EXTREMEPHYSBENCH: PHYSICAL STRUCTURE EVALUATION OF DATA-DRIVEN WEATHER PREDICTION MODELS UNDER EXTREMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Extreme weather events involve organized thermodynamic and dynamical structures, including thermal transport, moisture convergence, and vortex organization. Although data-driven weather prediction (DWP) has become competitive with numerical weather prediction (NWP) in aggregate forecast accuracy, its ability to reproduce these physical structures during extremes remains insufficiently characterized. We introduce **ExtremePhysBench**, a benchmark for systematically evaluating physical structures in forecasts of extreme weather events. The benchmark provides **(i) an event catalog** of 820 heat, cold, precipitation, and landfalling tropical cyclone events during 2022–2025, selected independently of the evaluated forecasts and yielding 6,317 forecast tasks; **(ii) physical diagnostics** comprising 25 universal and event-specific definitions that compare balance, transport, stability, and vortex organization with ERA5; and **(iii) Physics Scores and physical families** that quantify agreement with ERA5 and organize related diagnostics by physical process, supporting both process-level assessment and aggregate comparisons under a common evaluation protocol. Evaluating six DWP systems and operational GFS as the NWP baseline, we find that all six DWP systems achieve higher Overall Physics Scores than GFS. Physical family scores reveal model strengths and weaknesses in physical structure agreement across extreme weather event categories and processes, providing information complementary to conventional accuracy metrics and aggregate scores. Code is available at: `https://anonymous.4open.science/r/ExtremePhysBench-REVIEW/`.

1 INTRODUCTION

Extreme weather events, such as heavy precipitation and tropical cyclones, pose substantial risks to societies and critical infrastructure, making their accurate and timely prediction a central goal of weather forecasting. Yet forecasting them involves more than reproducing the magnitude of individual atmospheric variables. The development and evolution of extremes depend on coherent interactions among dynamical and thermodynamic processes. For example, moisture transport and convergence in heavy precipitation, and organized circulation, outflow, and warm-core structure in tropical cyclones (IPCC, 2021; Holton & Hakim, 2013). Therefore, a reliable forecast should capture the multivariate physical organization of the atmosphere that sustains the event, making physical structure fidelity an essential dimension for evaluating forecasts under extreme conditions.

Recent work has made extreme weather evaluation more systematic. HR-Extreme measures forecast field errors within masks for 17 extreme weather types (Ran et al., 2025); ExtremeWeatherBench standardizes high-impact cases, observations, and hazard-relevant metrics (McGovern et al., 2026); and TCBench evaluates tropical cyclone track and intensity (Gomez et al., 2026). Dedicated evaluations of recent high-impact events further examine how forecast models perform during extremes (Pasche et al., 2025). These studies establish essential comparisons of field accuracy and hazard characteristics. Their reported evaluation targets, however, do not provide a shared comparison of the physical structures underlying different extreme weather event categories. An error in precipitation or tropical cyclone intensity and track alone does not locate departures in moisture supply, circulation, or core warmth. The remaining need is an evaluation that compares these physical

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.

