CONFORMAL WEATHER FORECASTS: FIELD-COVERAGE AND LABEL-TIMING AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal calibration of a weather field must specify what is covered: an individual location, an average fraction of grid cells, or a field-level event. We study the event that at least 90% of unweighted grid cells lie inside their forecast intervals. A spatial order-statistic score gives the standard finite-sample split-conformal guarantee for this event under exchangeable forecast–verification pairs. A chronological Stormer audit instead tests where this assumption can fail: using 240 calibration cases from 2018–2019 and 336 test cases from 2020–2022, global calibration attains a 71.4% field-event rate. Six latitude bands attain 97.0% with mean width 4.27 K, but their worst-band score-event rate is only 48.2%; aggregate field coverage does not imply simultaneous regional validity. We also audit label timing using the same saved forecasts. For a fixed three-band rolling window, increasing an assumed label delay from zero to 90 days reduces field-event coverage from 98.2% to 89.6%, while width changes little. These are retrospective sensitivities using final archived targets, not recovered ERA5T vintages. The contribution is an explicit coverage-target and label-availability audit; operational validity under temporal shift remains outside the exchangeable guarantee.

1 INTRODUCTION

Weather forecasts drive evacuation decisions, grid dispatch, and crop protection, so getting the uncertainty right matters as much as getting the mean right. Over the past five years, machine learning models—GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), Stormer (Nguyen et al., 2024), and FuXi (Chen et al., 2023)—have surpassed operational numerical weather prediction (NWP) on standard bulk metrics such as RMSE of 500 hPa geopotential. The deterministic forecast configurations studied here emit a single forecast field rather than a calibrated prediction set. Their mean-squared-error (MSE) training objective creates a regression-to-the-mean bias that grows with lead time and with distance into the tails of the meteorological distribution. Strong point-forecast skill therefore leaves a separate question: which coverage event can a calibration procedure support, and under which assumptions?

The uncertainty-quantification gap. Operational ensembles and probabilistic neural systems such as ECMWF-ENS, GenCast (Price et al., 2025), and FuXi-ENS (Zhong et al., 2025) provide flow-dependent uncertainty and spatially coherent forecast alternatives. Conformal calibration addresses a different question: whether a chosen prediction set attains a specified frequentist coverage target under explicit assumptions. Recent work applies conformal prediction both to spatiotemporal surrogate models (Gopakumar et al., 2026) and to probabilistic AI weather forecasts (Asch et al., 2026). Our focus is the post-hoc calibration of deterministic neural weather backbones; we do not claim that this setting replaces or outperforms operational ensembles.

Our approach: post-hoc conformal prediction. We apply split conformal prediction (Vovk et al., 2005) to neural weather fields as a backbone-agnostic post-processing step. Empirically, a $\pm 1.645\sigma$ Gaussian residual interval reaches 88.5% coverage in the submitted 2020 experiment, while the corresponding conformal construction reaches 91.7% with 14% wider bands. The exact split-conformal statement requires exchangeability of calibration and test forecast–verification pairs. Here, "distribution free" means that no parametric residual distribution is assumed; it does not mean robustness to arbitrary temporal or climate shift. We therefore separate the exact theorem

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

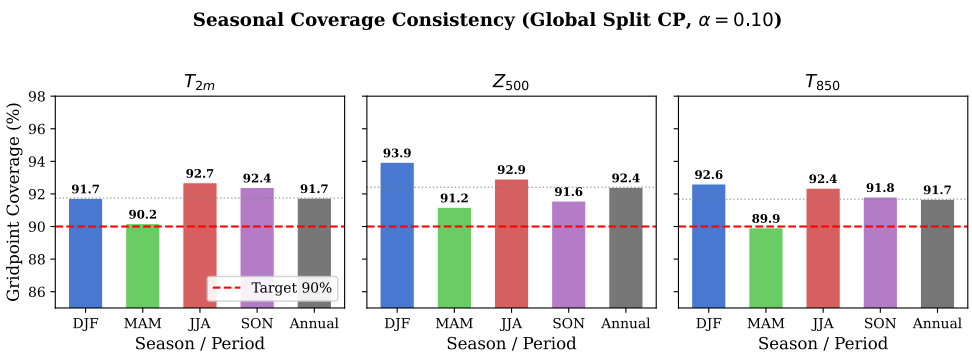

Figure 1: Seasonal mean grid-cell coverage of Global Split CP (90% reference, dashed) on three variables. This descriptive grid average is distinct from the field-event probability. ($n_{\text{cal}} = 240$, $n_{\text{test}} = 183$, full-year 2020.)

from chronological stress tests under temporal dependence. First, we introduce a *spatially-aware nonconformity score*—the q-th quantile (default $q = 0.90$) of the absolute error field across all grid points—which aggregates the spatial distribution of errors into a single scalar. This score aligns the conformal event with a field-level spatial-risk target. A spatial mean is also a valid conformal score for its own estimand, but it does not target the same event and yields approximately 50% empirical grid coverage in our submitted experiment. Second, we implement *Mondrian conformal prediction* stratified by latitude band and lead time. In the chronological experiments we treat this construction as empirical regional adaptation, rather than as a simultaneous or location-conditional guarantee.

Contributions. We make the field-event target explicit and derive its standard split-conformal construction; compare matched post-hoc calibrators using field, grid and regional endpoints; and measure how calibration-label delays change a rolling audit on the same forecasts. The theoretical statement and chronological observations answer separate questions: the former assumes exchangeability, while the latter measures its practical limitations in the retained years.

2 Related Work

2.1 Data-Driven Weather Forecasting

Transformer and graph neural network architectures have transformed medium-range weather prediction. Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), Stormer (Nguyen et al., 2024), and FuXi (Chen et al., 2023) all match or exceed ECMWF-IFS on WeatherBench 2 (Rasp et al., 2024) using ERA5 (Hersbach et al., 2020) as training data. NeuralGCM (Kochkov et al., 2024) hybridises a differentiable dynamical core with neural parameterisations for long-range stability. ClimaX (Nguyen et al., 2023) pioneered foundation-model pretraining on heterogeneous CMIP6 data. The deterministic configurations considered here produce point forecasts. Their forecast skill does not determine the coverage of a subsequently constructed prediction interval; calibration therefore requires a separate evaluation.

Extreme-weather and extreme-value modeling are discussed in the supplementary background. They motivate additional diagnostics, but are not assumptions of the spatial-quantile coverage guarantee.

2.2 Uncertainty Quantification for Weather

Ensemble NWP systems such as ECMWF-ENS have long been the gold standard for probabilistic weather forecasting. Among neural approaches, GenCast (Price et al., 2025) and FuXi-ENS (Zhong et al., 2025) generate ensembles via diffusion and variational inference, respectively, and provide flow-dependent uncertainty that is absent from deterministic backbones. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) provides an alternative: given a calibration set, it constructs prediction sets with *guaranteed* finite-sample coverage under exchangeability. Zaffran et al. (2022) extended conformal prediction to non-stationary time series via adaptive quantile estimation. Romano

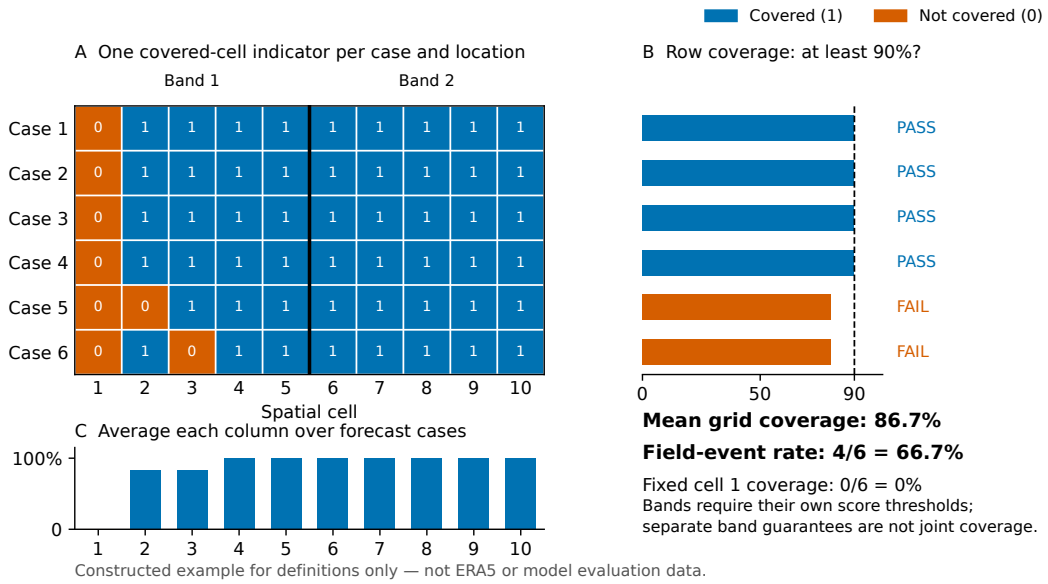

Figure 2: A constructed six-case, ten-cell example distinguishes coverage targets. (A) Entries indicate whether each interval covers its target. (B) Four of six rows cover at least 90% of cells, despite mean grid coverage of 86.7%. (C) The first cell is never covered. Grouping columns by latitude does not turn separately calibrated band guarantees into simultaneous validity. These illustrative values are not weather experiment results.

et al. (2019) introduced conformalized quantile regression (CQR), combining quantile regressors with a conformal correction. Gopakumar et al. (2026) preserve the tensor structure of spatiotemporal surrogate outputs and construct cell-wise conformal error bars, including applications to neural weather models. Their cell-wise construction targets marginal coverage for fixed output coordinates; our spatial-quantile construction instead targets a field-level event. Neither estimand dominates the other, and we compare them on identical Stormer forecasts in Section 6. More recently, Asch et al. (2026) apply online conformal prediction to probabilistic forecasts from GenCast, NeuralGCM, and AIFS-ENS. That setting retains ensemble-derived, flow-dependent uncertainty, whereas our experiments focus on post-hoc field-risk calibration for deterministic backbones.

