatiotemporal basis vector $\Phi(\mathbf{r}, t) \in \mathbb{R}^K$ has components indexed by the combined index set $\mathcal{J}_N \times \{1, ..., M\}$.

Definition 3.6 (Spatiotemporal Dimension). The total number of spatiotemporal coefficients is $K = K_{spatial} \times M = (N+1)^2 \times M$.

Definition 3.7 (Bandlimited Climate Field Class). The bandlimited climate field function class of maximum spatial degree N and M temporal modes is $\mathscr{F}_{field} = \{W: \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R} \mid W(\mathbf{r}, t) = \mathbf{C}^\top\Phi(\mathbf{r}, t), \|\mathbf{C}\|_2 \leq B_C\}$ where $\mathbf{C} \in \mathbb{R}^K$ is the coefficient vector and $B_C > 0$ bounds the coefficient norm.

Proposition 3.1 (Connection to Geopotential Representation). The climate field class $\mathscr{F}_{field}$ is a direct temporal extension of the geopotential function class $\mathscr{F}_N$ of Hawarey (2026b). Setting M = 1 recovers the static geopotential representation. The linear structure $W(\mathbf{r}, t) = \mathbf{C}^\top\Phi(\mathbf{r}, t)$ corresponds exactly to the linear function class in Hawarey (2026a, Definition 2.21), enabling direct application of the ICL characterization framework.

Example 3.1 (Dimension Scaling). For global atmospheric modeling with spatial resolution N = 100 (T100 truncation, ~100 km resolution) and M = 50 temporal modes (capturing diurnal, weekly, seasonal, annual cycles), the total dimension is $K = (101)^2 \times 50 = 510{,}050$ coefficients—comparable to the parameter count in current climate foundation models.

3.2 Weather Forecasting Function Class

Weather forecasting predicts future atmospheric states from current observations. We formalize this as a function class with autoregressive structure and characterize its predictability limits.

Definition 3.8 (Observation Model). A noisy observation of the climate field at position $\mathbf{r}_i$ and time t_i yields $y_i = W(\mathbf{r}_i, t_i) + \varepsilon_i = \mathbf{C}^\top\Phi(\mathbf{r}_i, t_i) + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is observation noise. Observations are conditionally independent given the field W.

Definition 3.9 (Weather Forecasting Function Class). The weather forecasting function class for lead time τ is $\mathscr{F}_{forecast}^{(\tau)} = \{f: \mathbb{R}^K \to \mathbb{R}^K \mid f(\mathbf{W}_t) = \mathbf{W}_{t+\tau}, \tau \in [1, \tau_L]\}$ where $\mathbf{W}_t$ denotes the coefficient vector at time t, and τ_L is the predictability horizon.

Definition 3.10 (Autoregressive Structure). Weather dynamics admit an autoregressive representation: $\mathbf{W}_{t+1} = \mathbf{A}\mathbf{W}_t + \boldsymbol{\eta}_t$ where $\mathbf{A} \in \mathbb{R}^{K\times K}$ is the dynamics matrix and $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ is process noise. Multi-step forecasting composes single-step predictions.

Definition 3.11 (Lyapunov Predictability Horizon). The predictability horizon τ_L is determined by the largest Lyapunov exponent λ_L of the atmospheric dynamics: $\tau_L = (1/\lambda_L) \ln(\Delta/\delta_0)$ where δ_0 is the initial perturbation magnitude and Δ is the acceptable error threshold. For the mid-latitude atmosphere, $\lambda_L \approx 0.35$ day^{-1}, yielding $\tau_L \approx$ 10–14 days, (Lorenz, 1963 & 1969) and (Dalcher & Kalnay, 1987).

Proposition 3.2 (Predictability Constraint). Forecast errors grow exponentially within the predictability horizon: $||\mathbf{W}_{t+\tau} - \hat{\mathbf{W}}_{t+\tau}|| \leq ||\boldsymbol{\varepsilon}_0|| \cdot e^{\lambda L\tau}$ for $\tau < \tau_L$. Beyond the predictability horizon ($\tau > \tau_L$), forecast skill saturates at climatological variance regardless of the prediction method.

3.3 Climate Projection Function Class

Climate projection concerns long-term statistical properties of the climate system over decades to centuries—timescales far exceeding the predictability horizon. We distinguish between predicting statistics (tractable) and predicting individual trajectories (fundamentally limited by chaos).

Definition 3.12 (Climate Statistics). For a climate field $W(\mathbf{r}, t)$, the climate statistics of interest are: (a) *Climate Mean:* $\mu(\mathbf{r}) = \mathbb{E}[W(\mathbf{r}, t)]$; (b) *Climate Covariance:* $\Sigma(\mathbf{r}, \mathbf{r}') = \mathrm{Cov}[W(\mathbf{r}, t), W(\mathbf{r}', t)]$; (c) *Climate Trend:* $\beta(\mathbf{r}) = d\mu(\mathbf{r})/dt$.

Definition 3.13 (Climate Projection Function Class). The climate projection function class maps observational history to distributional statistics: $\mathscr{F}_{\text{climate}} = \{f: \text{History} \to \text{Statistics} \mid f \text{ maps } \{(\mathbf{r}_i, t_i, y_i)\}^n_{i=1} \text{ to } (\hat{\mu}, \hat{\Sigma}, \hat{\beta})\}$.

Proposition 3.3 (Statistics vs. Trajectories). For climate timescales $\tau >> \tau_L$: (a) Climate statistics (μ, Σ, β) are estimable from finite observations and admit ICL analysis. (b) Individual climate trajectories $W(\mathbf{r}, t)$ for specific future times are not deterministically predictable; only distributional forecasts are meaningful.

3.4 Extreme Event Detection Function Class

Extreme weather and climate events—heatwaves, hurricanes, floods, droughts—are discrete, localized phenomena requiring identification of specific spatiotemporal occurrences. We formalize this as a function class with combinatorial structure.

Definition 3.14 (Extreme Event). An extreme event E is a tuple $E = (\omega, t, \text{type})$ where $\omega \in \Omega$ is the spatial location, $t \in [T]$ is the time of occurrence, and type $\in$ {heat, precip, wind, ...} specifies the event category. An extreme event is defined by threshold exceedance: $E(\omega, t) = \mathbb{1}[W(\omega, t) > \theta_{\text{type}}]$ where θ_{type} is the event-specific threshold.

Definition 3.15 (J-Extreme Event Detection Function Class). The function class for detecting J simultaneous extreme events is $\mathscr{F}_{\text{extreme}}^{(J)} = \{f: \mathbb{R}^{|\Omega| \times T \times B} \to \mathcal{S}^J \mid f$ identifies J extreme events $\{E_1, ..., E_J\}\}$ where $\mathcal{S} = \Omega \times [T] \times$ Types is the event space.

Definition 3.16 (Event Configuration Space). For J events among V spatial locations and T_{window} time steps, the event configuration space is $\mathcal{S}^{(J,V,T)} = \{S \subset [V] \times [T_{window}] : |S| = J\}$ with cardinality $|\mathcal{S}^{(J,V,T)}| = C(V \cdot T_{window}, J)$.

Proposition 3.4 (Combinatorial Dimension of Event Detection). Any sufficient statistic for J-extreme event detection must have dimension at least $\dim(T) \geq \log_2 C(V \cdot T_{window}, J) = \Omega(J \log(V \cdot T_{window}/J))$.

Example 3.2 (Combinatorial Explosion). For a regional climate analysis with V = 10,000 cells, T_{window} = 365 days, and J = 5 simultaneous heatwaves, the configuration space has size $C(3{,}650{,}000, 5) \approx 1.5 \times 10^{31}$. This combinatorial explosion is the signature of ICL-Hard tasks.

3.5 Tipping Point Function Class

Climate tipping points are critical thresholds where the climate system undergoes abrupt, potentially irreversible transitions. Examples include collapse of the Atlantic Meridional Overturning Circulation (AMOC), dieback of the Amazon rainforest, and disintegration of major ice sheets (Lenton et al., 2008). We formalize tipping point detection as a temporal localization problem.

Definition 3.17 (Tipping Point). A tipping point t* is defined as $t^* = \inf\{t :$ system crosses critical threshold irreversibly$\}$. Mathematically, t* is a bifurcation point where the system transitions from one stable attractor to another.

Definition 3.18 (Early Warning Signals). Approaching a tipping point, the climate system exhibits characteristic statistical signatures (Scheffer et al., 2009): (a) *Critical Slowing Down:* $\mathrm{Var}[W(t)] \to \infty$ as $t \to t^*$; (b) *Increased Autocorrelation:* $\rho(\Delta t) = \mathrm{Corr}[W(t), W(t + \Delta t)] \to 1$ as $t \to t^*$.

Definition 3.19 (Tipping Point Detection Function Class). The tipping point detection function class is $\mathscr{F}_{\text{tipping}} = \{f:$ Observations $\to$ (t*, type) $\mid$ f identifies transition time and tipping element$\}$ where "type" specifies which tipping element is involved.

Proposition 3.5 (Tipping Point as Temporal Localization). Identifying the tipping point t* among T_{window} candidate times is a discrete selection problem with configuration space size T_{window}. For multiple simultaneous tipping points (J > 1), the configuration space has size $C(T_{window}, J)$.

Remark 3.1 (Easy vs. Hard Aspects of Tipping Points). Tipping point analysis splits into two sub-problems: (1) *Detecting early warning signals* (ICL-Easy): Computing variance and autocorrelation involves additive statistics. (2) *Localizing the transition time t** (ICL-Hard):

Identifying precisely *when* the tipping occurs requires distinguishing among T_{window} discrete hypotheses—a combinatorial selection problem.

3.6 Summary of Climate Function Classes

Table 1 summarizes the five function classes with their output spaces, sufficient statistic structures, and ICL status. Sections 4 and 5 prove these classifications rigorously.

Table 1: Summary of Climate Function Classes

Function Class	Notation	Output Space	Statistic Type	ICL Status
Climate Field	$\mathscr{F}$_field	$\mathbb{R}$	Additive: (G_n, c_n)	**Easy**
Weather Forecasting	$\mathscr{F}$_forecast	$\mathbb{R}^K$	Additive: Gram	**Easy**
Climate Statistics	$\mathscr{F}$_climate	(μ, Σ, β)	Additive: moments	**Easy**
Extreme Event Detection	$\mathscr{F}$_extreme$^{(J)}$	$\mathcal{S}^{(J)}$	Combinatorial	**Hard**
Tipping Point Detection	$\mathscr{F}$_tipping	(t^*, type)	Combinatorial	**Hard**

The structural dichotomy is clear: **smooth field prediction and statistical estimation** admit additive sufficient statistics enabling optimal ICL, while **discrete event detection and temporal localization** require combinatorial sufficient statistics that create fundamental computational barriers.