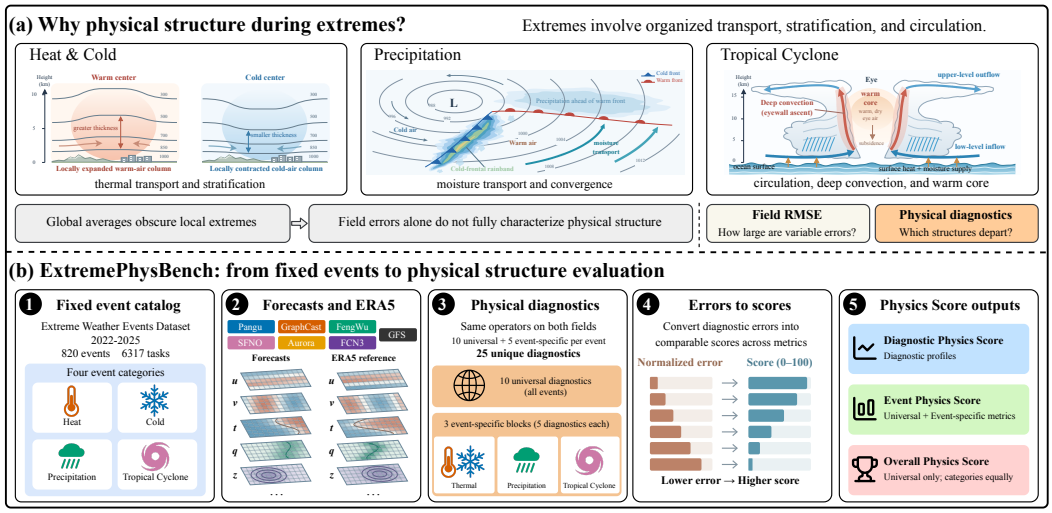

Figure 1: Motivation and overview. **(a)** Importance of physical structures in extreme weather events. **(b)** Evaluation from a fixed event catalog to physical structure evaluation.

structures within a common catalog of extremes and identifies model strengths and weaknesses by physical process.

Moving from hazard verification to physical structure evaluation within extremes presents three challenges. **(i) Event comparability.** Heat, cold, precipitation, and tropical cyclones differ in duration, extent, and motion, making it difficult to define comparable evaluation regions and target times without allowing individual forecasts to determine which samples are scored. **(ii) Diagnostic validity.** Physical diagnostics combine variables and spatial derivatives whose applicability depends on the atmospheric setting; cyclone structure must also be distinguished from position error. Even the reference atmosphere can depart from idealized balance, so a smaller balance residual does not necessarily indicate a more faithful forecast. **(iii) Score aggregation.** Diagnostic errors have different units and scales, and event categories contain different processes and numbers of forecast leads. The evaluation must make these choices consistent across models while preserving the process-level differences that an aggregate score can conceal.

We introduce **ExtremePhysBench** (Figure 1). Its **Extreme Weather Events Dataset** fixes 820 heat, cold, precipitation, and landfalling tropical cyclone (TC) events during 2022–2025, yielding 6,317 tasks independently of forecast outputs. Shared operators compare observable relationships within and between forecast and ERA5 fields, producing Physics Scores for physical structure agreement. We organize related diagnostics into **physical families**, each representing a physical process or structure such as moisture transport or geostrophic flow, and aggregate diagnostic scores within each physical family to obtain physical family Physics Scores. These scores reveal process-specific strengths and weaknesses that aggregate scores can conceal. Across six DWP systems and operational GFS, the leading model varies by event category and physical family (Table 5).

The benchmark makes three contributions. **First**, we construct a dataset of 820 extreme weather events and 6,317 forecast tasks, selected independently of the evaluated forecasts to support consistent comparisons. **Second**, we organize 25 physical diagnostics into physical families and provide Physics Scores that quantify agreement with ERA5 at both the process and aggregate levels. Reporting by physical family identifies which physical structures a model reproduces well or poorly, while hierarchical aggregation prevents processes with more diagnostic components from receiving disproportionate weight. **Third**, we provide a reproducible evaluation pipeline with explicit applicability rules, dimensional errors, and provenance records, enabling traceable comparisons across forecast systems, event categories, and leads.

2 Related Work

Existing work on extreme weather evaluation spans two complementary directions: verifying forecast fields and hazard characteristics, and assessing meteorological reasoning from weather informa-

tion. These efforts establish whether models predict or interpret extreme weather events successfully, motivating a further question: whether forecast fields preserve the thermodynamic and dynamical structures underlying those events. We review both directions below to position physical structure evaluation within this broader landscape.

Benchmarks for extreme weather forecasts. HR-Extreme evaluates forecast field errors within masks for 17 extreme weather types (Ran et al., 2025). ExtremeWeatherBench standardizes cases, observations, and metrics for weather hazards (McGovern et al., 2026). TCBench evaluates tropical cyclone track and intensity (Gomez et al., 2026). These benchmarks establish systematic comparisons of field accuracy and hazard characteristics.