Our method builds on Mondrian conformal prediction (Vovk et al., 2003) and adapts it for spatiotemporal forecast grids by stratifying calibration by latitude band and lead time, which allows interval widths to adapt to broad regional and lead-time differences. In our chronological experiments, we report this as empirical regional adaptation. Pooled performance does not imply simultaneous validity across bands or conditional coverage at a user-selected location.

3 Preliminaries

3.1 Problem Setting

Let $\mathcal{G} = \mathcal{S} \times \mathcal{T}$ denote the spatiotemporal forecast domain, where $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ is a set of $|\mathcal{S}|$ global grid points and $\mathcal{T} = \{1, \ldots, T\}$ is a set of T discrete lead times. A weather state at time t is a tensor $\mathbf{x}_t \in \mathbb{R}^{|\mathcal{S}| \times V}$ where V is the number of atmospheric variables (e.g., u_{10}, v_{10}, t_{2m}, z_{500}). A neural weather model $f_\theta : \mathbb{R}^{|\mathcal{S}| \times V} \to \mathbb{R}^{|\mathcal{S}| \times V \times T}$ maps an initial condition to a multi-step forecast $\hat{\mathbf{y}} = f_\theta(\mathbf{x}_t)$.

We are interested in the *marginal* distribution of the forecast residual $r_{s,\tau} = y_{s,\tau} - \hat{y}_{s,\tau}$ at grid point s and lead time τ. Standard MSE training minimizes $\mathbb{E}[r_{s,\tau}^2]$, which is dominated by the bulk of the distribution and systematically neglects the tail $\{r_{s,\tau} : |r_{s,\tau}| > u_s\}$ for a high threshold u_s.

Extreme-value modeling is background to a separate fine-tuning diagnostic, not an assumption of the conformal guarantee. Its peaks-over-threshold formulation is deferred to Appendix N.

3.2 Conformal Prediction

Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) constructs prediction sets $\mathcal{C}(\mathbf{x})$ with guaranteed coverage.

Definition 1 (Marginal Coverage). *$\mathcal{C}$ has $(1-\alpha)$-marginal coverage if $\mathbb{P}(\mathbf{y} \in \mathcal{C}(\mathbf{x})) \geq 1-\alpha$.*

Split conformal prediction. Given pre-trained f_θ, calibration set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_{\text{cal}}}$, and score $s(\mathbf{x}, \mathbf{y})$:

$$\hat{q} = \text{Quantile}\Big(\tfrac{\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil}{n_{\text{cal}}};\ \{s_i\}_{i=1}^{n_{\text{cal}}}\Big), \quad \mathcal{C}(\mathbf{x}) = \{\mathbf{y} : s(\mathbf{x}, \mathbf{y}) \leq \hat{q}\}. \tag{1}$$

This achieves $\mathbb{P}(\mathbf{y}_{\text{test}} \in \mathcal{C}(\mathbf{x}_{\text{test}})) \geq 1-\alpha$ under exchangeability (Vovk et al., 2005). Zaffran et al. (2022) extend this to non-stationary time series; Tibshirani et al. (2019) handle covariate shift.

Mondrian conformal prediction. Stratifying by category $\kappa(\mathbf{x})$ yields per-stratum coverage guarantees (Vovk et al., 2003) when exchangeability holds within each fixed stratum. Separate stratum-wise statements do not imply simultaneous validity over all strata; see Section 4.2.

4 Conformalized Weather Forecasting

Let f_θ be a pretrained deterministic neural weather model. Given an initial atmospheric state $\mathbf{x}_t \in \mathbb{R}^{|\mathcal{S}| \times V}$, it produces a forecast $\hat{\mathbf{y}}_t = f_\theta(\mathbf{x}_t) \in \mathbb{R}^{|\mathcal{S}| \times V}$ for a target lead time τ. Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_{\text{cal}}}$ be a held-out calibration set of n_{cal} non-overlapping forecast–analysis pairs. We impose *no distributional assumption* on the backbone errors and require exchangeability of the calibration and test pairs for the exact split-conformal guarantee. Temporal dependence and distribution shift can violate this condition; Section 6 therefore reports a chronological stress test separately from the theorem.

4.1 Spatially-Aware Nonconformity Score

The mean score targets a different event. A natural nonconformity score for a spatial forecast field is the spatially-averaged absolute error:

$$s_t^{\text{mean}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} |\hat{y}_{t,s} - y_{t,s}|. \tag{2}$$

The conformal quantile $\hat{q}$ is then the $(1-\alpha)$-th quantile of $\{s_i^{\text{mean}}\}_{i=1}^{n_{\text{cal}}}$, and the prediction set at test time is the set of fields $\mathbf{y}$ such that $s^{\text{mean}}(\hat{\mathbf{y}}, \mathbf{y}) \leq \hat{q}$. This is a legitimate conformal score, but its prediction set is defined by a bound on mean field error rather than by the fraction of covered grid cells. No spatial independence or law of large numbers is needed for this distinction. For example, ten errors all equal to one and nine errors equal to zero with one equal to ten both have mean one, yet a threshold of one covers 100% and 90% of their cells, respectively. A bound on the mean therefore does not specify the covered fraction. Concretely, when the resulting scalar threshold is used as a uniform grid-cell interval at target $1-\alpha = 0.90$, empirical grid coverage in our submitted experiment is approximately 50%. This does not invalidate mean-score conformal prediction for its own mean-error estimand; it shows that the estimand is mismatched to the field-level event studied here.

The spatially-aware score. We instead define the nonconformity score as the q-th spatial quantile of the per-gridpoint absolute error field:

$$s_t = \text{Quantile}_q\Big(\big\{|\hat{y}_{t,s} - y_{t,s}|\big\}_{s \in \mathcal{S}}\Big), \tag{3}$$

where we set $q = 0.90$ by default and use the order statistic of rank $\lceil q|\mathcal{S}| \rceil$ (not an interpolated sample quantile). This score is a single scalar per forecast time t that summarises the 90th percentile of the spatial error distribution. This uses the entire spatial field. Grid cells are correlated and must not be counted as independent calibration samples; the calibration sample unit is one forecast case. The score is the smallest threshold covering at least a fraction q of the grid cells.

Proposition 1 (Field-level coverage with a spatial quantile score)**.** *Let* $\{s_i\}_{i=1}^{n_{\text{cal}}}$ *be conformal scores computed via* (3) *on the calibration set, and let*

$$\hat{q} = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n_{\text{cal}}+1)\rceil}{n_{\text{cal}}};\ \{s_i\}_{i=1}^{n_{\text{cal}}}\right). \tag{4}$$

Here the conformal quantile means the order statistic of rank $k = \lceil (n_{\text{cal}}+1)(1-\alpha)\rceil$*; if* $k > n_{\text{cal}}$*, set* $\hat{q} = +\infty$ *rather than clipping the rank to the largest observed score. Define the prediction set at test time as*

$$\mathcal{C}_{1-\alpha}(\mathbf{x}) = \{\mathbf{y} : s(\hat{\mathbf{y}}, \mathbf{y}) \le \hat{q}\}. \tag{5}$$

Then under exchangeability of $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_{n_{\text{cal}}}, \mathbf{y}_{n_{\text{cal}}}), (\mathbf{x}, \mathbf{y})$*,*

$$\mathbb{P}\big(\mathbf{y} \in \mathcal{C}_{1-\alpha}(\mathbf{x})\big) \ge 1 - \alpha. \tag{6}$$

Proof. The prediction set (5) is defined by the event $\{s_{n_{\text{cal}}+1} \le \hat{q}\}$ where $s_{n_{\text{cal}}+1} = s(\hat{\mathbf{y}}_{\text{test}}, \mathbf{y}_{\text{test}})$. Under exchangeability, a randomly tie-broken rank of $s_{n_{\text{cal}}+1}$ among $s_1, \ldots, s_{n_{\text{cal}}+1}$ is uniform on $\{1, \ldots, n_{\text{cal}}+1\}$; including ties with $\le$ makes the deterministic coverage conservative. By the standard split-CP order-statistics argument (Vovk et al., 2005), coverage (6) follows directly. Appendix C states explicitly why the same argument does not automatically extend to the chronological stress test. □

Remark 1 (Interpretation of the prediction set)**.** *Under exchangeability, the prediction set* (5) *guarantees that the 90th-percentile spatial error does not exceed* $\hat{q}$ *with probability at least* $1-\alpha$*. Equivalently, the probability that at least 90% of grid cells have absolute error no greater than* $\hat{q}$ *is at least* $1-\alpha$*. This is a field-level spatial-risk statement. It is not location-conditional coverage, a guarantee for every grid point simultaneously, or a decomposition of epistemic model uncertainty.*

Calibration in practice. We use ERA5 reanalysis (Hersbach et al., 2020) 5-day forecasts with $n_{\text{cal}} = 240$ samples as the calibration set, generating scores $\{s_i\}$ via (3) and computing $\hat{q}$ at target $1-\alpha = 0.90$ via (4). The frozen Stormer backbone achieves a bulk RMSE of 3.10 K for $T_{2\text{m}}$ on the 183-sample test split used here (the EVT fine-tuning experiments in Appendix B use a slightly different evaluation split with baseline RMSE 3.39 K).