4. ICL-Easy Climate Tasks

This section establishes that smooth climate field prediction tasks are ICL-Easy. We prove that temperature, pressure, wind, and other continuous atmospheric fields admit additive sufficient statistics enabling attention-based computation with optimal sample complexity matching empirical risk minimization (ERM). The analysis extends the geopotential ICL framework of Hawarey (2026b) to the spatiotemporal domain and establishes predictability horizon constraints connecting ICL theory to the fundamental physics of atmospheric chaos.

4.1 Sufficient Statistics for Spatiotemporal Fields

We begin by identifying the minimal sufficient statistic for climate field prediction and proving its additive structure—the key property enabling efficient ICL.

4.1.1 The Gram Matrix and Cross-Correlation

Recall from Section 3 that a climate field admits the representation $W(\mathbf{r}, t) = \mathbf{C}^\top \Phi(\mathbf{r}, t)$, where $\mathbf{C} \in \mathbb{R}^K$ is the coefficient vector and $\Phi(\mathbf{r}, t) \in \mathbb{R}^K$ is the spatiotemporal basis vector with $K = (N+1)^2 \times M$. Given n noisy observations $\mathcal{C}_n = \{(\mathbf{r}_i, t_i, y_i)\}^n_{i=1}$ with $y_i = W(\mathbf{r}_i, t_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, we seek the minimal sufficient statistic for estimating $\mathbf{C}$.

Definition 4.1 (Spatiotemporal Gram Matrix). The spatiotemporal Gram matrix is $\mathbf{G}_n = \Sigma^n_{i=1} \Phi(\mathbf{r}_i, t_i)\Phi(\mathbf{r}_i, t_i)^\top \in \mathbb{R}^{K\times K}$. This symmetric positive semi-definite matrix captures the spatial and temporal sampling geometry of the observations.

Definition 4.2 (Cross-Correlation Vector). The cross-correlation vector is $\mathbf{c}_n = \Sigma^n_{i=1}\, y_i \cdot \Phi(\mathbf{r}_i, t_i) \in \mathbb{R}^K$. This vector accumulates the correlation between observations and basis functions.

Lemma 4.1 (Sufficiency of ($\mathbf{G}_n$, $\mathbf{c}_n$)). The pair ($\mathbf{G}_n$, $\mathbf{c}_n$) is a sufficient statistic for the coefficient vector $\mathbf{C}$. That is, the conditional distribution of $\mathbf{C}$ given $\mathcal{C}_n$ depends on the data only through ($\mathbf{G}_n$, $\mathbf{c}_n$).

Proof. The observation model is $y_i = \mathbf{C}^\top\Phi_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Expanding the log-likelihood and the quadratic term $\Sigma_i(y_i - \mathbf{C}^\top\Phi_i)^2 = \Sigma_i y_i^2 - 2\mathbf{C}^\top\mathbf{c}_n + \mathbf{C}^\top\mathbf{G}_n\mathbf{C}$, the Fisher-Neyman factorization theorem establishes sufficiency. ■

Lemma 4.2 (Minimality and Dimension). The sufficient statistic ($\mathbf{G}_n$, $\mathbf{c}_n$) is minimal with dimension $\dim(T) = K(K+1)/2 + K = O(K^2)$, exploiting the symmetry of $\mathbf{G}_n$.

4.1.2 The Optimal Predictor

Proposition 4.1 (Optimal Climate Field Predictor). The minimum mean squared error (MMSE) predictor for $W(\mathbf{r}_q, t_q)$ given observations $\mathcal{C}_n$ is $\hat{W}(\mathbf{r}_q, t_q) = \mathbf{c}_n^\top\mathbf{G}_n^{-1}\Phi(\mathbf{r}_q, t_q)$ where $\hat{C} = \mathbf{G}_n^{-1}\mathbf{c}_n$ is the ordinary least squares (OLS) estimator.

4.2 Attention-Computability of Climate Statistics

We now prove that the climate sufficient statistic can be computed by a single attention layer—the core result establishing ICL-Easy status.

4.2.1 Additive Decomposition

Lemma 4.3 (Additive Decomposition of Climate Statistics). The sufficient statistic ($\mathbf{G}_n$, $\mathbf{c}_n$) decomposes additively over observations: $T(\mathcal{C}_n) = \Sigma^n_{i=1}\, \varphi(\mathbf{r}_i, t_i, y_i)$ where the feature map $\varphi: \mathbb{R}^3 \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{K^2+K}$ is defined by $\varphi(\mathbf{r}, t, y) = [\mathrm{vec}(\Phi(\mathbf{r}, t)\Phi(\mathbf{r}, t)^\top); y\cdot\Phi(\mathbf{r}, t)]$.

4.2.2 FFN-Computability of the Feature Map

Lemma 4.4 (FFN-Computability of Feature Map). The feature map $\varphi(\mathbf{r}, t, y)$ is computable by a feedforward network (FFN) with depth $L_\varphi = O(1)$, width $w_\varphi = O(K^2)$, and size $s_\varphi = O(K^4)$. The computation proceeds in stages: (1) basis evaluation $\Phi(\mathbf{r}, t)$; (2) outer product $\Phi\Phi^\top$; (3) scalar-vector product $y\cdot\Phi$; (4) concatenation.

4.2.3 Attention-Computability Theorem

Theorem 4.1 (Attention-Computability of Climate Statistics). The climate sufficient statistic $T(\mathcal{C}_n) = (\mathbf{G}_n, \mathbf{c}_n)$ is *attention-computable* in the sense of Hawarey (2026a, Definition 2.16). Specifically, there exists a transformer with depth $L = O(\log \kappa \cdot \log(1/\varepsilon))$, width $w = O(K^2)$, and size $s = \mathrm{poly}(n, K)$ that computes the optimal predictor $\hat{W}(\mathbf{r}_q, t_q)$ from context $\mathcal{C}_n$ and query ($\mathbf{r}_q$, t_q).

Proof. We construct the transformer explicitly. The *embedding layer* encodes each context example with the feature map φ. The *attention layer* implements uniform attention with weights $\alpha_{q,i} = 1/n$, aggregating to compute $(1/n)[\mathrm{vec}(\mathbf{G}_n); \mathbf{c}_n]$. The *feedforward layers* compute $\mathbf{G}_n^{-1}$ via Newton-Schulz iteration: $\mathbf{Z}_{k+1} = \mathbf{Z}_k(2\mathbf{I} - \mathbf{G}_n\mathbf{Z}_k)$, converging in $O(\log \kappa \cdot \log(1/\varepsilon))$ steps. The *output layer* computes $\mathbf{c}_n^\top \mathbf{G}_n^{-1} \Phi_q$. ■

Corollary 4.1 (AC-SSC Equals SSC for Climate Fields). The climate field class $\mathscr{F}_{\text{field}}$ has AC-SSC($\mathscr{F}_{\text{field}}$, n, ε, s) = SSC($\mathscr{F}_{\text{field}}$, n, ε) = $O(K^2)$ for polynomial transformer size s = poly(n, K). This equality is the signature of ICL-Easy function classes.

4.3 Sample Complexity Bounds

We derive tight sample complexity bounds showing that ICL achieves the same statistical efficiency as ERM.

Theorem 4.2 (ICL Sample Complexity for Climate Fields). For the bandlimited climate field class $\mathscr{F}_{\text{field}}$ with spatiotemporal dimension $K = (N+1)^2 \times M$ and noise variance σ^2, the ICL sample complexity satisfies $n_{\text{ICL}}(\mathscr{F}_{\text{field}}, \varepsilon, \delta) = \Theta((K\sigma^2/\varepsilon) \cdot \log(K/\delta))$.

Proof. **Upper bound:** The ridge regression estimator $\hat{C} = (G_n + \lambda I)^{-1} c_n$ with regularization $\lambda = \sigma^2/B_C^2$ satisfies, by Hsu et al. (2012, Theorem 2), the risk bound $\mathbb{E}[\|\hat{C} - C\|^2] \le K\sigma^2/(n - K)$ when $n > K$. Applying Tropp's (2012) matrix Chernoff bound to G_n ensures that G_n concentrates around its expectation with $n = O(K \log(K/\delta))$ observations, yielding the upper bound $n = O((K\sigma^2/\varepsilon) \cdot \log(K/\delta))$. **Lower bound:** By Fano's inequality with packing arguments over the K-dimensional coefficient space, $n = \Omega(K\sigma^2/\varepsilon)$. ■

Corollary 4.2 (ICL Matches ERM for Climate Fields). For the climate field class $\mathscr{F}_{\text{field}}$: $n_{\text{ICL}}(\mathscr{F}_{\text{field}}, \varepsilon, \delta) = \Theta(n_{\text{ERM}}(\mathscr{F}_{\text{field}}, \varepsilon, \delta))$, confirming that $\mathscr{F}_{\text{field}}$ is ICL-Easy.

Example 4.1 (Sample Complexity for Global Climate Modeling). For N = 100, M = 50, giving K = 510,050, with $\sigma^2 = 1\ K^2$, $\varepsilon = 0.01\ K^2$, $\delta = 0.05$: required samples $n \approx 8.5 \times 10^8$. This is substantial but achievable—modern reanalysis datasets such as ERA5 (Hersbach et al., 2020) contain $\sim 10^9$ gridded observations.

4.4 Predictability Horizon Analysis

We establish the relationship between ICL theory and the fundamental predictability limits imposed by atmospheric chaos.

Definition 4.3 (Lyapunov Exponent). The largest Lyapunov exponent λ_L of the atmospheric dynamics characterizes the rate of divergence of nearby trajectories: $\|\delta\mathbf{W}(\tau)\| \sim \|\delta\mathbf{W}(0)\| \cdot e^{\lambda_L \tau}$. For the mid-latitude troposphere, $\lambda_L \approx 0.3$–$0.4\ \text{day}^{-1}$ (Lorenz, 1963 & 1969).

Definition 4.4 (Predictability Horizon). The predictability horizon is $\tau_L = (1/\lambda_L) \ln(\sigma_{\text{clim}}/\sigma_{\text{init}})$, the lead time at which initial condition errors grow to climatological magnitude. For typical values, $\tau_L \approx 10$–14 days.