Reasoning about extreme weather. WeatherQA tests whether multimodal models can identify affected areas and severe convection potential from weather maps and contextual information (Ma et al., 2024). It evaluates meteorological reasoning, whereas ExtremePhysBench derives and verifies physical diagnostics from forecast fields.

Table 1 summarizes benchmark coverage. Existing efforts support systematic evaluation of extremes, including derived quantities used for hazard verification. ExtremePhysBench adds direct comparison of physical structure across event categories, with shared evaluation rules and scores organized by process. Appendix A reviews broader evaluation infrastructure.

Table 1: Capabilities relevant to evaluating physical structure in extreme weather forecasts.

Benchmark	Multiple hazards	Forecast eval.	Physical structure	Physical family scores
HR-Extreme	✓	✓	×	×
ExtremeWeatherBench	✓	✓	×	×
TCBench	×	✓	×	×
WeatherQA	✓	×	×	×
ExtremePhysBench	✓	✓	✓	✓

3 Extreme Weather Events Dataset

We construct the 2022–2025 catalog from thermal stress and precipitation products and TC best track records, independently of the evaluated forecasts. Table 2 summarizes the events and tasks; Appendix B specifies source versions, tracking parameters, and selection counts.

Table 2: Event catalog and forecast task inventory.

Event category	Catalog rule	Events	Tasks	Leads (h)
Heat	ERA5-HEAT UTCI; daily max $> p_{95}$; tracked; $\geq$ 3 d and $\geq$ 1600 cells	314	2198	24:24:168
Cold	ERA5-HEAT UTCI; daily min $< p_{05}$; tracked; $\geq$ 3 d and $\geq$ 1600 cells	106	742	24:24:168
Precipitation	GPM IMERG V07B; daily total $\geq p_{95}$ and 10 mm d^{-1}; tracked; $\geq$ 3 d and $\geq$ 300 cells	275	1925	24:24:168
TC	IBTrACS v04r01; offshore wind $\geq$ 34 kt; target 24 h before selected offshore fix	125	1452	6:6:120
Total	–	**820**	**6317**	–

Heat and cold events. ERA5-HEAT provides daily maximum and minimum Universal Thermal Climate Index (UTCI) fields (Di Napoli et al., 2021; Di Napoli, 2020). Heat masks exceed the local calendar day 95th percentile of daily maximum UTCI, and cold masks fall below the 5th percentile of daily minimum UTCI. Thresholds use 1991–2020 and a nominal centered window of 15 days. Tracking these masks yields 314 extreme heat stress and 106 extreme cold stress events. The masks identify thermal stress events; the benchmark then evaluates physical structure of the atmosphere inside them.

Precipitation events. Daily precipitation comes from Version 07B of the Integrated Multi-satellitE Retrievals for the Global Precipitation Measurement mission (GPM IMERG) (Huffman et al., 2020; 2023). We remap the product from $0.1°$ to $0.25°$ and require both the local calendar day 95th percentile over 2001–2020 and 10 mm day^{-1}. The resulting catalog contains 275 events. These precipitation observations define where evaluation takes place; ERA5 supplies the atmospheric reference for evaluating water vapor, transport, and convergence.

From daily masks to events. For heat, cold, and precipitation, masks cover land between $60°$S and $60°$N on a common $0.25°$ grid. TempestExtremes links connected regions across days and

calendar boundaries (Ullrich et al., 2021). Retained regions persist for at least three days and meet the declared size thresholds. Each event has one fixed target at 00 UTC on its second retained day and seven forecasts with leads from 24 to 168 h. This common anchor avoids selecting a peak using forecast performance. Daily masks define the region of an extreme weather event, while the scored fields represent the prescribed instantaneous state; the target need not coincide with the daily extremum.