4.2 Mondrian Conformal Prediction by Latitude Band

Motivation. The marginal coverage guarantee in Proposition 1 is global: it holds on average over all test forecast dates but does not guarantee coverage for any specific geographic region. In practice, forecast error structure varies strongly with latitude—polar regions exhibit larger absolute errors than tropics, mid-latitude storm tracks show pronounced seasonality, and monsoon regions have distinct error modes. A single global quantile $\hat{q}$ will over-cover tropical grid points (prediction intervals wider than necessary) while potentially under-covering polar grid points (intervals too narrow for the actual error regime).

Latitude-band stratification. We partition the globe into six latitude bands:

$$\mathcal{B} = \big\{[-90^\circ, -60^\circ),\ [-60^\circ, -30^\circ),\ [-30^\circ, 0^\circ),\ [0^\circ, 30^\circ),\ [30^\circ, 60^\circ),\ [60^\circ, 90^\circ]\big\}. \tag{7}$$

For each band $b \in \mathcal{B}$ and lead time τ, we define a Mondrian category $\kappa = (b, \tau)$ and compute a band- and lead-time-specific nonconformity score:

$$s_t^{(b)} = \text{Quantile}_q\Big(\big\{|\hat{y}_{t,s} - y_{t,s}|\big\}_{s \in \mathcal{S}_b}\Big), \tag{8}$$

where $\mathcal{S}_b = \{s \in \mathcal{S} : \text{lat}(s) \in b\}$ is the set of grid points within band b. We then apply the Mondrian conformal quantile independently per (b, τ) stratum:

$$\hat{q}_{b,\tau} = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n_{b,\tau}+1)\rceil}{n_{b,\tau}};\ \big\{s_i^{(b,\tau)}\big\}_{i \in \mathcal{I}_\tau}\right), \tag{9}$$

Here $\mathcal{I}_\tau$ indexes calibration cases with the forecast and verification field available at lead τ, and $n_\tau = |\mathcal{I}_\tau|$. Each such case contributes to every band, so $n_{b,\tau} = n_\tau$; bands do not divide the number of calibration dates.

Algorithm 1 Post-Hoc Mondrian Conformal Calibration for Weather Fields

Require: Frozen backbone f_θ; calibration initial states $\mathbf{x}_i$ and lead-indexed verification fields $\mathbf{y}_{i,\tau}$ for $i \in \mathcal{I}_\tau$; spatial quantile level q; latitude bands $\mathcal{B}$; lead times $\mathcal{T}$; miscoverage level α; $|\mathcal{I}_\tau| \geq 1$ for every evaluated lead.

// Phase 1: Compute calibration nonconformity scores
for each lead $\tau \in \mathcal{T}$, case $i \in \mathcal{I}_\tau$ **do**
  $\hat{\mathbf{y}}_{i,\tau} \leftarrow f_\theta^{(\tau)}(\mathbf{x}_i)$
  for each band $b \in \mathcal{B}$ **do**
    $s_i^{(b,\tau)} \leftarrow \mathrm{Quantile}_q\big(\{|\hat{y}_{i,s,\tau} - y_{i,s,\tau}|\}_{s \in \mathcal{S}_b}\big)$ ▷ Eq. (8)
  end for
end for
// Phase 2: Mondrian quantile per stratum
for each stratum $(b, \tau) \in \mathcal{B} \times \mathcal{T}$ **do**
  $n_{b,\tau} \leftarrow |\mathcal{I}_\tau|$
  $\hat{q}_{b,\tau} \leftarrow \mathrm{Quantile}\Big(\lceil (1-\alpha)(n_{b,\tau}+1) \rceil / n_{b,\tau};\ \{s_i^{(b,\tau)}\}_{i \in \mathcal{I}_\tau}\Big)$ ▷ Eq. (9)
end for

Ensure: Calibrated quantile table $\{\hat{q}_{b,\tau}\}_{b \in \mathcal{B},\, \tau \in \mathcal{T}}$

// Inference (test time):
$\hat{\mathbf{y}}_\tau \leftarrow f_\theta^{(\tau)}(\mathbf{x}_{\mathrm{test}})$ for each τ
for each band b, lead time τ **do**
  $I_{s,\tau}(\mathbf{x}) \leftarrow \big[\hat{y}_{s,\tau} - \hat{q}_{b,\tau},\ \hat{y}_{s,\tau} + \hat{q}_{b,\tau}\big]$ for all $s \in \mathcal{S}_b$
end for
return the indexed interval array $\{I_{s,\tau}(\mathbf{x}) : s \in \mathcal{S}, \tau \in \mathcal{T}\}$

Interpretation under exchangeability. Define the displayed interval at grid cell $s \in \mathcal{S}_b$ and lead time τ as $I_{s,\tau}(\mathbf{x}) = [\hat{y}_{s,\tau} - \hat{q}_{b,\tau}, \hat{y}_{s,\tau} + \hat{q}_{b,\tau}]$. If calibration and test forecast–verification pairs are exchangeable, the fixed-band order statistic gives the event guarantee

$$\mathbb{P}\left\{ \frac{1}{|\mathcal{S}_b|} \sum_{s \in \mathcal{S}_b} \mathbf{1}\{y_{s,\tau} \in I_{s,\tau}(\mathbf{x})\} \geq q \right\} \geq 1 - \alpha, \tag{10}$$

This collection of separate statements is not a simultaneous guarantee across all bands, nor does it provide conditional coverage at each location. In the chronologically separated evaluation, exchangeability is not assumed and we describe the observed results only as empirical regional adaptation.

Remark 2 (Sample efficiency)**.** *Each calibration forecast contributes a score to every fixed latitude band at each evaluated lead time. The relevant sample size for a band–lead score sequence is therefore the number of calibration forecasts, not the number of grid cells in that band. Finer partitions can nevertheless make regional estimates noisier and increase multiplicity concerns, so we report sensitivity to 3, 6, and 9 bands rather than claiming that six bands are universally optimal.*

Lead-time stratification. Lead time is part of the main banded construction. We stratify by lead time $\tau \in \{1\text{-day}, 3\text{-day}, 5\text{-day}\}$, recognizing that forecast skill and error variance scale strongly with τ. The full Mondrian grid is $|\mathcal{B}| \times |\mathcal{T}| = 6 \times 3 = 18$ strata. Each stratum receives its own calibration quantile, so that the prediction interval width adapts jointly to geographic region and lead time.

Exploratory EVT fine-tuning. In a separate experiment, fine-tuning the last 25% of Stormer's layers with a composite MSE and Generalized Pareto Distribution (GPD) loss increased 5-day $T_{2\mathrm{m}}$ RMSE from 3.39 K to 3.56 K. This negative result is not part of the conformal method; we move its setup, ablations, and possible explanations to Appendix B.

5 Experiments

Data. We use ERA5 reanalysis (Hersbach et al., 2020) at $1.40625°$ resolution for three atmospheric variables: 2-metre temperature ($T_{2\mathrm{m}}$), 500 hPa geopotential (Z_{500}), and 850 hPa temperature (T_{850}).

Calibration set: $n_{\text{cal}} = 240$ daily forecast–analysis pairs. Test set: $n_{\text{test}} = 183$ samples spanning all seasons of 2020 (held out). The calibration set covers a preceding period; the test set covers the full calendar year 2020. This chronological separation does not establish exchangeability, so its coverage values are empirical rather than a verification of the finite-sample theorem. 2020 was a strong La Niña year with anomalously low Arctic sea ice, so single-year coverage statistics should not be over-generalised (see Limitations, Section 7).

Multi-year chronological stress test. Using the same official Stormer checkpoint and WeatherBench2 ERA5 source, we sample 240 daily 00 UTC calibration initializations from 2018–2019 and 112 held-out initializations from each of 2020, 2021, and 2022 (336 test cases; fixed seed). Each forecast uses five successive 24-hour steps. All methods use the same day-5 $T_{2\text{m}}$ forecasts and cases. Interval thresholds, climatology, and extreme-event thresholds use calibration data only. Because calibration and testing are chronologically separated, we report this as a stress test for which exchangeability is not assumed, rather than as a verification of Proposition 1.