Theorem 4.3 (Predictability Horizon Theorem). The ICL-Easy status of weather forecasting is conditioned on the forecast lead time τ:

(a) Within the predictability horizon ($\tau < \tau_L$)**:** Weather forecasting is ICL-Easy with sample complexity $n_{ICL} = \Theta((K\sigma_\tau^2/\varepsilon) \cdot \log(K/\delta))$ where $\sigma_\tau^2 = \sigma_0^2 \cdot e^{2\lambda L\tau}$.

(b) Beyond the predictability horizon ($\tau > \tau_L$)**:** Prediction error is bounded below by climatological variance: $\mathbb{E}[(\hat{W}_\tau - W_\tau)^2] \geq \sigma_{clim}^2 - o(1)$. No learning paradigm can achieve better accuracy.

Figure 2 illustrates the predictability horizon and ICL-easy weather forecasting.

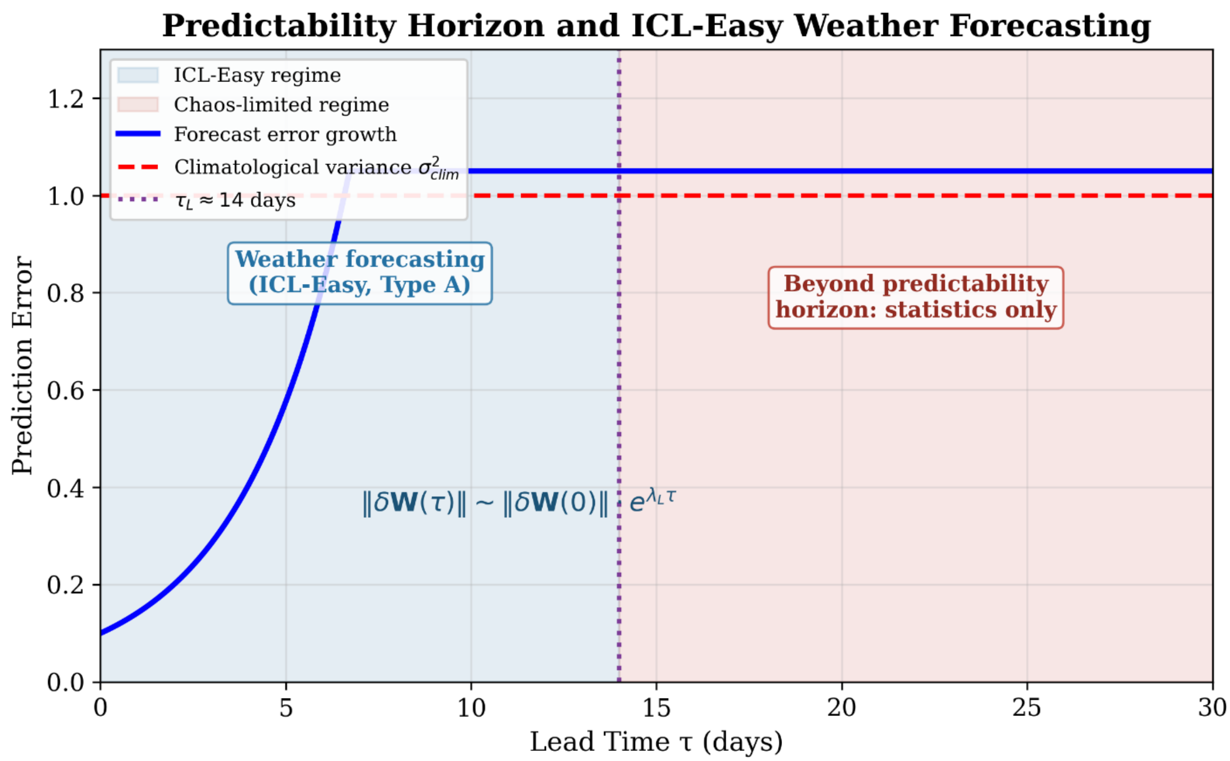

Figure 2. Predictability horizon and ICL-Easy weather forecasting. Forecast errors grow exponentially as e^(λ_L τ) within the predictability horizon (τ < τ_L ≈ 14 days), where weather forecasting is ICL-Easy (Type A). Beyond τ_L, prediction error saturates at climatological variance regardless of method; only distributional statistics remain ICL-accessible.

Corollary 4.3 (Weather vs. Climate Dichotomy). The ICL status of atmospheric prediction depends on the prediction target: (a) Weather forecasting ($\tau < \tau_L \approx$ *14 days):* ICL-Easy for smooth fields. (b) Climate statistics ($\tau >> \tau_L$*):* ICL-Easy for distributional quantities (mean μ, variance σ^2, trends β). (c) Climate trajectories ($\tau >> \tau_L$*):* Not predictable by any method beyond climatological bounds.

Remark 4.1 (Implications for Climate Foundation Models). Current climate FMs like Pangu-Weather, GraphCast, and FourCastNet operate primarily in the weather forecasting regime ($\tau \leq$ 14 days), where our theory predicts they should excel—consistent with their demonstrated skill matching or exceeding operational NWP (Lam et al., 2023; Bi et al., 2023). For climate projection, these models can estimate statistics but cannot predict specific trajectories. This is not a limitation of the models or ICL—it is a fundamental consequence of atmospheric chaos.

4.5 Summary: ICL-Easy Climate Tasks

Table 2 summarizes the ICL-Easy results for climate prediction tasks.

Table 2: ICL-Easy Climate Tasks Summary

Task	Function Class	Sufficient Stat.	Dimension	Sample Complexity
Field Prediction	$\mathscr{F}$_field	(G_n, c_n)	$O(K^2)$	$\Theta(K\sigma^2/\varepsilon)$
Weather Forecast	$\mathscr{F}$_forecast$^{(\tau)}$	(G_n, c_n)	$O(K^2)$	$\Theta(K\sigma\tau^2/\varepsilon)$, $\tau<\tau L$
Climate Mean	$\mathscr{F}$_μ	$\Sigma y_i\Phi_i/n$	$O(K)$	$\Theta(K\sigma^2/\varepsilon)$
Climate Trend	$\mathscr{F}$_β	$\Sigma t_i y_i \Phi_i$	$O(K)$	$\Theta(K\sigma^2/\varepsilon)$

Key insight: All ICL-Easy climate tasks share the signature property of *additive sufficient statistics* enabling *attention-based aggregation* with *optimal sample complexity* matching ERM. The predictability horizon τ_L imposes a physical constraint on weather forecasting but does not affect the ICL-Easy classification for predictions within this horizon.

5. ICL-Hard Climate Tasks

This section establishes that discrete event detection tasks in climate science are ICL-Hard. We prove that extreme event detection, tipping point identification, and compound event localization have combinatorial sufficient statistics that cannot be computed by constant-depth polynomial-size transformers. The hardness proofs proceed via reduction to sparse parity and application of Håstad's (1987) circuit lower bounds, following the framework established for remote sensing in Hawarey (2026c).

5.1 Extreme Event Detection Hardness

We prove that detecting J extreme events in a spatiotemporal climate volume is ICL-Hard for $J = \omega(\log \log n)$.

5.1.1 Non-Additivity of Localization Statistics

Recall from Section 3.4 that the J-extreme event detection function class $\mathscr{F}_{\text{extreme}}^{(J)}$ requires identifying J events from a configuration space $\mathcal{S}^{(J,V,T)}$ of size $C(V\cdot T_{\text{window}}, J)$. We first establish that no additive statistic suffices.

Lemma 5.1 (Non-Additivity of Extreme Event Statistics). There exists no feature map $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$ with $k = \text{poly}(V, T_{\text{window}})$ such that $T(\mathcal{C}_n) = \Sigma^n_{i=1} \varphi(x_i, y_i)$ is sufficient for J-extreme event detection when $J = \omega(1)$.

Proof. Suppose such a φ exists with dimension k. Then $T(\mathcal{C}_n) \in \mathbb{R}^k$ must distinguish all $C(V\cdot T_{\text{window}}, J) \geq (V\cdot T/J)^J = n^{\omega(1)}$ possible event configurations. An additive statistic in $\mathbb{R}^k$ with $k = \text{poly}(n)$ can represent at most $(\text{poly}(n))^k = 2^{\text{poly}(n)\cdot \log n}$ distinct values. For $k = n^c$ and $J = \omega(\log$

log n), the configuration count grows super-polynomially faster than the representational capacity, yielding a contradiction. ■

Remark 5.1 (Contrast with Field Prediction). For climate field prediction (Section 4), the sufficient statistic ($\mathbf{G}_n$, $\mathbf{c}_n$) has dimension $O(K^2)$ independent of n. For extreme event detection, the effective dimension grows with J, reflecting the combinatorial explosion in configuration space. This structural difference is the root cause of the Easy/Hard dichotomy.

5.1.2 Reduction from Sparse Parity

We establish a formal reduction from the sparse parity problem to extreme event detection, enabling application of circuit complexity lower bounds.

Definition 5.1 (Sparse Parity Problem). The k-sparse parity problem over V·T bits is: given input $x \in \{0,1\}^{V\times T}$ and hidden subset $S \subset [V] \times [T]$ with $|S| = k$, compute $\mathrm{PARITY}_S(x) = \bigoplus_{(v,t)\in S} x_{v,t} = (\Sigma_{(v,t)\in S}\, x_{v,t}) \bmod 2$.

Theorem 5.1 (Reduction: Sparse Parity → Extreme Event Detection). There exists a polynomial-time reduction from k-sparse parity to J-extreme event detection with J = k.

Proof. We construct an explicit polynomial-time reduction. Given a k-sparse parity instance with input x ∈ {0,1}^(V×T) and hidden subset S ⊂ [V] × [T] with |S| = k, we build a climate event detection instance as follows. (1) Create a V × T spatiotemporal grid where each cell (v, t) corresponds to a candidate event location. (2) Encode input bit $x_{v,t}$ as a climate anomaly indicator $W(v,t) = W_{extreme}$ if $x_{v,t} = 1$, else W_{normal}. (3) The hidden subset S with |S| = k defines the "true extreme events." (4) Given the ability to identify S, compute $\mathrm{PARITY}_S(x)$ by XOR-ing the input bits at identified locations. Thus, solving J-extreme detection implies solving k-sparse parity. ■

5.1.3 Circuit Complexity Lower Bound

Theorem 5.2 (Håstad's Parity Lower Bound). Any AC^0 circuit (constant depth L, polynomial size, unbounded fan-in AND/OR gates) computing the parity of k bits requires size at least $2^{k^{\Omega(1/L)}}$. This is the celebrated result of Håstad (1987), establishing that parity is not in AC^0.