Landfalling tropical cyclone events. We select 125 storms from IBTrACS version 04r01 (Knapp et al., 2010; Gahtan et al., 2024). Each has an eligible transition from an offshore fix with reported wind of at least 34 kt to a landed fix 6 h later. One transition per storm defines a target 24 h before the offshore fix. Leads span 6–120 h, but initializations before the first synoptic fix with wind of at least 25 kt are excluded. This gives 1,452 tasks, with one to 20 per storm. All 125 targets remain in the catalog; the smaller event cohorts at longer leads follow from this initialization rule.

4 EXTREMEPHYSBENCH: PHYSICAL STRUCTURE EVALUATION

4.1 EVALUATION WITHIN EXTREME WEATHER EVENTS

The benchmark evaluates the physical structure of the atmosphere within a prescribed extreme weather event. Each task fixes an event, an initialization time, a forecast lead, a valid time, and an evaluation region. For heat, cold, and precipitation, the region is the catalog mask. For TCs, the catalog position seeds separate searches for the forecast and ERA5 centers; physical structures are compared relative to their own centers. This separates the representation of the vortex from its displacement.

A diagnostic combines forecast variables into the physical structure of interest. For example, moisture convergence depends jointly on wind, humidity, and their spatial derivatives. To evaluate that physical structure, we apply the same operator $\mathcal{D}_i$ to the forecast state X_k^{F} and ERA5 state X_k^{E} for task k, then compare its outputs:

$$\Delta D_{i,k} = \mathcal{D}_i(X_k^{\mathrm{F}}) - \mathcal{D}_i(X_k^{\mathrm{E}}). \tag{1}$$

This comparison retains relationships between variables before reducing them to an error. Its reference is the physical structure represented in ERA5. Even for a balance residual, the target is the ERA5 residual, which need not be zero.

A common region for comparison. Each model must be evaluated on the same reference samples. The catalog region, ERA5 surface pressure, derivative stencil, and diagnostic availability determine the final support. A forecast cannot remove difficult samples by producing missing values; such output is recorded as invalid. Geostrophic diagnostics exclude the equatorial band, and pressure levels below the reference surface are omitted. Appendix D specifies these rules and the geometry used for TC profiles.

4.2 PHYSICAL DIAGNOSTICS

Table 3 lists the diagnostics and evaluated processes; Appendix C.1 gives the detailed inventory. There are 25 definitions: ten universal diagnostics and five event-specific diagnostics for each of thermal stress, precipitation, and TCs. Heat and cold share the thermal stress configuration, so every event category has 15 applicable diagnostic definitions. Physical families group related diagnostics for reporting; they also keep a process with several diagnostic components from receiving extra weight simply because it has more entries.

Physical structures shared across extreme weather events. The universal physical families distinguish hydrostatic structure (U1), geostrophic flow and shear inferred from geopotential (U2–U3), flow kinematics (U4–U8), and thermodynamic state (U9–U10). Together they resolve the thickness–temperature relationship, circulation geometry, stability, and energy state that individual variable errors cannot identify.

Heat and cold. Transport diagnostics (E1–E3) measure advection of temperature, humidity, and moist static energy. Stability (E4) and frontogenesis (E5) distinguish stratification from changes in temperature gradients induced by horizontal flow, separating processes that a temperature error alone does not resolve.

Table 3: Physical structure diagnostics for extreme weather events.