Backbone. We use the publicly released Stormer (Nguyen et al., 2024) pretrained checkpoint at $1.40625°$ resolution. The backbone is *entirely frozen*; no fine-tuning is performed for the conformal prediction experiments. The EVT fine-tuning results reported in Appendix B use a separate experimental branch and are not used in the coverage evaluation.

Methods compared.

- **Stormer backbone**: deterministic forecast; no uncertainty quantification. Bulk 5-day $T_{2\text{m}}$ RMSE $= 3.10\,\text{K}$ on the full-year 2020 test set. It is a forecast source, not a UQ baseline.
- **Pointwise Gaussian residual interval**: a matched post-hoc UQ baseline using calibration residual scale and a $\pm 1.645\sigma$ interval.
- **Cell-wise residual CP**: a matched location-wise conformal baseline that retains a separate absolute-error calibration distribution for each grid cell (Gopakumar et al., 2026).
- **Global CP (mean score)**: split CP with the spatially-averaged absolute error as the nonconformity score (2). It is valid for its mean-error estimand under exchangeability, but is empirically mismatched to the grid-coverage metric used here.
- **Global CP (spatial quantile score, ours)**: split CP with the 90th-percentile spatial score (3). Under exchangeability, it targets the event that at least 90% of grid cells are covered with probability at least $1 - \alpha$.
- **Mondrian CP (ours)**: latitude-band-stratified conformal prediction per Algorithm 1. Evaluated as empirical regional adaptation in the chronological stress test.

Quantile regression is not included as a matched post-hoc baseline because it requires fitting or modifying a predictive model. Conformalized quantile regression (Romano et al., 2019) is compatible with our framework when conditional quantile estimates are available. We also do not evaluate probabilistic ensembles on matched initializations. Applying an absolute-error score only to an ensemble mean would discard useful spread; more suitable extensions include conformalized ensemble quantiles or a spread-normalized residual.

Evaluation. We distinguish three quantities. *Grid coverage* is the cosine-latitude-weighted fraction of grid cells inside their intervals, averaged over test dates. The *field-event rate* is the fraction of test dates on which at least $q = 0.90$ of grid cells are covered, counted *without* latitude weights; for Global CP this is equivalent to $s_{\text{test}} \leq \hat{q}$. Area-weighted coverage and the unweighted field event are different estimands. We also report mean interval width in Kelvin and regional diagnostics. These empirical summaries should not be conflated with the finite-sample probability statement, which additionally requires exchangeability.

Online updating. The calibration order statistic can be recomputed as new forecast–verification pairs arrive, including with a rolling window. Under drift, an ordinary rolling window does not retain the exchangeable finite-sample guarantee. Adaptive online conformal methods instead target long-run error control under their own assumptions (Zaffran et al., 2022; Asch et al., 2026).

Table 1: Chronological day-5 $T_{2\mathrm{m}}$ stress test. Calibration uses 240 cases from 2018–2019; evaluation uses 336 cases from 2020–2022. These are empirical results under temporal shift, not guaranteed coverage.

Method	Grid coverage	Field event	Width (K)
Pointwise Gaussian	85.3%	0.3%	2.96
Cell-wise residual CP	87.0%	3.0%	3.05
Global spatial-risk CP	95.2%	71.4%	5.68
6-band spatial-risk CP	92.2%	97.0%	4.27

Table 2: Sensitivity to the latitude partition in the chronological stress test. Six bands give the smallest width among the evaluated partitions; this does not establish universal optimality.

Partition	Grid cov.	Field event	Width (K)	Worst-band grid cov.
Global	95.2%	71.4%	5.68	–
3 bands	93.5%	94.9%	5.29	90.6%
6 bands	92.2%	97.0%	4.27	89.8%
9 bands	92.2%	97.6%	4.33	90.2%

6 RESULTS

Chronological stress-test results. Table 1 compares matched post-hoc methods on the 336 test forecasts. Grid coverage and width are cosine-latitude weighted. The field event is the fraction of dates on which at least 90% of all grid cells are covered. The Global spatial-risk method achieves only a 71.4% field-event rate, showing that its nominal exchangeable guarantee does not carry over to this temporal split. The six-band construction has a 97.0% aggregate field-event rate and a smaller mean width than Global CP, but this pooled number is not simultaneous validity across bands: its worst-band score-event rate is 48.2%.

Cell-wise CP targets fixed-location marginal coverage, whereas the spatial-risk methods target a field-level event; the table therefore exposes an estimand and width tradeoff rather than establishing that one method dominates. Across 2020, 2021, and 2022, six-band grid coverage is 92.3%, 92.1%, and 92.2%, and its aggregate field-event rate is 99.1%, 95.5%, and 96.4%, respectively. The fifth-percentile location coverage is 80.4% for cell-wise CP, 68.5% for Global spatial-risk CP, and 73.8% for six-band spatial-risk CP, further demonstrating that field-level and location-wise coverage are different targets.

Lead-time stratification is part of the main banded construction. In a separate matched audit of saved 2020 forecasts, six-band empirical grid coverage and width are 92.3%/1.94 K at day 1, 92.9%/3.03 K at day 3, and 93.4%/4.40 K at day 5.

Calibration-label availability. We reanalyze all 336 cases with the latest 60 labels eligible at forecast initialization. Eligibility requires the forecast's five-day verification time plus an assumed publication delay to precede initialization. The three-band partition, window and delays were fixed before this reanalysis. At zero delay, rolling calibration increases the field-event rate from 94.94% to 98.21% relative to fixed three-band calibration, a paired difference of 3.27 pp (descriptive 95% calendar-block interval [0.88,6.41]); width increases by 0.387 K [0.318,0.456]. This is a coverage–width tradeoff.

Figure 3 isolates the delay sensitivity within that rolling rule. A 90-day delay lowers field-event coverage to 89.58% even though mean width remains close to 5.68 K. The same interval width can therefore conceal a materially different field-event rate. All cases and five delays are retained. The targets are final-vintage ERA5 fields: delaying them changes eligibility, but does not reconstruct the values historically available in ERA5T. Appendix J gives all methods and block-length sensitivities.

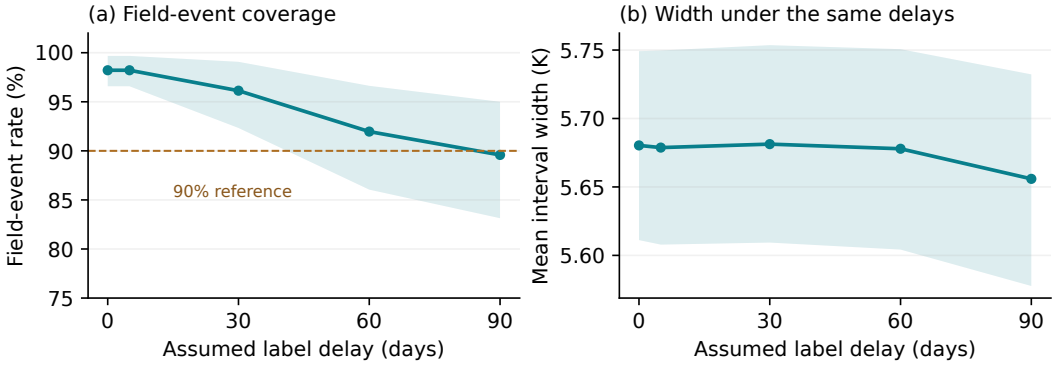

Figure 3: Existing-forecast label-delay sensitivity for the fixed three-band, 60-case rolling rule. Both panels retain the same 336 cases and all five assumed delays. Shading gives descriptive pointwise 95% intervals from the saved 30-calendar-day block bootstrap. The dashed line is a 90% reference, not a shift-valid guarantee. These are final-vintage targets, not recovered historical ERA5T releases.

6.1 Temporal uncertainty in the coverage–width contrast

Using the same saved forecasts and per-case metrics, we recompute 10,000 nonoverlapping calendar-block bootstrap replicates for 30, 60, and 90 days. The archive includes all five label delays (0, 5, 30, 60, and 90 days), three individual years and pooled evaluation, six methods, and paired rolling-minus-fixed contrasts. The following zero-delay pooled band3 contrast illustrates sensitivity without changing point estimates: area-weighted grid-cell coverage improves by 0.00399 and width increases by 0.3873 K.

Block days	Coverage difference 95% CI	Width difference 95% CI (K)
30	[0.00222, 0.00564]	[0.3181, 0.4561]
60	[0.00180, 0.00607]	[0.3065, 0.4770]
90	[0.00103, 0.00652]	[0.2858, 0.4881]

The coverage–width tradeoff remains in this contrast. Longer blocks widen some intervals and do not remove the need to justify temporal dependence assumptions. This is a retrospective fixed-final-vintage study with simulated label availability; it does not recover historical weather analysis vintages or prove coverage under arbitrary dependence.