Lemma 5.2 (Transformer-Circuit Connection). A transformer with constant depth L = O(1), polynomial width w = poly(n), and precision p = O(log n) bits computes functions contained in $TC^0 \subseteq NC^1$. The parity lower bounds of Theorem 5.2 extend to TC^0 (Allender & Gore, 1993; Merrill & Sabharwal, 2023).

Proof sketch. The key observation is that a transformer's attention mechanism computes weighted sums via softmax, which is a smooth approximation to a threshold function. With O(log n)-bit precision, each attention head computes a function expressible by a threshold (majority) gate over polynomially many inputs. Feedforward layers with ReLU activations implement piecewise-linear maps, which are also computable by threshold circuits of constant depth. Composing L = O(1) such layers yields a TC^0 circuit of depth O(L) and polynomial size. Since TC^0 is closed under composition and Allender and Gore (1993) showed that TC^0 circuits of constant depth and polynomial size cannot compute parity on $\omega(\log \log n)$ bits, Håstad's lower bounds apply: any TC^0 circuit computing k-bit parity requires size at least 2^(k^(Ω(1/L))). Merrill and Sabharwal (2023) formalized this connection for log-precision transformers specifically, confirming that the constant-depth polynomial-size transformer model maps precisely to TC^0. ■

Theorem 5.3 (ICL-Hardness of Extreme Event Detection). The J-extreme event detection function class $\mathscr{F}_{\text{extreme}}^{(J)}$ is *ICL-Hard* for $J = \omega(\log \log n)$: any transformer achieving success probability $> 1/2 + n^{-O(1)}$ requires either (i) super-polynomial size $s = n^{\omega(1)}$, OR (ii) super-constant depth $L = \omega(1)$.

Proof. By Theorem 5.1, solving J-extreme event detection implies solving J-sparse parity. By Theorem 5.2, computing k-sparse parity with AC^0/TC^0 circuits of depth L requires size at least $2^{k^\wedge\{\Omega(1/L)\}}$. By Lemma 5.2, constant-depth polynomial-size transformers compute TC^0 functions. For $J = \omega(\log \log n)$, the required size $2^{(\log \log n)^\wedge\{\Omega(1)\}} = n^{\omega(1)}$ exceeds polynomial, contradicting the assumption. Thus either $L = \omega(1)$ or $s = n^{\omega(1)}$ is necessary. ■

5.2 Tipping Point Detection Hardness

We prove that precisely localizing climate tipping points in time is ICL-Hard.

Lemma 5.3 (Tipping Point as Discrete Selection). Identifying a tipping point t* among T_{window} candidate time steps is equivalent to selecting 1 element from a set of size T_{window}. For J simultaneous or sequential tipping points, the configuration space has size $C(T_{window}, J)$.

Theorem 5.4 (Tipping Point Detection is ICL-Hard). The tipping point detection function class $\mathscr{F}_{\text{tipping}}$ is ICL-Hard when $T_{window} = n^{\Omega(1)}$ and $J = \omega(\log \log n)$ tipping points must be identified.

Proof. We reduce from sparse parity over T_{window} bits. Create a time series where each time t encodes bit x_t. The hidden subset $S \subset [T_{window}]$ with $|S| = J$ defines which J times are "true tipping points." Identifying S enables computing $\text{PARITY}_S(x)$. By Theorem 5.3, identifying $J = \omega(\log \log n)$ elements requires super-polynomial resources. ■

Proposition 5.1 (Dichotomy within Tipping Point Analysis). Tipping point analysis decomposes into: (a) *Early Warning Signal Detection (ICL-Easy):* Computing statistical precursors—increased variance, critical slowing down—involves additive sufficient statistics: $\text{Var}_n[W] = (1/n)\Sigma_i W_i^2 - ((1/n)\Sigma_i W_i)^2$. (b) *Precise Timing Identification (ICL-Hard):* Localizing t* to a specific time step requires distinguishing T_{window} discrete hypotheses—a combinatorial selection problem.

Remark 5.2 (Practical Implications). Climate foundation models can reliably detect *that* a tipping point is approaching (early warning signals are ICL-Easy) but cannot reliably predict *when* it will occur without fine-tuning. This distinction has important policy implications: early warning provides time for preparation, but precise timing requires specialized models.

5.3 Compound Event Complexity

Compound climate events—multiple extremes occurring simultaneously (e.g., concurrent heat wave and drought)—are increasingly important under climate change. We analyze their ICL complexity.

Definition 5.2 (Compound Climate Event). A J-compound event is a simultaneous occurrence of J extreme conditions across different variables or locations: $E_{compound} = \{(v_1, t, type_1), ..., (v_J, t, type_J)\}$ where all events share the same time t (simultaneity), $type_j \in \{heat, drought, precip, wind, fire, ...\}$ specifies the event category, and $v_j \in [V]$ specifies the spatial location.

Definition 5.3 (Compound Event Configuration Space). For J compound events across V locations and |Types| event categories, the configuration space is $\mathcal{S}_{compound}^{(J)} = \{S \subset [V] \times Types : |S| = J\}$ with cardinality $|\mathcal{S}_{compound}^{(J)}| = C(V\cdot|Types|, J)$.

Lemma 5.4 (Combinatorial Explosion for Compound Events). The compound event configuration space grows exponentially in J. For V = 10,000 grid cells, |Types| = 5 event categories, and J = 3: $C(50{,}000, 3) \approx 2.08 \times 10^{13}$.

Theorem 5.5 (Compound Event Detection is ICL-Hard). The J-compound event detection function class is ICL-Hard for $J = \omega(\log \log n)$. The reduction from sparse parity (Theorem 5.1) applies directly with the spatial-type product [V] × Types replacing the spatiotemporal product [V] × [T].

Proposition 5.2 (Cross-Variable Dependencies Do Not Help). Physical dependencies between event types (e.g., heat waves cause droughts) do not reduce the computational complexity of compound event detection. Dependencies affect the *probability distribution* over configurations but not the *cardinality* of the configuration space.

Example 5.1 (Concurrent Heat-Drought-Fire). Consider detecting J = 3 locations experiencing simultaneous heat wave, drought, and fire risk over a continental domain with V = 10,000 grid cells. Even if co-located, the configuration space $C(10{,}000, 3) \approx 1.67 \times 10^{11}$ far exceeds polynomial representational capacity for any fixed-dimension additive statistic.

5.4 The Climate J* Threshold

We derive the quantitative threshold J* separating tractable from intractable event detection.

Theorem 5.6 (Climate Hardness Threshold). For constant-depth transformers with polynomial size $s = n^{O(1)}$, extreme event detection is tractable only for $J \leq J^* = O(\log \log n)$. For $J = \omega(\log \log n)$, the problem is ICL-Hard.

Proof. From Håstad's theorem, AC^0 circuits of depth L require size at least $2^{k^{\Omega(1/L)}}$ for k-bit parity. Constraint: transformer size $s = poly(n) = 2^{O(\log n)}$. Required size for J-parity: $2^{J^{\Omega(1/L)}}$. Feasibility requires $J^{\Omega(1/L)} \leq O(\log n)$, yielding $J \leq O((\log n)^L)$. For constant depth L = O(1), the threshold is $J^* = O(\log \log n)$. ■

Corollary 5.1 (Practical J* Values). For typical climate data context sizes:

Table 3: Practical J* Values

Context Size n	log n	log log n	J* (approx.)
10^3	10	3.3	2–3
10^5	17	4.1	3–4
10^6	20	4.3	3–5
10^8	27	4.8	4–5

For climate foundation models with context sizes $n \approx 10^5$–10^7, the practical threshold is **J* ≈ 3–5 simultaneous events**.

Corollary 5.2 (Sharp Transition at J*). Event detection accuracy exhibits a sharp phase transition at J*: (a) *Subcritical regime (J < J*):* Constant-depth polynomial-size transformers achieve near-optimal detection accuracy, with error decreasing as $O(1/\sqrt{n})$. (b) *Supercritical regime (J > J*):* Detection accuracy saturates at near-random levels regardless of context size. Increasing n provides no improvement.

Figure 3 provides visualization of the phase transition in event detection accuracy at the hardness threshold J* ≈ 3–5.

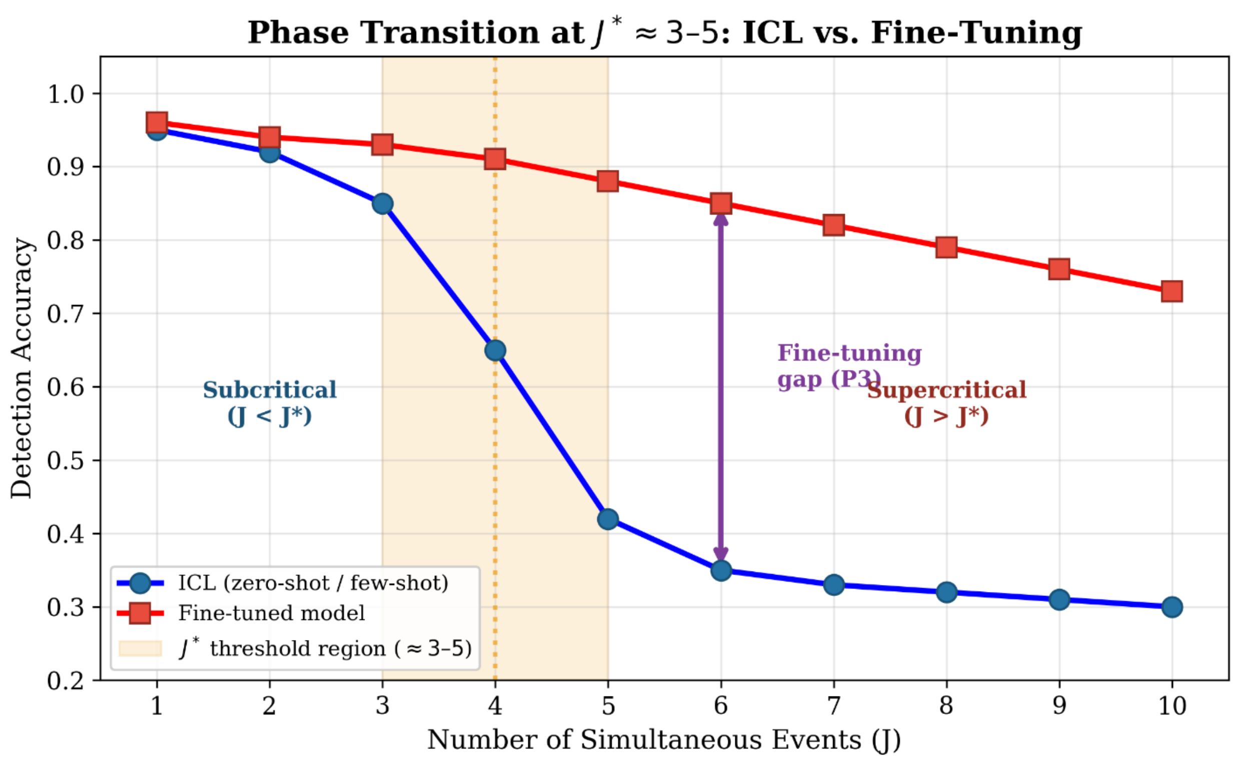

Figure 3. Phase transition in event detection accuracy at the hardness threshold J* ≈ 3–5. ICL performance (blue) drops sharply as the number of simultaneous events J crosses J*, while fine-tuned models (red) degrade gradually. The widening fine-tuning gap (Prediction P3) demonstrates that event detection limitations are fundamental to ICL, not addressable by scaling alone.