Physical family	Code	Diagnostic	Evaluated physical structure
Universal			
Hydrostatic structure	U1	Normalized signed hydrostatic residual components	Thickness–temperature consistency by sublayer
Geostrophic flow	U2	Geostrophic wind components, 500	Flow implied by geopotential gradients
	U3	Geostrophic shear components, 850 and 500	Vertical shear of inferred geostrophic flow
Flow kinematics	U4	Relative vorticity, 850	Lower tropospheric rotation
	U5	Horizontal divergence, 850	Lower tropospheric convergence/divergence
	U6	Relative vorticity, 500	Middle tropospheric rotation
	U7	Horizontal divergence, 500	Middle tropospheric convergence/divergence
	U8	Deformation components, 850	Horizontal stretching and shearing
Thermodynamics	U9	Dry N^2, 850–500	Layer dry static stability
	U10	Moist static energy, 850	Thermal, potential, and latent energy state
Heat and cold			
Transport	E1	Temperature advection, 850	Horizontal temperature transport
	E2	Moisture advection, 850	Horizontal humidity transport
	E3	Horizontal MSE advection, 850–500	Partial column MSE advection
Stability	E4	Dry N^2, 850–700	Lower tropospheric dry stability
Frontogenesis	E5	Kinematic temperature frontogenesis, 850	Flow-induced temperature gradient change
Precipitation			
Water amount	E1	Water vapor, 850–300	Partial column water vapor content
Moisture transport and convergence	E2	Vapor transport components, 850–300	Integrated vapor transport vector
	E3	Vapor flux convergence, 850–300	Integrated vapor flux convergence
	E4	Moisture flux convergence, 850	Lower tropospheric moisture convergence
Thermodynamics	E5	MSE difference, 850 and 500	Signed vertical MSE contrast
Tropical cyclones			
Inner core circulation	E1	Tangential wind profile, 850	Circulation within 150 km of the center
Potential vorticity and rotation	E2	Total Ertel potential vorticity, 700	Rotation, stratification, and tilting
	E3	Radial vorticity profile, 500	Radial distribution of rotation
Upper outflow	E4	Radial divergence profile, 200	Outflow within 500 km of the center
Warm core structure	E5	Thickness anomaly, 300–700	Core thickness relative to the environment

Precipitation. Water content (E1), vapor transport (E2), and moisture convergence (E3–E4) distinguish available water from its supply to a region. The vertical moist static energy contrast (E5) adds thermodynamic structure. These diagnostics evaluate the event environment, not rainfall accumulation.

Tropical cyclones. Tangential wind (E1), potential vorticity and radial rotation (E2–E3), upper divergence (E4), and thickness anomaly (E5) resolve circulation, outflow, and core warmth. Comparing these physical structures about independently diagnosed centers separates vortex representation from displacement; track and intensity require additional verification.

4.3 Diagnostic Errors and Physics Scores

Diagnostic errors differ in units and magnitude, so they require normalization before aggregation. We summarize Equation 1 as RMSE $E_{i,k}$, using area weights for fields and declared component or radial weights. Each error is normalized by a fixed scale $S_i = C_i\tau_i$, where C_i is a characteristic magnitude and τ_i a declared tolerance fraction.

An exponential mapping of normalized mean squared error to a bounded score is used by Hodson et al. (2021). We adopt this mapping method to give every diagnostic the same score direction and range:

$$q_{i,k} = 100 \exp\left[-(E_{i,k}/S_i)^2\right], \qquad S_i = C_i\tau_i. \tag{2}$$

Closer agreement with ERA5 gives a higher score, reaching 100 at exact agreement. The bounded map limits the influence of large normalized errors while the outputs retain dimensional errors. Appendix C.2 illustrates the mapping; score changes are not dimensional error growth rates.

The scales and tolerances are fixed across models, events, and leads within each declared regime. None is fitted to the evaluated forecasts or event statistics. The exponential map and tolerance values are benchmark design choices; they are not probabilities or calibrated uncertainty bounds. Appendix C.2 documents the scale construction and consolidated parameter table, and Appendix D gives the error calculations.

4.4 Physical Family and Aggregate Reports

Diagnostic and physical family reports identify processes hidden by aggregate scores. We map errors to scores before averaging leads for each diagnostic, diagnostics within each physical family, and physical families within each universal or event group. The Event Physics Score weights the two groups equally; Category Physics Scores weight events equally. This hierarchy prevents physical families with more diagnostics or events with more forecast leads from receiving greater weight solely because they contain more terms. The weighting scheme and score mapping are explicit benchmark design choices, consistent with guidance on composite indicators (OECD et al., 2008).