The bootstrap resamples dates in blocks rather than treating individual grid cells as independent observations. A large spatial grid therefore does not supply an equally large number of independent calibration or evaluation cases.

7 Conclusion

We formulate deterministic neural-weather calibration as control of a field-level spatial-risk event. Under exchangeability, split conformal prediction gives finite-sample coverage for this explicit target; it is neither location conditional nor robust to arbitrary temporal shift. The chronological stress test shows that regional adaptation can improve empirical coverage, while the detailed results and assumption boundaries are reported in the appendix. A label-matured rolling-window reanalysis improves empirical coverage at the cost of wider intervals (Appendix J); it does not establish shift-robust coverage.

Limitations. Field coverage averages one event per forecast; it cannot certify a user-selected location or simultaneous coverage of all latitude bands. The chronological comparison also leaves calibration-defined extremes under-covered, as detailed in Appendix L. Rolling calibration changes the set of eligible residuals; it neither recovers historical analysis vintages nor establishes exchangeability. Operational use would require evaluating the actual labels available at each initialization, together with the forecast system and target region used in deployment.

Reproducibility Statement

The manuscript and appendix specify the coverage targets, calibration and test splits, score construction, and evaluation units. Retained evaluation code and saved-prediction analyses distinguish mean grid coverage from the field-level event. Historical backbone and quantile plots are regenerated from archived summary records and remain descriptive. The coverage-target matrix is a constructed definition example, not an ERA5 measurement. Available reanalyses do not establish validity under arbitrary temporal shifts or independently reproduce all backbone training.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anna Asch, Raphael Rossellini, Pedram Hassanzadeh, and Rebecca Willett. Rigorous uncertainty quantification of probabilistic AI weather forecasts with conformal prediction. *arXiv preprint arXiv:2606.19642*, 2026.

August A Balkema and Laurens de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974.

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

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

Vignesh Gopakumar, Ander Gray, Joel Oskarsson, Lorenzo Zanisi, Daniel Giles, Matt J. Kusner, Stanislas Pamela, and Marc Peter Deisenroth. Uncertainty quantification of surrogate models using conformal prediction. *Machine Learning: Science and Technology*, 7:015025, 2026. doi: 10.1088/2632-2153/ae2e7b.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Claudio Peubey, Raluca Radu, Dirk Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

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

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

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, pp. 25904–25938. PMLR, 2023.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Olivier C Pasche and Sebastian Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics*, 18(4):2818–2839, 2024.

James Pickands, III. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1): 119–131, 1975.

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

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Bauer, Iulian Sandu, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sager, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. Technical Report 4, Royal Holloway, University of London, 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

Shan Zhao, Zhitong Xiong, Jie Zhao, and Xiao Xiang Zhu. Exebench: Benchmarking foundation models on extreme earth events. *arXiv preprint arXiv:2505.08529*, 2025.

Xiaohui Zhong, Lei Chen, Hao Li, Roberto Buizza, Jun Liu, Jie Feng, Zijian Zhu, Xu Fan, Kan Dai, Jing-Jia Luo, Jie Wu, and Bo Lu. FuXi-ENS: A machine learning model for efficient and accurate ensemble weather prediction. *Science Advances*, 11(44):eadu2854, 2025. doi: 10.1126/sciadv.adu2854. URL `https://doi.org/10.1126/sciadv.adu2854`.

A Extreme Value Theory Background

Theorem 1 (Pickands-Balkema-de Haan (Pickands, 1975; Balkema & de Haan, 1974)). *Let X be a random variable with distribution F. For a wide class of distributions F (all max-domain-of-attraction distributions), there exist functions $\sigma(u)$ such that*

$$\lim_{u \to x^*} \sup_{z \geq 0} \left| \mathbb{P}(X - u \leq z \mid X > u) - \mathrm{GPD}_{\xi, \sigma(u)}(z) \right| = 0, \tag{11}$$

where $x^* = \sup\{x : F(x) < 1\}$.

Theorem 1 justifies modelling the conditional distribution of exceedances $Z = X - u \mid X > u$ as a GPD for sufficiently large u. In our setting, $X = r_{s,\tau}$ (or its absolute value, for two-sided extremes) and the threshold u_s is set to the empirical 90th percentile of training residuals per grid cell.

B EVT Fine-Tuning: Full Experimental Details

Before presenting the conformal prediction results, we document a negative finding that shaped the paper's direction and carries independent methodological value for the community.

Experimental setup. We fine-tuned the last 25% of Stormer's transformer layers—the final three of twelve attention blocks—using a composite loss:

$$\mathcal{L}_{\mathrm{tail}}(\theta;\, \xi, \sigma) = \mathcal{L}_{\mathrm{MSE}} + \lambda\, \mathcal{L}_{\mathrm{POT}}(\theta;\, \xi, \sigma), \tag{12}$$

where $\mathcal{L}_{\mathrm{MSE}}$ is the standard mean-squared error over all grid points and variables, and $\mathcal{L}_{\mathrm{POT}}$ is the GPD negative log-likelihood applied to residuals exceeding the 90th-percentile spatial threshold $u_{s,v}$ (see Section N for the GPD definition and Theorem 1). We swept $\lambda \in \{0.01, 0.05, 0.1, 0.5\}$, trained with AdamW at learning rate 3×10^{-5} with 500-step linear warmup, and evaluated on ERA5 5-day $T_{2\mathrm{m}}$ forecasts over full-year 2020 (183 test samples). The partial fine-tuning strategy (only the top layers) was chosen deliberately to reduce the risk of catastrophic forgetting.

Results. Table 3 summarises the outcome. Fine-tuning with the GPD tail loss *increased* 5-day RMSE from 3.39 K (baseline Stormer) to 3.56 K—a statistically consistent $+5\%$ degradation across all λ values tested. No configuration of λ recovered baseline performance, and larger λ values produced monotonically worse RMSE.

Table 3: 5-day $T_{2\mathrm{m}}$ RMSE on ERA5 full-year 2020 (183 samples). All EVT fine-tuned variants degrade skill relative to the frozen baseline.

Model	**RMSE (K)**	**Δ vs. Baseline**
Stormer (baseline, frozen)	3.39	—
Stormer + EVT fine-tune ($\lambda = 0.01$)	3.47	$+2.4\%$
Stormer + EVT fine-tune ($\lambda = 0.05$)	3.51	$+3.5\%$
Stormer + EVT fine-tune ($\lambda = 0.10$)	3.56	$+5.0\%$
Stormer + EVT fine-tune ($\lambda = 0.50$)	3.71	$+9.4\%$

Table 4: Detailed forecast skill comparison for 5-day $T_{2\mathrm{m}}$ on ERA5 full-year 2020 (183 test samples). Tail-RMSE$_{95}$ and Tail-RMSE$_{99}$ are computed over grid points whose absolute error exceeds the 95th and 99th percentile of the error distribution, respectively. Lower is better (↓).

Method	RMSE (K) ↓	Tail-RMSE$_{95}$ (K) ↓	Tail-RMSE$_{99}$ (K) ↓
Stormer (baseline)	3.39 ± 0.36	4.40	4.39
GPD-FT Stormer ($\lambda = 0.10$)	3.56 ± 0.38	4.66	4.74

Analysis: three failure mechanisms. We identify three structural reasons for the degradation, each of which generalizes beyond this specific experiment.

(1) Lead-time mismatch. The fine-tuning data consisted of ERA5 5-day forecasts, but Stormer was originally pretrained on a curriculum spanning lead times from 6 hours to 10 days, with a carefully tuned loss weighting that balances short- and long-range skill. The GPD tail loss was applied only to the 5-day residuals, creating a gradient signal that implicitly prioritized the tail behavior at day 5 at the expense of shorter lead-time accuracy. Because Stormer's autoregressive rollout accumulates errors from earlier steps, degraded day-1 and day-2 skill propagated forward to amplify the day-5 error. A lead-time-aware composite loss that re-weights the GPD penalty at each lead time proportionally could partially address this, but would require substantially more calibration data.

(2) Catastrophic forgetting of pretrained dynamics. Even restricting fine-tuning to the top 25% of layers, the GPD gradient signal conflicted with the pretrained weight structure. Stormer's top layers encode multi-scale synoptic dynamics built up over hundreds of millions of training steps on 40 years of ERA5. The tail loss, which activates only for residuals exceeding the 90th percentile ($\sim$10% of all spatiotemporal points per batch), provides a sparse and noisy gradient that disrupts these learned representations without providing a clear replacement signal. The layer-wise learning rate decay (0.85^k for layer k from the top) we employed was insufficient to prevent overwriting.

(3) Multi-variable interference. The composite loss sums $\mathcal{L}_{\mathrm{POT}}$ contributions across all variables ($T_{2\mathrm{m}}$, T_{850}, Z_{500}, u_{10}, v_{10}, q_{700}, precipitation). Each variable has different error variance and tail shape ξ, so a single loss coefficient λ applied uniformly creates conflicting gradient directions. Variables with larger-magnitude residuals (e.g., Z_{500}) dominate the GPD gradient, causing the backbone to adjust its representations in ways that reduce geopotential tail errors while increasing $T_{2\mathrm{m}}$ errors—the metric we care about most here. Per-variable λ tuning would require a 7-dimensional sweep beyond our computational budget.