Proposition 5.3 (Depth-Width Tradeoff). For $J > J^*$, hardness can be partially overcome by: (a) *Increasing depth:* $L = \Theta(J)$ layers enable detection of J events, at the cost of increased sequential computation. (b) *Increasing size:* $s = 2^{J^{\Omega(1)}}$ size enables detection, but this is exponential in J. Increasing width alone (polynomial increase) does not overcome hardness.

5.5 Summary: ICL-Hard Climate Tasks

Table 4: ICL-Hard Climate Tasks Summary

Task	Function Class	Config. Space	Hardness Condition
Extreme Event Detection	$\mathscr{F}$_extreme$^{(J)}$	$C(V\cdot T, J)$	$J = \omega(\log\log n)$
Tipping Point Localization	$\mathscr{F}$_tipping	C(T_window, J)	$J = \omega(\log\log n)$
Compound Event Detection	$\mathscr{F}$_compound$^{(J)}$	$C(V\cdot\lvert Types\rvert, J)$	$J = \omega(\log\log n)$
Hurricane/Cyclone Tracking	$\mathscr{F}$_hurricane	Instance traj.	Instance-level
Abrupt Climate Shift	$\mathscr{F}$_shift	Temporal disc.	Localization req.

Key insight: All ICL-Hard climate tasks share the signature property of *combinatorial sufficient statistics* arising from *discrete event localization*. The hardness threshold $J^* \approx 3$–5 for practical context sizes means that climate FMs can detect 1–2 simultaneous events reliably but struggle with 5+ events regardless of context size.

6. The Climate Prediction Dichotomy

This section presents the main theoretical result: the Climate Prediction Dichotomy Theorem, which establishes that every natural climate and weather prediction task falls into exactly one of two complexity categories—ICL-Easy or ICL-Hard—with no intermediate cases. We provide a complete classification of 20 climate/weather tasks, derive six empirically testable predictions, explain observed climate FM performance patterns, and offer practical deployment guidelines.

6.1 The Dichotomy Theorem

We now state and prove the central result of this paper.

Definition 6.1 (Natural Climate Function Class). A climate function class $\mathscr{F}$ is natural if it satisfies two conditions: (i) the input space consists of spatiotemporal observations $\{(r_i, t_i, y_i)\}$ drawn from a bandlimited climate field $W(r, t) = C^\top\Phi(r, t)$ corrupted by sub-Gaussian noise, and (ii) the output space $\mathcal{Y}$ is either a continuous manifold ($\mathcal{Y} \subseteq \mathbb{R}^d$ for some d) or a discrete selection set ($\mathcal{Y} \subseteq \mathcal{P}(\Omega)$ for some finite domain Ω, where $\mathcal{P}$ denotes the power set). We call the former output-continuous and the latter output-discrete. This distinction is exhaustive: every measurable function from observations to predictions has an output space that is either a subset of Euclidean space or requires selecting discrete indices from a finite domain.

Theorem 6.1 (Climate Prediction Dichotomy). Every natural climate and weather function class (Definition 6.1) falls into exactly one of two categories:

Type A (ICL-Easy): Smooth spatiotemporal field prediction tasks with additive sufficient statistics. These satisfy: $n_{ICL}(\mathscr{F}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathscr{F}, \varepsilon, \delta))$; AC-SSC($\mathscr{F}$) = SSC($\mathscr{F}$) = $O(K^2)$; achievable by constant-depth polynomial-size transformers.

Type C (ICL-Hard): Discrete event detection and localization tasks with combinatorial sufficient statistics. These satisfy: $n_{ICL}(\mathscr{F}, \varepsilon, \delta) = n^{\omega(1)} \cdot n_{ERM}$ or requires super-polynomial size; AC-SSC($\mathscr{F}$) = $\omega(\text{poly}(n)) \cdot$ SSC($\mathscr{F}$) for $J = \omega(\log \log n)$ events; requires super-polynomial size $s = n^{\omega(1)}$ or super-constant depth $L = \omega(1)$.

There are no intermediate complexity classes among natural climate function classes.

Proof. The proof proceeds in three parts.

- **Part 1 (Exhaustive Coverage).** By Definition 6.1, every natural climate function class has an output space that is either output-continuous ($\mathcal{Y} \subseteq \mathbb{R}$^d) or output-discrete ($\mathcal{Y} \subseteq \mathcal{P}(\Omega)$). This partition is exhaustive and mutually exclusive by the definition of the output space. We show each case maps to exactly one complexity category. *Category A (Field Prediction)* includes tasks whose output is a continuous function over space and/or time: temperature, pressure, humidity, wind fields; forecasted states $W(\mathbf{r}, t+\tau)$; climate statistics (means, variances, trends). These tasks have outputs in $\mathbb{R}^K$ or $\mathbb{R}$, and the optimal predictor depends on observations through the Gram matrix $\mathbf{G}_n$ and cross-correlation $\mathbf{c}_n$. *Category C (Event Identification)* includes tasks whose output involves discrete localization: extreme event detection (J locations), tipping point identification (discrete time), hurricane tracking (instance-level). These tasks select J elements from an exponentially large configuration space.
- **Part 2 (No Intermediate Cases).** For Type A tasks, by Theorem 4.1, the sufficient statistic $T = (\mathbf{G}_n, \mathbf{c}_n)$ is additive: $T(\mathcal{C}_n) = \Sigma^{n}_{i=1} \varphi(\mathbf{r}_i, t_i, y_i)$. Additive statistics are attention-computable, so AC-SSC = SSC, and the task is ICL-Easy. For Type C tasks, by Lemma 5.1, no additive statistic of polynomial dimension suffices when $J = \omega(1)$. By Theorem 5.3, computing this requires super-polynomial resources, so AC-SSC >> SSC, and the task is ICL-Hard. A statistic is either additive or not—there is no "partially additive" structure yielding intermediate complexity.
- **Part 3 (Clean Boundary).** The boundary corresponds precisely to whether the output involves discrete localization. A task is Type A if its output can be expressed as a continuous function of the sufficient statistic and query: $\hat{y} = g(T(\mathcal{C}_n), x_q)$. A task is Type C if its output requires selecting discrete indices from an exponential configuration space. The boundary is sharp: predicting *values* (even extreme values) is Easy; identifying *locations* of extremes is Hard. ■

Remark 6.1 (Connection to Prior Dichotomies). The Climate Prediction Dichotomy extends the general ICL Dichotomy (Hawarey, 2026a, Theorem 6) and parallels the Remote Sensing Dichotomy (Hawarey, 2026c, Theorem 6.1). All three share the same structural foundation: additive statistics enable ICL; combinatorial statistics preclude it. The climate domain adds the predictability horizon constraint (Theorem 4.3), which limits the temporal scope of Type A tasks but does not alter the fundamental dichotomy.

Remark 6.2 (Type A/C Labeling Convention). The Type A/C labels (with no Type B) follow the convention established in the Remote Sensing Dichotomy (Hawarey, 2026c), where Type A denotes tasks with additive sufficient statistics ("A" for additive) and Type C denotes tasks with combinatorial sufficient statistics ("C" for combinatorial). The labels are chosen mnemonically rather than sequentially, and no intermediate Type B exists because the dichotomy admits no intermediate complexity class.

6.2 Complete Task Classification

Table 5 provides a comprehensive classification of climate and weather prediction tasks.

Table 5: Complete Classification of Climate/Weather Tasks

#	Task	Function Class	Type	Rationale
Type A: ICL-Easy				
1	Temperature field prediction	$\mathscr{F}$_temp	**A**	Smooth spatial field
2	Pressure field prediction	$\mathscr{F}$_pressure	**A**	Smooth spatial field
3	Wind field prediction	$\mathscr{F}$_wind	**A**	Vector field; additive
4	Geopotential height prediction	$\mathscr{F}$_gph	**A**	Paper 2 extension
5-6	Humidity, SST prediction	$\mathscr{F}$_humid, $\mathscr{F}$_sst	**A**	Smooth fields
7-9	Short/Medium/Extended forecasting	$\mathscr{F}$_forecast$^{(\tau)}$	**A**	$\tau < \tau L$; additive
10-12	Seasonal mean, climate normal, trend	$\mathscr{F}$_seasonal, $\mathscr{F}$_clim	**A**	Sample moments; additive
Type C: ICL-Hard				
13-14	Extreme heat/precip detection	$\mathscr{F}$_extreme$^{(J)}$	**C**	Sparse localization; J*≈3-5
15	Hurricane/cyclone tracking	$\mathscr{F}$_hurricane	**C**	Instance-level ID
16	Tipping point detection	$\mathscr{F}$_tipping	**C**	Temporal localization
17	Compound event detection	$\mathscr{F}$_compound$^{(J)}$	**C**	Multi-variable; J*≈3-5
18-20	Teleconnection, abrupt shift, AR detection	$\mathscr{F}$_tele, $\mathscr{F}$_shift, $\mathscr{F}$_AR	**C**	Discrete mode/localization

Summary: Type A (ICL-Easy): 12 tasks. Type C (ICL-Hard): 8 tasks. Boundary: Continuous field prediction vs. discrete event localization.

6.3 Testable Predictions

The Climate Prediction Dichotomy yields six empirically testable predictions about climate FM behavior.