The Overall Physics Score has a narrower scope: it averages only the universal group (U1–U10), giving heat, cold, precipitation, and TC equal weight. The event-specific diagnostics describe different physical objects across event categories and are retained in their own category and physical family reports. Arithmetic averaging permits stronger agreement in one physical structure to offset weaker agreement in another, which is why the physical family report is part of the main result. Complete aggregation equations and applicability rules appear in Appendix C.3.

5 Experimental Setup

5.1 Forecast Systems and Shared Fields

We evaluate six DWP systems: Pangu-Weather (**Pangu**) (Bi et al., 2023), GraphCast Operational (**GraphCast**) (Lam et al., 2023; Google DeepMind, 2023), **FengWu** (Chen et al., 2025), Spherical Fourier Neural Operator (**SFNO**) (Bonev et al., 2023), **Aurora** (Bodnar et al., 2025), and FourCastNet 3 (**FCN3**) (Bonev et al., 2025). Operational **GFS** provides the NWP baseline (NCEP, NOAA, n.d.). We do not include ECMWF's Integrated Forecasting System (IFS) because a temporally continuous 0.25° forecast archive covering the full 2022–2025 evaluation period was not available to us. Earth2Studio runs the released DWP checkpoints and retrieves archived GFS forecasts (Earth2Studio Contributors, 2024). Appendix E identifies the model artifacts and inference configurations.

The DWP systems use initial states sampled every 6 h from hourly ARCO-ERA5 (Hersbach et al., 2020; Carver & Merose, 2023); GFS retains its operational initial state. Forecast fields are linearly interpolated to the valid time ERA5 grid at 0.25° before evaluating diagnostics. All inputs to scored diagnostics belong to the 69-variable intersection of outputs available from the seven systems. The same operators, reference support, scales, and aggregation apply to every system. Differences between the DWP and GFS initialization sources remain part of this comparison.

5.2 Tasks and Evaluation Protocol

Evaluation begins in 2022, after the latest disclosed data exposure of the selected DWP systems, including GraphCast's HRES fine tuning through 2021. All systems use the tasks and lead ranges in Table 2, comparing forecasts initialized at different times for each fixed event target.

Every system has a terminal evaluation record for every requested task. Terminal coverage records whether evaluation finished with a declared status; it does not guarantee a finite score for every diagnostic. Formal scoreboard eligibility requires at least 95% terminal coverage in each category. Physical family reports retain their corresponding event counts rather than treating an inapplicable diagnostic as a score of zero. For model differences in the main scoreboard, paired event bootstrap intervals are reported at the nominal 95% level without adjustment for multiple comparisons.

6 Results

We first compare overall and event category performance across forecast systems, then examine physical families to identify process-specific strengths and weaknesses hidden by aggregate scores. Supplementary analyses then examine the relationship to conventional metric RMSE, variation with forecast lead, and sensitivity of the GFS to the verifying analysis field.

6.1 Overall and Event Category Performance

Table 4 compares Overall scores from universal diagnostics and Category scores that include event-specific diagnostics. Overall therefore compares physical structures shared across categories, while Category scores also assess the processes specific to each hazard; Overall is not the mean of the four Category scores. Appendix F.2 provides complementary aggregate views.

FengWu leads Overall (46.03); paired differences for all five other DWP systems have nominal 95% intervals below zero, including Aurora at -3.44 points $[-3.90, -2.93]$. All six DWP systems exceed GFS by 8.69–19.21 points under ERA5 verification, with initialization differences limiting broader interpretation. FengWu leads heat, cold, and precipitation, while Aurora leads TC. Event distributions overlap substantially (Appendix F.3); a higher category mean does not imply an advantage on every event.

Table 4: Overall and Category Physics Scores by forecast system.