Implication for design. These observations are specific to the evaluated Stormer fine-tuning setup, although they suggest broader questions about adapting pretrained models to sparse tail objectives. The GPD likelihood is a high-variance, sparse signal that is fundamentally at odds with the dense, low-variance MSE signal that drove pretraining. For the configuration tested here, post-hoc conformal calibration preserves the frozen backbone skill whereas the tail-loss variants do not. Conformal prediction requires no gradient through the backbone and has the order-statistics guarantee in Proposition 1 when exchangeability holds. This experiment does not establish that tail fine-tuning must fail for every architecture, loss, or dataset.

C Proof of Proposition 1 and Mondrian Coverage

Setup. Let $\mathcal{D}_{\mathrm{cal}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_{\mathrm{cal}}}$ be the calibration set. Denote by $s_i = s(f_\theta(\mathbf{x}_i), \mathbf{y}_i)$ the spatially-aware nonconformity score (3) for calibration point i, and let $s_{n_{\mathrm{cal}}+1} = s(f_\theta(\mathbf{x}), \mathbf{y})$ be the test score.

Step 1: Coverage under exchangeability. If $((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_{n_{\mathrm{cal}}+1}, \mathbf{y}_{n_{\mathrm{cal}}+1}))$ are exchangeable, the rank of $s_{n_{\mathrm{cal}}+1}$ among $\{s_1, \ldots, s_{n_{\mathrm{cal}}+1}\}$ is uniformly distributed on $\{1, \ldots, n_{\mathrm{cal}} + 1\}$. By the standard order-statistics argument (Vovk et al., 2005),

$$\mathbb{P}\big(s_{n_{\mathrm{cal}}+1} \le \hat{q}\big) \ge \frac{\lceil (1-\alpha)(n_{\mathrm{cal}}+1) \rceil}{n_{\mathrm{cal}}+1} \ge 1 - \alpha, \tag{13}$$

which is precisely the coverage guarantee (6).

Step 2: Scope under temporal dependence. The rank argument in Step 1 does not by itself apply when calibration and test forecast–verification pairs are temporally dependent or drawn from different distributions. Autocorrelation estimates and effective sample sizes are diagnostics, not replacement theorems. We therefore make no numerical β-mixing correction and do not claim exact validity for the chronologically separated 2018–2019 to 2020–2022 evaluation. Its observed coverage is reported directly in Table 1.

Step 3: Fixed-band statement under exchangeability (10)**.** For each stratum (b, τ), the calibration scores $\{s_i^{(b,\tau)}\}$ form an exchangeable sub-sequence because a fixed band/lead score is a deterministic function of each exchangeable forecast–verification pair. Independence between bands is not required. Applying Step 1 within each stratum gives

$$\mathbb{P}\big(s_{n_{\mathrm{cal}}+1}^{(b,\tau)} \le \hat{q}_{b,\tau}\big) \ge 1 - \alpha, \tag{14}$$

establishing (10) for that fixed band-level event. This does not establish joint validity over every band, conditional coverage at individual locations, or validity under the chronological split. □

D Additional Experimental Details

Hardware and compute. The submitted forecast-generation experiments used a single NVIDIA A100-SXM4-80GB GPU; EVT fine-tuning (Section B) used 4 GPUs for 10 epochs (≈2 h wall-clock). We separately benchmarked postprocessing overhead on the same Stormer workload. A batch-one, five-step Stormer model forward pass took a median of 1.928 s on one H100 over ten timed runs after two warm-up runs. Using one CPU thread, computing the six-band calibration thresholds from 240 saved fields took 61.0 ms once, and constructing intervals for one new saved forecast took 0.023 ms. We report these raw timings separately because model inference and conformal postprocessing were measured on different hardware. This is a within-backbone latency measurement, not a matched compute comparison with probabilistic ensemble systems.

Data. ERA5 at $1.5°$ resolution streamed from the WeatherBench2 Zarr store on Google Cloud Storage (`gs://weatherbench2/datasets/era5/`), bilinearly interpolated to the Stormer grid 128×256 at $1.40625°$. Normalisation constants (per-variable mean and standard deviation) are taken from the public Stormer release (Nguyen et al., 2024).

Forecast generation. We use 24-hour autoregressive rollout with the 24-hour Stormer head, yielding 5 steps for a 120-hour (5-day) forecast. ClimaX uses its native 24-hour head with the same 5-step rollout.

Conformal calibration. The split conformal quantile is computed at level $\lceil (1-\alpha)(n_{\mathrm{cal}}+1) \rceil / n_{\mathrm{cal}}$ with $\alpha = 0.10$. Per-variable climatology μ_s^{clim} is estimated as the pointwise mean over calibration targets. The spatial score is the 0.90 quantile of $|r_{s,\tau}|$ over the 128×256 grid per forecast.

E Additional Results and Visualisations

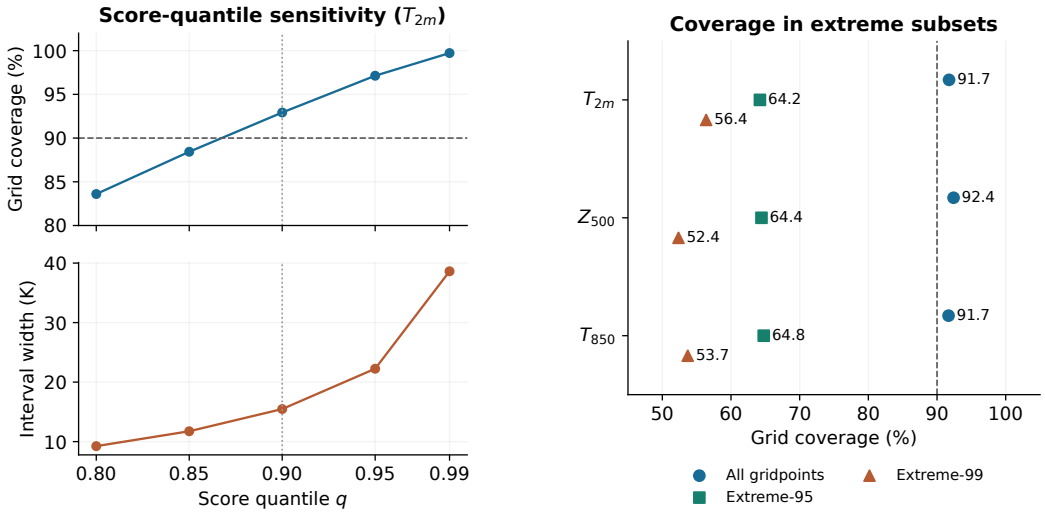

Figure 4: **Left:** Ablation of score quantile q for T_{2m} (Stormer). $q = 0.90$ is the tested quantile associated with $\geq$90% coverage. **Right:** Coverage gap between all gridpoints and extreme-event gridpoints (obs. anomaly $>$ 95th/99th pct). Standard CP substantially under-covers extremes, motivating EVT-augmented conformal scores as future work.

Table 5: Backbone comparison: RMSE and CP coverage ($1-\alpha = 0.90$) for Stormer and ClimaX on 5-day ERA5 forecasts (183 test samples, 2020). Despite ClimaX having substantially higher RMSE, the CP framework achieves ∼90% coverage on both backbones by automatically widening the prediction intervals to match each backbone's error distribution.

	ClimaX (n=60)		**Stormer** (n=183)	
Variable	**RMSE**	**Coverage [95% CI]**	**RMSE**	**Coverage [95% CI]**
$T_{2\mathrm{m}}$	17.5 K	91.7% [82.0, 96.4]	3.1 K	91.7% [86.8, 94.9]
Z_{500}	3373 $\mathrm{m}^2/\mathrm{s}^2$	89.5% [79.2, 95.0]	1042 $\mathrm{m}^2/\mathrm{s}^2$	92.4% [87.6, 95.4]
T_{850}	20.4 K	92.3% [82.7, 96.8]	4.6 K	91.7% [86.8, 94.9]

Wilson score 95% confidence intervals. All ClimaX CIs include 90%.

Table 6: Ablation: effect of spatial quantile level q on $T_{2\mathrm{m}}$ interval width and empirical coverage ($n_{\mathrm{test}} = 183$, 5-day forecasts, target $1-\alpha = 0.90$). Bold row is the recommended default.

q	**Width (K)**	**Coverage**	**≥90%?**
0.80	9.24	83.6%	×
0.85	11.74	88.4%	×
0.90	**15.47**	**92.9%**	✓
0.95	22.25	97.1%	✓
0.99	38.62	99.7%	✓

For Z_{500}: multiply widths by ≈297 ($\mathrm{m}^2/\mathrm{s}^2$/K ratio).
For T_{850}: widths ≈ 1.25× larger than T_{2m}.