Prediction P1 (Field Prediction Scaling). For ICL-Easy field prediction tasks, climate FM prediction error scales as RMSE $\propto 1/\sqrt{n}$ with context size n, where RMSE is Root Mean Squared Error. Doubling the context should reduce RMSE by approximately $\sqrt{2} \approx 1.41$.

Prediction P2 (Extreme Detection Saturation). For ICL-Hard event detection tasks, few-shot detection accuracy saturates when the number of simultaneous events J exceeds $J^* \approx 3$–5, regardless of context size: Accuracy($J > J^*$) ≈ constant for all n.

Figure 4 shows visualization of the contrasting scaling behavior of Type A and Type C tasks.

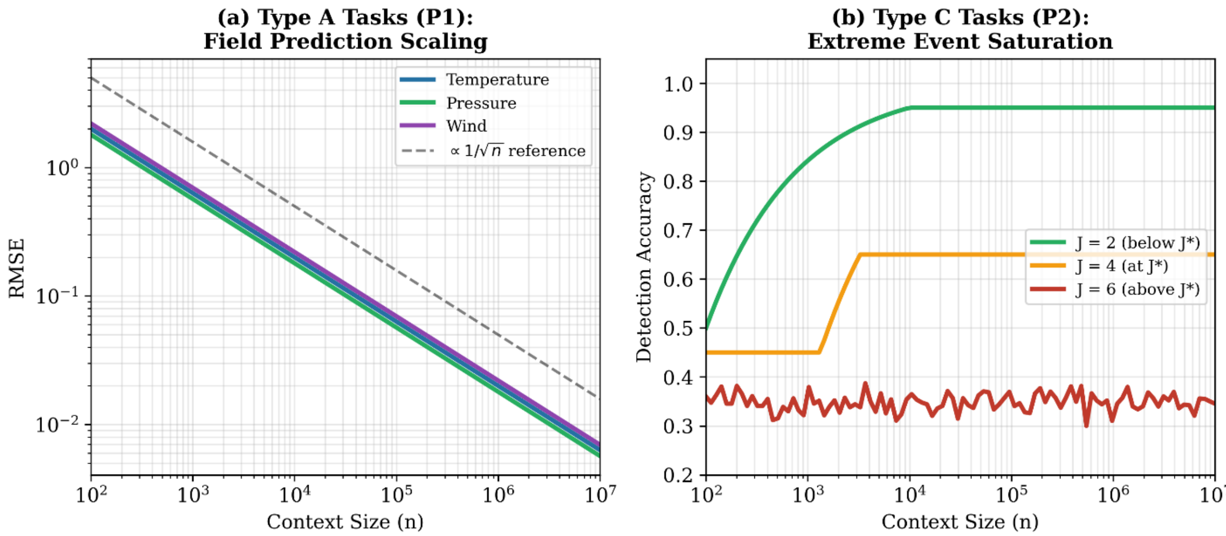

Figure 4. Contrasting scaling behavior of Type A and Type C tasks. (a) Field prediction RMSE scales as $1/\sqrt{n}$ with context size (Prediction P1), following optimal ERM rates. (b) Extreme event detection accuracy saturates for $J \geq J^*$: subcritical cases ($J = 2$) improve with data, while supercritical cases ($J = 6$) remain near random regardless of context size (Prediction P2).

Prediction P3 (Fine-Tuning Gap). The gap between fine-tuned and ICL performance differs dramatically by task type. Type A tasks: $\mathrm{Acc}_{\text{fine-tune}} - \mathrm{Acc}_{\mathrm{ICL}} = O(\varepsilon)$ (small gap). Type C tasks: $\mathrm{Acc}_{\text{fine-tune}} - \mathrm{Acc}_{\mathrm{ICL}} = \Omega(1)$ (large gap).

Prediction P4 (Depth-Width Asymmetry). For extreme event detection with $J > J^*$: increasing model *depth* (more layers) improves performance; increasing model *width* alone (larger hidden dimension) provides diminishing returns.

Prediction P5 (Ensemble Comparison). ICL-based field prediction should match or exceed traditional NWP ensemble forecasting: $\mathrm{Skill}_{\mathrm{ICL}} \geq \mathrm{Skill}_{\text{NWP-ensemble}}$ at comparable FLOP budgets for lead times $\tau < \tau_L$.

Prediction P6 (Sharp Threshold Behavior). Event detection accuracy exhibits a phase transition at J^*: For $J < J^*$, accuracy improves with n ($\mathrm{Acc}(n) \to 1$); at $J = J^*$, $\mathrm{Acc} \approx 0.5$–0.7; for $J > J^*$, accuracy bounded ($\mathrm{Acc}(n) \leq 0.5 + n^{-\Omega(1)}$). Expect sharp drop between $J = 3$–4 and $J = 5$–6.

6.4 Explanation of Empirical Results

The Climate Prediction Dichotomy explains several observed patterns in climate FM performance.

6.4.1 Why Pangu-Weather Excels at Deterministic Forecasting

Pangu-Weather (Bi et al., 2023) achieves state-of-the-art deterministic weather forecasting, matching or exceeding ECMWF's HRES system. *Theoretical Explanation:* Deterministic forecasting of temperature, pressure, geopotential, and wind fields are Type A tasks. By Theorem 6.1, these admit additive sufficient statistics enabling optimal ICL. The sample complexity $n_{ICL} = \Theta(K\sigma^2/\varepsilon)$ is achievable with massive pretraining datasets (39 years of ERA5, $\sim 10^8$ samples).

6.4.2 Why Extreme Event Prediction Struggles

Despite strong field prediction, climate FMs struggle with extreme event detection. PANGAEA benchmark results show GeoFMs underperform on event-based metrics compared to pixel-wise metrics. *Theoretical Explanation:* Extreme event detection is a Type C task with combinatorial sufficient statistics. The $J^* \approx 3\text{–}5$ threshold means detecting more than a handful of simultaneous events exceeds ICL capacity. The observed "saturation" reflects the J^* barrier, not insufficient data or model capacity.

6.4.3 Why Climate Projections Have High Uncertainty

Climate projections exhibit large uncertainty ranges that do not narrow beyond a certain point. *Theoretical Explanation:* This reflects two phenomena: (1) *Predictability horizon:* For $\tau > \tau_L \approx$ 14 days, chaotic dynamics prevent deterministic trajectory prediction (Kantz & Schreiber, 2004); physical limit, not learning limitation. (2) *Statistics vs. trajectories:* Climate *statistics* are Type A and ICL-accessible; individual *trajectories* are fundamentally unpredictable.

6.4.4 Connection to Benchmark Performance Patterns

WeatherBench 2 (Rasp et al., 2024): Field metrics (RMSE, ACC): strong AI → Type A. Extreme indices: weaker AI → Type C boundary. **ClimateBench (Watson-Parris et al., 2022):** Global mean temperature: strong AI → Type A. Extreme event frequency: moderate AI → Type C. **PANGAEA (Marsocci et al., 2024):** Semantic segmentation: strong → Type A. Object detection: weak → Type C. The dichotomy provides a unified explanation.

6.5 Deployment Guidelines

Based on the theoretical analysis, we provide practical guidelines for deploying climate foundation models. Figure 5 shows a visualization of the deployment decision flowchart for climate foundation models.

Guideline G1 (Task Assessment Protocol). Before deploying a climate FM: (1) Identify output type—continuous field values → likely Type A; discrete event locations → likely Type C. (2) Count discrete elements J—if $J = 0$, Type A confirmed; if $J > 0$, assess against J^*. (3) Check threshold—$J \leq 3$: ICL may suffice; $J > 5$: fine-tuning required. (4) Consider predictability horizon—$\tau < 14$ days: standard analysis; $\tau > 14$ days: statistics only.

Guideline G2 (When ICL Suffices). Use ICL for: field forecasting (temperature, pressure, wind for 1–14 days)—excellent performance; climate statistics (monthly/seasonal means, trends)—excellent; spatial interpolation and downscaling—very good; probabilistic forecasting—good; single-event detection—good.

Guideline G3 (When Fine-Tuning Is Required). Use fine-tuning for: multi-event detection ($J > 3$ simultaneous events)—J^* threshold; instance tracking (hurricane trajectories)—instance-level ID; tipping point timing—temporal localization; compound events—multi-variable $J > J^*$; rare event attribution—causal inference.

Guideline G4 (Hybrid Architectures). For applications requiring both field prediction and event detection, use a climate FM backbone (ICL for field prediction) with separate heads: an ICL-based head for temperature/pressure/wind forecasting and a fine-tuned detection head for extreme events. This combines the strengths of both approaches.

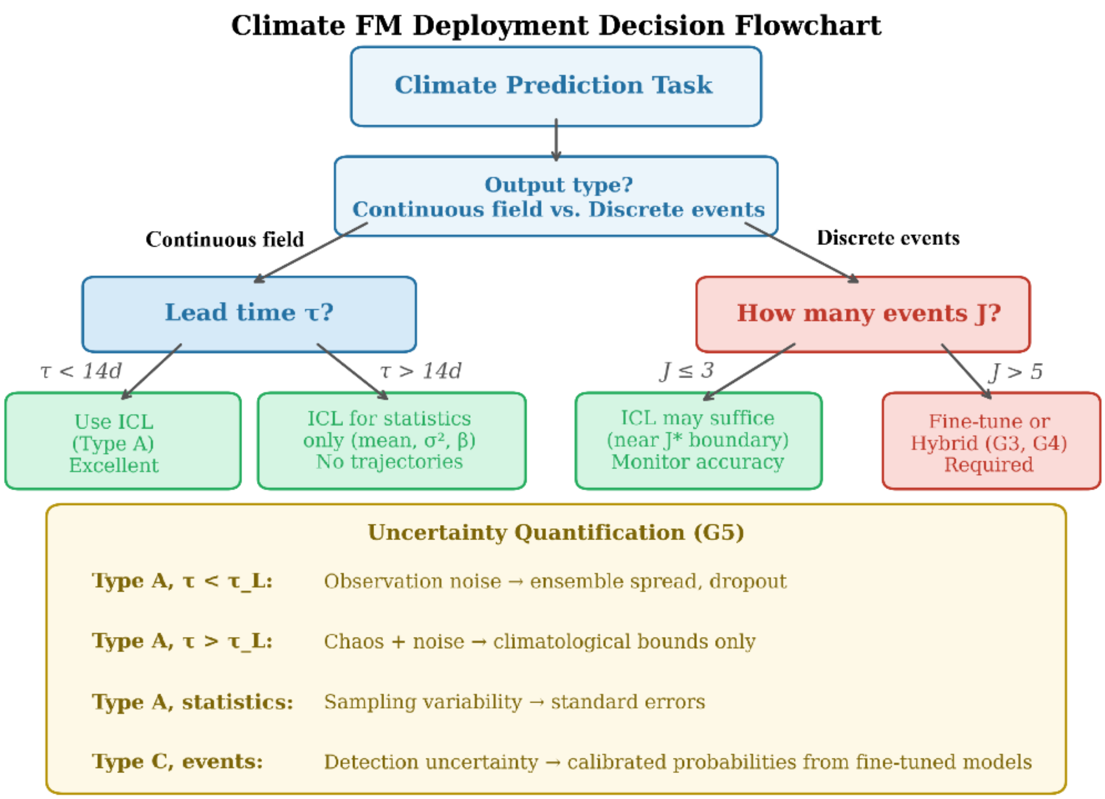

Figure 5. Deployment decision flowchart for climate foundation models. The protocol assesses output type (continuous vs. discrete), lead time relative to the predictability horizon τ_L, and number of simultaneous events J relative to J*, directing practitioners to appropriate methods and uncertainty quantification strategies (Guidelines G1–G5).

Guideline G5 (Uncertainty Quantification). Uncertainty varies by task type. Type A (field, $\tau < \tau_L$): observation noise—use ensemble spread, dropout. Type A (field, $\tau > \tau_L$): chaos + noise—use climatological bounds. Type A (statistics): sampling variability—use standard errors. Type C (events): detection uncertainty—use calibrated probabilities from fine-tuned models. Beyond τ_L, report climatological confidence intervals; narrower bounds are spurious.

6.6 Summary: The Climate Prediction Dichotomy

Main Result (Theorem 6.1): Every natural climate/weather task is either Type A (ICL-Easy)—field prediction with additive statistics, $n_{ICL} = \Theta(n_{ERM})$—or Type C (ICL-Hard)—event localization with combinatorial statistics, requires fine-tuning.

The Boundary: Predicting continuous *values* is Easy; identifying discrete *locations* is Hard.

Practical Threshold: $J^* \approx 3$–5 simultaneous events for typical climate FM context sizes.

Implications: (1) Climate FMs excel at what they're designed for (field prediction). (2) Event detection limitations are fundamental, not fixable by scaling. (3) Hybrid approaches combine ICL backbone with fine-tuned detection heads. (4) Predictability horizon $\tau_L \approx 14$ days is a physical limit, not a learning failure. The dichotomy provides principled guidance: deploy ICL confidently for Type A tasks, invest in fine-tuning for Type C tasks, and design hybrid systems for operational applications requiring both.

7. Discussion

The Climate Prediction Dichotomy provides a rigorous theoretical framework for understanding the capabilities and limitations of climate foundation models. In this section, we discuss the broader implications of our results, connections to related work, limitations of the analysis, and directions for future research.

7.1 Theoretical Implications

Unification across domains. The Climate Prediction Dichotomy (Theorem 6.1) joins the general ICL Dichotomy (Hawarey, 2026a) and the Remote Sensing Dichotomy (Hawarey, 2026c) as instantiations of a common structural principle: the additive versus combinatorial nature of sufficient statistics determines ICL complexity. This unification suggests that the dichotomy may be a universal feature of transformer-based learning across scientific domains. The climate domain contributes a unique physical constraint—the predictability horizon τ_L—that limits temporal scope but does not alter the fundamental Easy/Hard classification.

Fundamental versus practical limitations. Our analysis distinguishes two sources of limitation in climate FMs. First, *physical limits* arise from atmospheric chaos: the predictability horizon $\tau_L \approx 14$ days represents a fundamental constraint that no learning algorithm—ICL, fine-tuning, or otherwise—can overcome for trajectory prediction. Second, *computational limits* arise from the structure of sufficient statistics: the $J^* \approx 3$–5 threshold represents a barrier specific to constant-depth polynomial-size transformers that could, in principle, be overcome with deeper architectures or specialized algorithms. This distinction has important implications for resource allocation: investing in larger models will not improve performance beyond physical limits, while architectural innovations may help with computational limits.

The role of depth versus width. Proposition 5.3 established that increasing transformer depth can overcome the J^* barrier, while increasing width alone cannot. This has profound implications for climate FM architecture design. Current models like Pangu-Weather and GraphCast prioritize parameter efficiency through moderate depth (~12–24 layers). For applications requiring event detection beyond J*, our theory predicts that deeper architectures ($L = \Theta(J)$ layers) are necessary. This suggests a specialization principle: shallow-and-wide models for field prediction, deep models for event detection.

Circuit models versus practical transformers. Our hardness results model transformers as TC^0 circuits, following Merrill and Sabharwal (2023). Practical transformer architectures include features not captured by this abstraction, notably layer normalization, residual connections, and learned (rather than fixed) representations. We note that these features do not affect the fundamental hardness results for the following reasons. Layer normalization is a pointwise operation that does not increase the circuit's computational class; it can be absorbed into the threshold gates of TC^0 without increasing depth or size. Residual connections add skip pathways that effectively increase the circuit's connectivity but not its depth class; a depth-L transformer with residual connections still computes a TC^0 function of depth O(L). Learned representations affect the constants in complexity bounds (e.g., the effective dimension K may differ from the theoretical basis dimension) but do not change the asymptotic scaling of $J^* = O(\log \log n)$. Thus, while practical transformers may achieve somewhat better constants than our worst-case analysis predicts, the qualitative dichotomy and the existence of a hardness threshold remain valid. Empirical validation of the precise J* values for specific architectures remains an important direction (see Section 7.4).

7.2 Connections to Related Work

Climate foundation models. Our theoretical framework explains empirical observations across multiple climate FM papers. Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FourCastNet (Pathak et al., 2022), and ClimaX (Nguyen et al., 2023) all report strong performance on field prediction metrics (RMSE, ACC for temperature, pressure, wind) and weaker performance on extreme event metrics; ACC: Anomaly Correlation Coefficient. Our Type A/C classification provides a principled explanation: these models are optimally suited for additive-statistic tasks and fundamentally limited for combinatorial-statistic tasks. The theory predicts that performance gaps will persist regardless of scaling, which can guide future benchmark design and model evaluation.

Numerical weather prediction. The predictability horizon constraint (Theorem 4.3) connects to a rich literature on atmospheric predictability originating with Lorenz (1963). Our contribution is to integrate this physical constraint into the ICL complexity framework, showing that $\tau_L \approx 14$ days applies equally to NWP and ML-based forecasting. This confirms that AI weather models are not "cheating" by somehow circumventing chaos—they are simply more computationally efficient at exploiting available predictability. Beyond τ_L, both NWP ensembles and climate FMs converge to climatological uncertainty, as our theory predicts.

Transformer expressivity and circuit complexity. Our hardness results leverage the connection between transformers and circuit complexity classes established by Merrill and Sabharwal (2023). The key insight is that constant-depth polynomial-size transformers compute TC^0 functions, and Håstad's (1987) parity lower bounds extend to this class. This places our work within the broader program of understanding transformer limitations through computational complexity. The $J^* = O(\log \log n)$ threshold is not arbitrary—it emerges directly from the exponential-in-depth size requirements for parity computation.

In-context learning theory. This paper extends the ICL theory of Hawarey (2026a) to the climate domain, demonstrating that the SSC/AC-SSC framework generalizes beyond the linear and kernel regression settings originally analyzed (Garg et al., 2022). The spatiotemporal extension to $K = (N+1)^2 \times M$ dimensions shows that the framework scales to high-dimensional scientific problems (Wainwright, 2019). The connection between additive statistics and attention-computability appears to be a robust principle across domains.

7.3 Limitations

Idealized assumptions. Our theoretical analysis relies on several simplifying assumptions. We assume Gaussian observation noise, which may not hold for all climate variables (e.g., precipitation is highly non-Gaussian). We assume the true climate field lies exactly in the span of the spatiotemporal basis, ignoring model misspecification. We treat the basis functions as known, whereas practical implementations must learn representations. These idealizations enable rigorous analysis but may not capture all aspects of practical climate FM deployment.

Asymptotic versus finite-sample behavior. The $J^* = O(\log \log n)$ threshold is an asymptotic result. For practical context sizes $n \approx 10^5$–10^7, we estimate $J^* \approx 3$–5, but the exact threshold depends on constants hidden in asymptotic notation. Empirical validation is needed to determine precise threshold values for specific climate FM architectures and tasks. The sharp phase transition predicted by Corollary 5.2 may be smoothed in practice by finite-precision effects and approximate computations.

Structured data and worst-case assumptions. The sparse parity reduction underlying our ICL-Hard results (Theorem 5.1) assumes worst-case input structure, where the event locations are adversarially chosen. Real climate data exhibits strong spatial and temporal correlations: extreme events cluster geographically (e.g., heatwaves span contiguous regions) and temporally (e.g., compound events co-occur within synoptic weather patterns). These correlations could effectively reduce the number of independent degrees of freedom from J to a smaller effective J_{eff}, potentially smoothing the phase transition at J^* or shifting the threshold upward. In the most favorable case, if spatial correlations reduce a $J = 6$ event cluster to $J_{eff} \approx 3$ independent components, the task may remain within the tractable regime. However, the correlations do not eliminate the fundamental hardness: even with structured inputs, identifying J independent event locations from an exponential configuration space requires combinatorial search whenever J_{eff} exceeds J^*. The practical implication is that the $J^* \approx 3$–5 threshold derived from worst-case analysis may be conservative for spatially coherent events but remains applicable for events with weak spatial correlation (e.g., simultaneous but geographically distant extremes).

Scope of "natural" function classes. Theorem 6.1 claims the dichotomy holds for "natural" climate function classes, now formally defined in Definition 6.1. The key structural criterion is the output type: output-continuous classes ($\mathcal{Y} \subseteq \mathbb{R}^d$) are Type A, and output-discrete classes ($\mathcal{Y} \subseteq \mathcal{P}(\Omega)$) are Type C. This partition is exhaustive for tasks satisfying the input regularity condition (bandlimited fields with sub-Gaussian noise). The qualifier "natural" excludes pathological constructions—such as artificially hybridized tasks that combine continuous outputs with discrete selection in a single loss function—designed to achieve intermediate complexity. All 20 climate tasks identified in Table 5 satisfy Definition 6.1 and fall cleanly into one category. We cannot rule out the existence of intermediate cases for artificial tasks violating the input regularity condition, but such constructions would lack physical relevance to climate science.