Table 7: Submitted-2020 diagnostic stratified by event intensity. Although aggregate grid coverage is above 90%, coverage conditioned on observed anomalies above the 95th/99th percentile is substantially lower. This table is descriptive and does not imply an extreme-conditional guarantee.

Variable	Global (all)	Extreme-95	Extreme-99
$T_{2\mathrm{m}}$	91.7%	64.2%	56.4%
Z_{500}	92.4%	64.4%	52.4%
T_{850}	91.7%	64.8%	53.7%

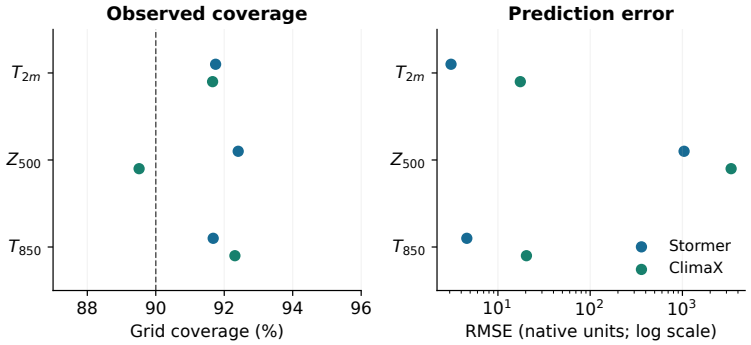

Figure 5: Backbone-agnostic demonstration. Despite a 5.6× RMSE gap, Stormer and ClimaX achieve near-identical gridpoint coverage (≈91%) because the conformal calibration automatically adjusts interval width to match each backbone's error distribution.

Table 8: Empirical grid coverage on calibration-defined extreme-anomaly subsets in the chronological stress test. No method has an extreme-conditional guarantee.

Method	95th percentile	99th percentile
Pointwise Gaussian	82.3%	79.9%
Cell-wise residual CP	82.8%	79.8%
Global spatial-risk CP	73.5%	66.7%
6-band spatial-risk CP	84.5%	80.0%

F Extreme-Event Analysis

Extreme event analysis. The method does not provide an extreme-conditional guarantee. In the multi-year stress test, using anomaly thresholds defined from calibration data only, empirical coverage at the 95th/99th anomaly percentiles is 82.3%/79.9% for Gaussian intervals, 82.8%/79.8% for cell-wise CP, 73.5%/66.7% for Global spatial-risk CP, and 84.5%/80.0% for six-band CP. All methods under-cover this conditioned subset. The result also shows that Global spatial-risk CP can perform worse than the matched baselines on extremes, so the field-level construction should not be presented as a remedy for rare-event calibration.

G Detailed Coverage Results

Coverage results. For completeness, Table 9 reports empirical coverage at target $1 - \alpha = 0.90$ on the 183-sample test set (full year 2020), calibrated on $n_{\text{cal}} = 240$ samples:

- **Mean-score CP is mismatched to grid coverage**: global split CP with the spatially averaged nonconformity score (2) achieves approximately 50% empirical grid coverage. The mean score remains conformally valid for its own mean-error estimand, but is empirically mismatched to the spatial coverage event evaluated here.
- **Spatial-quantile CP reaches the empirical grid-coverage target**: Global Split CP with the 90th-percentile spatial score (3) achieves **91.7%** annual gridpoint coverage for $T_{2\text{m}}$ at the 90% target, with an average interval width of $15.47\,\text{K}$. Corresponding results for Z_{500} and T_{850} are 92.4% and 91.7%, respectively (Table 9). Seasonal coverage is consistent: DJF, MAM, JJA, and SON all remain within 1–3 percentage points of the 90% nominal target.
- **Latitude-banded CP has higher empirical grid coverage**: after stratifying by latitude band and lead time (Algorithm 1), $T_{2\text{m}}$ gridpoint coverage rises to **94.2%**, with adaptive widths that are narrower in the tropics (lower error variance) and wider in the polar bands (higher error variance). Mondrian coverage for Z_{500} and T_{850} is 94.4% and 93.8%, respectively.
- **EVT fine-tuning widens intervals**: because the EVT fine-tuned backbone has higher RMSE ($3.56\,\text{K}$ vs. $3.39\,\text{K}$ on the EVT evaluation subset of 183 samples, $+5\%$; see Appendix B), the calibration nonconformity scores are larger and the conformal threshold $\hat{q}$ is wider—the opposite of the intended effect.

Naive Gaussian vs. conformal intervals. The naïve Gaussian interval with $\pm 1.645\sigma$ (theoretically 90% for Gaussian errors) achieves only **88.5%** empirical gridpoint coverage, falling short of the 90% target. Real atmospheric residuals have fatter tails than a Gaussian, so the $\pm 1.645\sigma$ quantile systematically underestimates the required interval width. Our Global Split CP achieves **91.7%** coverage for $T_{2\text{m}}$ with only 14% wider intervals ($15.47\,\text{K}$ vs. $13.52\,\text{K}$). The split-conformal field-event guarantee requires exchangeability but makes no parametric assumption on the error distribution. Latitude-banded CP has higher empirical $T_{2\text{m}}$ grid coverage in this dataset, at **94.2%**, with widths adapted by latitude. The corresponding Global/latitude-banded empirical coverages are 92.4%/94.4% for Z_{500} and 91.7%/93.8% for T_{850}. We report the 88.5% Gaussian comparison only for $T_{2\text{m}}$ and do not extrapolate that value to the other variables. All results use $n_{\text{cal}} = 240$ calibration samples and $n_{\text{test}} = 183$ test samples on 5-day forecasts.

Temporal-dependence caveat. Autocorrelation and effective sample size are useful diagnostics, but they do not replace exchangeability or establish an exact finite-sample guarantee. We therefore

Table 9: Conformal prediction coverage for 5-day forecasts at target $1-\alpha = 0.90$. Calibration: $n_{\text{cal}} = 240$ samples; test: $n_{\text{test}} = 183$ samples spanning all four seasons of 2020. Global CP is empirical annual gridpoint coverage with the spatial-quantile score (3). The Mondrian column reports empirical coverage after latitude-band stratification; it is not simultaneous validity across bands. DJF/MAM/JJA/SON are seasonal coverages under Global CP. RMSE is the frozen Stormer backbone error. † Naïve $\pm 1.645\sigma$ achieves only 88.5% empirically due to heavy-tailed error distributions (theoretical 90% holds only for Gaussian errors).

Variable	RMSE	Global CP	Mondrian	DJF	MAM	JJA	SON
$T_{2\text{m}}$ (K)	3.10 K	**91.7%**	**94.2%**	91.7%	90.2%	92.7%	92.4%
Z_{500} (m^2/s^2)	1042 m^2/s^2	**92.4%**	**94.4%**	93.9%	91.2%	92.9%	91.6%
T_{850} (K)	4.60 K	**91.7%**	**93.8%**	92.6%	89.9%	92.4%	91.8%
Naïve $\pm 1.645\sigma^{\dagger}$	—	88.5%	—	—	—	—	—

remove the submitted numerical β-mixing slack and interpret all chronologically separated results as empirical stress tests. Methods for covariate shift or adaptive online conformal control require their own assumptions (Tibshirani et al., 2019; Zaffran et al., 2022).

H Full Coverage Table

The backbone-agnostic comparison (Appendix E) confirms that ClimaX achieves 91.7% $T_{2\text{m}}$ coverage with $\hat{q} = 36.2\,\text{K}$, vs. Stormer's 91.7% at $\hat{q} = 7.7\,\text{K}$.

I Additional Robustness Checks

Backbone-agnosticity. To demonstrate backbone-agnosticity, we apply the identical CP framework to ClimaX (Nguyen et al., 2023), a ViT foundation model with substantially lower forecast accuracy than Stormer. Appendix E shows that despite ClimaX's T_{2m} RMSE being 5.6× larger (17.5 K vs. 3.1 K), our framework achieves 91.7% gridpoint coverage on both backbones. The CP calibration automatically widens the intervals to compensate for the weaker backbone ($\hat{q} = 36.2\,\text{K}$ for ClimaX vs. 7.7 K for Stormer), showing empirical backbone portability in this evaluation. Exact coverage for either backbone still requires exchangeability.

Extreme-event coverage (global vs. 95th/99th pct anomalies) is discussed in Appendix E.

The ablation over q, the extreme-event coverage gap, and the backbone-agnostic coverage comparison (Stormer vs. ClimaX) are reported in Appendix E.

Ablation: spatial quantile level q. Appendix E reports empirical coverage and interval width for $T_{2\text{m}}$ as q varies from 0.80 to 0.99. The relationship between q and coverage is strictly monotone: higher q produces higher nonconformity scores, a higher conformal threshold, and wider but more conservative intervals. The default $q = 0.90$ specifies the scientific estimand that 90% of the spatial domain should be covered; it is not selected to manufacture a particular test-set coverage value. The ablation illustrates how changing the spatial-risk tolerance changes empirical grid coverage and width.