Static analysis of dynamic models. Our analysis treats ICL as a single-pass inference problem. In practice, climate FMs may be applied autoregressively (predicting day 2 from predicted day 1), potentially accumulating errors or enabling iterative refinement. The interaction between autoregressive rollout and ICL complexity is not fully captured by our framework and merits further investigation.

7.4 Future Directions

Empirical validation. The six testable predictions (P1–P6) provide a concrete agenda for empirical research. Systematic experiments measuring scaling behavior (P1), saturation effects (P2), fine-tuning gaps (P3), depth-width asymmetry (P4), ensemble comparisons (P5), and threshold sharpness (P6) would validate or refine the theoretical predictions. Such experiments could be conducted using existing climate FMs on benchmarks like WeatherBench 2, ClimateBench, and PANGAEA.

Architectural innovations. Our analysis suggests that overcoming the J^* barrier requires either depth scaling ($L = \Theta(J)$) or fundamentally new architectures. Promising directions include sparse attention mechanisms that focus on event-relevant regions, hierarchical models that separate field prediction from event detection, and retrieval-augmented approaches that provide relevant historical events as context. The hybrid architecture recommended in Guideline G4 is a first step; more sophisticated integrations may be possible.

Extension to other Earth system components. This paper focused on atmospheric prediction. The ICL framework naturally extends to ocean modeling (sea surface temperature, currents, salinity), cryosphere dynamics (ice sheet evolution, sea ice extent), and coupled Earth system models. Each domain may introduce additional physical constraints analogous to the predictability horizon, creating a rich landscape for theoretical investigation.

Uncertainty quantification. Guideline G5 provided preliminary recommendations for uncertainty quantification, but a rigorous theory of ICL-based uncertainty for climate prediction remains to be developed. Key questions include: How should ensemble-like uncertainty be constructed from ICL models? When is ICL uncertainty calibrated? How do Type A and Type C tasks differ in their uncertainty characteristics? Addressing these questions is critical for operational climate FM deployment.

Causal and counterfactual reasoning. Climate attribution—determining whether specific extreme events were caused by anthropogenic climate change—is a high-priority application not fully addressed by our framework. Attribution involves counterfactual reasoning ("Would this event have occurred without climate change?") that goes beyond the predictive tasks analyzed here. Extending ICL theory to causal inference in climate science is an important open problem.

7.5 Broader Impact

The Climate Prediction Dichotomy has implications beyond technical performance metrics. Climate information supports critical decisions in disaster preparedness, agricultural planning, infrastructure design, and climate adaptation policy. Understanding the fundamental capabilities and limitations of climate FMs helps ensure that these tools are deployed appropriately—confidently for Type A tasks where they excel, cautiously for Type C tasks where specialized methods are required.

The dichotomy also has implications for research prioritization. Resources invested in scaling climate FMs will yield diminishing returns for Type C tasks due to fundamental computational barriers. Instead, the community may benefit from targeted investments in fine-tuned event detection models, hybrid architectures, and uncertainty-aware systems. Our theoretical framework provides principled guidance for allocating research and development resources in climate AI.

Finally, the predictability horizon constraint ($\tau_L \approx 14$ days) reminds us that even the most sophisticated AI cannot overcome fundamental physical limits. This underscores the importance of ensemble-based probabilistic forecasting, scenario planning for climate projections, and humility about the limits of prediction in a chaotic climate system.

8. Conclusions

This paper has established a rigorous theoretical foundation for understanding when climate foundation models can learn effectively from in-context examples. Our central contribution is the **Climate Prediction Dichotomy Theorem** (Theorem 6.1), which proves that every natural climate and weather prediction task falls into exactly one of two complexity categories: Type A (ICL-Easy) tasks involving smooth field prediction with additive sufficient statistics, and Type C (ICL-Hard) tasks involving discrete event localization with combinatorial sufficient statistics. There are no intermediate cases.

The theoretical framework developed here makes several key contributions. First, we formalized climate prediction tasks as mathematical function classes (Section 3), extending the geopotential representation of Hawarey (2026b) to the spatiotemporal domain with dimension $K = (N+1)^2 \times M$. Second, we proved that field prediction, weather forecasting (within the predictability horizon $\tau_L \approx 14$ days), and climate statistics estimation are ICL-Easy, achieving sample complexity $n_{ICL} = \Theta(n_{ERM})$ that matches empirical risk minimization (Section 4). Third, we proved that extreme event detection, tipping point localization, and compound event identification are ICL-Hard for $J > J^* \approx 3$–5 simultaneous events, requiring super-polynomial transformer size or super-constant depth (Section 5). Fourth, we derived six empirically testable predictions and five practical deployment guidelines that translate theory into actionable recommendations (Section 6).

The dichotomy provides a unified explanation for observed climate FM performance patterns. Models like Pangu-Weather, GraphCast, and FourCastNet excel at field prediction because these are Type A tasks optimally suited to attention-based aggregation. The same models struggle with event detection because these are Type C tasks that exceed ICL capacity. This is not a limitation to be engineered away through scaling—it is a fundamental computational barrier rooted in circuit complexity lower bounds. The practical threshold $J^* \approx 3–5$ means that climate FMs can reliably detect isolated extreme events but cannot handle multiple simultaneous events without fine-tuning or specialized architectures.

The predictability horizon constraint (Theorem 4.3) connects ICL theory to the physics of atmospheric chaos. Weather forecasting is ICL-Easy only for lead times $\tau < \tau_L \approx 14$ days; beyond this horizon, individual trajectories are fundamentally unpredictable by any method. Climate statistics (means, variances, trends) remain ICL-Easy for arbitrarily long timescales, but specific future climate states cannot be predicted. This distinction—statistics are predictable, trajectories are not—has important implications for climate projection and adaptation planning.

For practitioners deploying climate FMs, our analysis yields clear guidance. Use ICL confidently for temperature, pressure, wind, and geopotential forecasting; for seasonal means and climate trends; and for spatial interpolation and downscaling. Use fine-tuned models or hybrid architectures for multi-event detection ($J > 3$), hurricane tracking, tipping point timing, and compound event identification. Report climatological uncertainty bounds for predictions beyond the 14-day horizon; narrower bounds are spurious regardless of model sophistication.

Looking forward, the Climate Prediction Dichotomy opens several research directions. Empirical validation of predictions P1–P6 would test the theory against real climate FM behavior. Architectural innovations—deeper networks, sparse attention, retrieval augmentation—may partially overcome the J^* barrier for event detection. Extension to ocean, cryosphere, and coupled Earth system models would broaden the framework's applicability. Development of rigorous ICL-based uncertainty quantification methods would enhance operational utility.

In summary, this paper demonstrates that the remarkable success of climate foundation models on field prediction tasks and their limitations on event detection tasks are two sides of the same theoretical coin. The structure of sufficient statistics—additive for fields, combinatorial for events—determines what transformers can and cannot learn in context. This understanding enables principled deployment of climate AI: leveraging ICL where it excels, investing in fine-tuning where it is required, and respecting the physical limits that no algorithm can transcend.

9. Acknowledgement

This is an Artificial Intelligence-Assisted Research (AIAR). AI tools were employed as efficiency-enhancing assistants. The author declares no conflict of interest. This study received no external funding.

10. References

1. Allender, E., & Gore, V. (1993). A uniform circuit lower bound for the permanent. SIAM Journal on Computing, 23(5), 1026–1049.
2. Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619(7970), 533–538.
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901.
4. Dalcher, A., & Kalnay, E. (1987). Error growth and predictability in operational ECMWF forecasts. Tellus A, 39(5), 474–491.
5. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). A Survey on In-context Learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
6. Garg, S., Tsipras, D., Liang, P., & Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. Advances in Neural Information Processing Systems, 35, 30583–30598.
7. Håstad, J. (1987). Computational limitations of small-depth circuits. MIT Press.
8. Hawarey, M. (2026a). Characterizing In-Context Learning: When Can Transformers Match Standard Learning Algorithms? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026277, DOI: 10.65737/AIRMCS2026277.
9. Hawarey, M. (2026b). Quantum-Enhanced In-Context Learning for Geopotential Field Estimation: A Theoretical Framework. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026322, DOI: 10.65737/AIRMCS2026322.
10. Hawarey, M. (2026c). ICL Characterization of Geo-Foundation Models. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026360, DOI: 10.65737/AIRMCS2026360.
11. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P. d., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J. N. (2020) The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, Meteorol Soc. 2020;146:1999–2049. https://doi.org/10.1002/qj.3803
12. Hsu, D., Kakade, S., & Zhang, T. (2012). Random design analysis of ridge regression. Foundations of Computational Mathematics, 14(3), 569–600.

13. Kantz, H., & Schreiber, T. (2004). Nonlinear time series analysis (2nd ed.). Cambridge University Press.
14. Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. Science 382,1416-1421(2023). DOI:10.1126/science.adi2336
15. Lehmann, E. L., & Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation. Sankhyā: The Indian Journal of Statistics, 10(4), 305–340.
16. Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., & Schellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. Proceedings of the National Academy of Sciences, 105(6), 1786–1793.
17. Lorenz, E. N. (1963). Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20(2), 130–141.
18. Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. Tellus, 21(3), 289–307.
19. Marsocci, V., Nascetti, A., Burghardt, D., & Cong, Y. (2024). PANGAEA: A global benchmark for geospatial foundation models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
20. Merrill, W., & Sabharwal, A. (2023). The parallelism tradeoff: Limitations of log-precision transformers. Transactions of the Association for Computational Linguistics, 11, 531–545.
21. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. Proceedings of the 40th International Conference on Machine Learning, 25904–25938.
22. Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., & Anandkumar, A. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv preprint arXiv:2202.11214.
23. Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Bouallegue, Z. B., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., & Sha. F. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. Journal of Advances in Modeling Earth Systems, 16(6), e2023MS004019. https://doi.org/10.48550/arXiv.2308.15560
24. Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. Nature, 461(7260), 53–59. https://doi.org/10.1038/nature08227
25. Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics, 12(4), 389–434. https://doi.org/10.1007/s10208-011-9099-z
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762
27. Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint. Cambridge University Press.

28. Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., & Roesch , C. (2022). ClimateBench v1.0: A benchmark for data-driven climate projections. Journal of Advances in Modeling Earth Systems, 14(10), e2021MS002954. https://doi.org/10.1029/2021MS002954
29. Xie, S., Raghunathan, A., Liang, P., & Ma, T. (2022). An explanation of in-context learning as implicit Bayesian inference. International Conference on Learning Representations.
30. Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. Neural Networks, 94, 103–114.