The full ablation over score quantile $q \in \{0.80, 0.85, 0.90, 0.95, 0.99\}$ is reported in Appendix E. At the pre-specified $q = 0.90$ estimand, the saved-forecast audit gives 92.9% empirical grid coverage with interval width $15.47\,\text{K}$.

Extreme-event analysis. Stress tests on calibration-defined extremes are reported in Appendix F; they do not provide extreme-conditional guarantees.

J LABEL-MATURED ROLLING CALIBRATION AUDIT

This retrospective analysis uses the same 576 forecasts as the fixed-calibration comparison; it changes calibration eligibility rather than the forecast model. At each test initialization, fixed calibration includes only 2018–2019 cases with verification time no later than that initialization. Rolling calibration uses the latest 60 cases satisfying that same rule, including earlier test cases only after their five-day labels mature.

The first test initialization and latest fixed-calibration verification time are both 2020-01-05 00 UTC. Thus no fixed-calibration labels are excluded as immature in this particular sample. This assumes labels are available at their verification time; operational observation latency and reanalysis publication delays are not modeled in this zero-delay result. The sensitivity analysis below varies hypothetical delays; neither analysis is a complete operational availability audit.

Table 10: All 336 test cases. Grid coverage and interval width are area weighted; the field event requires coverage of at least 90% of unweighted grid cells.

Method	Grid coverage	Field-event rate	Width (K)
Fixed global	95.16%	71.43%	5.681
Rolling global	95.98%	87.80%	6.247
Fixed cell-wise	87.01%	2.98%	3.053
Rolling cell-wise	89.17%	28.27%	3.483
Fixed three-band	93.46%	94.94%	5.293
Rolling three-band	93.86%	98.21%	5.680

For three bands, the paired rolling-minus-fixed field-event difference is 3.27 percentage points (descriptive 95% block-bootstrap interval $[0.88, 6.41]$ points), while width increases by 0.387 K ($[0.318, 0.456]$ K). We resample nonoverlapping 30-calendar-day blocks with all their observed cases, using 10,000 replicates and seed 10669. Per-year results and paired case metrics are archived. The window of 60 and three-band comparison were fixed before this analysis, without tuning on outcomes. This is a coverage–width tradeoff, not an equal-width improvement or an arbitrary-dependence guarantee. Cell-wise calibration targets a different marginal event, so its low field-event rate does not by itself refute its intended target.

J.1 PRESPECIFIED LABEL-RELEASE DELAY SENSITIVITY

We repeat all six methods on the same 336 test cases under assumed release delays of 0, 5, 30, 60, and 90 days. A calibration label is eligible only when its verification time plus the assumed delay is no later than forecast initialization. The rolling window remains the latest 60 eligible cases, with the original 30-day block bootstrap. Delays and methods were fixed before this analysis; no cases or delay settings are selected using outcomes. The zero-delay case metrics reproduce the preceding analysis exactly, and all saved rolling-source timestamps pass the delayed-availability check.

Table 11: Rolling three-band calibration under hypothetical publication delays. Coverage columns are percentages; field events use unweighted grid cells.

Delay (days)	Area grid coverage	Field-event rate	Width (K)
0	93.86	98.21	5.680
5	93.84	98.21	5.679
30	93.75	96.13	5.681
60	93.66	91.96	5.678
90	93.58	89.58	5.656

These scenarios quantify sensitivity to stale calibration labels; they are not measurements of actual ERA5 release dates. ECMWF documents approximately five-day latency for ERA5T and a final release two to three months later[1]. The five-day scenario uses final archived labels, not recovered historical ERA5T vintages; preliminary and final values can differ. Consequently, none of these

[1] ECMWF ERA5 dataset documentation, accessed September 2026.

delay scenarios certifies an operationally available historical label stream. All six methods, paired comparisons, and block intervals are retained in the artifact. The results remain retrospective and do not establish a guarantee under arbitrary temporal dependence.

J.2 Time-block sensitivity

We additionally recompute descriptive intervals using nonoverlapping 60- and 90-calendar-day blocks, retaining the original 30-day calculation. All five assumed publication delays, six methods, rolling-minus-fixed pairs and individual years are reported. Each calculation retains all 336 pooled cases; changing blocks does not change point estimates. Few large blocks within a year make these intervals unstable, not guaranteed under unrestricted temporal dependence. This reanalysis does not create new forecasts or historical vintages.

K Expanded Conclusion

We formulate post-hoc calibration of deterministic neural weather forecasts as control of a field-level spatial-risk event. A spatial-quantile score makes the target explicit: under exchangeability, split conformal prediction guarantees that at least a chosen fraction of the domain is within the calibrated threshold with probability at least $1 - \alpha$, without a parametric residual model. The guarantee is neither location conditional nor robust to arbitrary temporal shift. On the submitted 2020 Stormer evaluation, Global Split CP obtains 91.7% empirical grid coverage. In the additional 2018–2019 to 2020–2022 chronological stress test, the six-band construction obtains 92.2% grid coverage, a 97.0% aggregate field-event rate, and a mean width of 4.27 K. Global spatial-risk CP reaches only a 71.4% field-event rate in the same stress test, demonstrating the practical importance of the assumption boundary. The latitude-banded method is best interpreted here as empirical regional adaptation. Cell-wise calibration remains more appropriate when the scientific question concerns a fixed location, while operational ensembles provide flow-dependent uncertainty that our deterministic-backbone study does not replace.

L Detailed Limitations

Limitations. **(i)** Exact finite-sample validity requires exchangeability; ordinary split CP need not retain nominal coverage under temporal or climate shift. **(ii)** The spatial-risk target is field level and does not give conditional coverage at a user-selected location or simultaneous validity for all bands. **(iii)** All evaluated methods under-cover calibration-defined extreme subsets; at the 99th anomaly percentile, the best of the matched stress-test methods reaches 80.0% coverage. **(iv)** We do not evaluate GenCast, FuXi-ENS, ECMWF-ENS, NeuralGCM, or AIFS-ENS on matched initializations, so no empirical efficiency or accuracy advantage over probabilistic systems is claimed. **(v)** The 2020–2022 evaluation is broader than the submitted single-year test but remains insufficient for operational deployment or long-term climate-shift conclusions.

M Additional evidence boundary

The retrospective delay experiments use final-vintage verification fields. A real-time revision study requires preserved initial bytes, verifiable first-availability times, and later final versions; file modification times or present-day downloads of final ERA5 do not recover that information. Calibration-label availability and forecast-initialization availability are separate requirements.

N Extreme Value Theory: Peaks Over Threshold

Definition 2 (Generalized Pareto Distribution)**.** *A random variable $Z \geq 0$ follows a Generalized Pareto Distribution $Z \sim \mathrm{GPD}(\xi, \sigma)$ if its survival function is*

$$\mathbb{P}(Z > z) = \begin{cases} \left(1 + \xi \frac{z}{\sigma}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ \exp\left(-\frac{z}{\sigma}\right) & \text{if } \xi = 0, \end{cases} \tag{15}$$

where $\sigma > 0$ is the scale parameter and $\xi \in \mathbb{R}$ is the shape parameter ($\xi > 0$: heavy tail; $\xi < 0$: bounded tail; $\xi = 0$: exponential tail).

The Pickands–Balkema–de Haan theorem (Pickands, 1975; Balkema & de Haan, 1974) justifies modelling excess $Z = X - u \mid X > u$ as a $\mathrm{GPD}(\xi, \sigma)$ for large u; see Appendix A for the formal statement. In our setting, $X = r_{s,\tau}$ (or its absolute value, for two-sided extremes) and the threshold u_s is set to the empirical 90th percentile of training residuals per grid cell.

O Extreme Weather Forecasting

Several works have sought to improve machine learning forecasts of extremes. Pasche & Engelke (2024) applied Extreme Value Theory to neural network quantile regression for flood risk, fitting GPDs to model residuals but without integrating EVT into training of a large pretrained backbone. Taillardat et al. (2016) calibrated ensemble forecasts using quantile regression forests with model output statistics, demonstrating the value of post-hoc tail calibration for operational NWP. ExEBench (Zhao et al., 2025) benchmarked foundation models on seven categories of extreme Earth events, documenting systematic underestimation of extreme intensities across all tested models. NeuralGCM (Kochkov et al., 2024) hybridised a differentiable dynamical core with learned neural parameterisations for stable long-range simulations. Separately from the conformal method, we report a controlled case study in which GPD-based fine-tuning of a pretrained Stormer backbone degrades forecast skill (Appendix B).

P Extreme Value Theory in Machine Learning

EVT has been applied to anomaly detection (Coles, 2001) and neural network generalisation, but its integration into the training objectives of large-scale weather models remains largely unexplored. The Pickands-Balkema-de Haan theorem (Pickands, 1975; Balkema & de Haan, 1974) motivates the use of GPD-based losses for tail calibration: exceedances above a sufficiently high threshold converge to a GPD in distribution. Appendix B reports a negative result for one such setup: the evaluated GPD loss degrades overall skill when applied to the pretrained Stormer backbone. We treat this as a configuration-specific case study rather than a general claim about EVT fine-tuning.