System	Overall	Heat	Cold	Precipitation	TC	Score − GFS	Δ vs. FengWu [95% CI]
FengWu	**46.03**	**59.64**	**39.85**	**34.80**	57.80	+19.21	–
Aurora	42.59	54.37	36.82	29.81	**62.13**	+15.77	−3.44 [−3.90,−2.93]
GraphCast	42.42	56.22	37.57	31.98	60.91	+15.59	−3.61 [−4.20,−2.95]
Pangu	42.05	55.55	34.85	31.29	58.36	+15.22	−3.98 [−4.52,−3.28]
SFNO	37.96	50.73	31.03	26.99	58.60	+11.14	−8.07 [−8.66,−7.47]
FCN3	35.51	45.11	28.72	23.50	54.36	+8.69	−10.52 [−11.16,−9.79]
GFS	26.82	35.22	22.86	18.18	41.66	–	–

6.2 Physical Families Reveal Different Strengths

Table 5 uses common complete events across models, weighting applicable events equally within each physical family. Applicability varies, notably for geostrophy and TC warm core structure, so averaging the displayed physical family means need not reproduce Category scores. Comparisons use fixed physical family scales and are descriptive; no physical family confidence intervals were computed. Figure 2 resolves physical family differences into individual diagnostics; Appendix F.4 provides detailed diagnostic views.

Heat: a broad advantage across processes. FengWu leads all seven physical families, including transport, stability, and frontogenesis. Its category advantage therefore spans both the redistribution of heat and moisture and the stratification and deformation of temperature fields. Agreement across these physical families shows that its heat ranking is not driven by a single physical family.

Cold: different leaders across physical families. FengWu leads the three physical families for thermal processes, but Aurora leads hydrostatic structure (12.26 versus 11.95). Frontogenesis scores are close for FengWu and GraphCast (26.55 versus 26.43). All systems score lower on the physical families for thermal processes in cold than in heat under shared scales; this sample comparison alone does not establish a difference in predictability.

Precipitation: water amount and supply differ. GraphCast leads water content (40.23), whereas FengWu leads moisture transport and convergence (34.74) and vertical energy contrast (20.81). A category score thus conceals differences between available vapor, its supply, and thermodynamic structure. The event-specific physical family for thermodynamics measures vertical energy contrast; the universal physical family measures stability and energy state.

Tropical cyclones: circulation, outflow, and warmth separate. Aurora leads circulation (87.23) and potential vorticity and rotation (61.12); GraphCast leads warm core structure (86.34). Outflow scores are close for FengWu (44.16), Pangu (44.10), and Aurora (44.03). Common applicable samples contain 54 events for circulation, 55 for potential vorticity and outflow, and 33 for warm core structure, whose reference environment requirement limits support. These comparisons concern independently centered physical structures and complement track and intensity verification.

6.3 Supplementary Analyses

Relationship to field accuracy. We compare Physics Scores with RMSE for u, v, t, q, z at 850, 700, and 500 hPa on the same event support. Matching the evaluation regions isolates differences between

Table 5: Physics Scores by event category and physical family.

Model	**Universal**				**Event-specific**		
Heat	Hyd.	Geos.	Kinem.	Thermo.	Adv.	Stab.	Frontog.
GFS	12.75	38.39	22.72	45.20	41.07	58.62	22.74
Pangu	28.10	56.30	51.41	59.82	60.40	76.81	49.10
GraphCast	24.09	49.56	54.40	63.32	62.75	78.93	51.51
FengWu	**33.18**	**60.01**	**56.22**	**64.28**	**64.55**	**80.20**	**52.52**
SFNO	21.88	45.87	50.18	50.20	57.81	71.53	48.49
Aurora	30.13	47.42	52.06	57.91	59.60	75.95	49.42
FCN3	20.09	47.27	38.81	49.94	46.54	69.22	37.63
Cold	Hyd.	Geos.	Kinem.	Thermo.	Adv.	Stab.	Frontog.
GFS	3.51	18.01	22.10	34.73	29.29	38.13	10.79
Pangu	8.33	28.03	41.53	45.89	40.81	52.55	22.58
GraphCast	9.16	25.49	46.62	48.01	46.04	55.57	26.43
FengWu	11.95	**32.07**	**48.73**	**49.56**	**47.35**	**58.14**	**26.55**
SFNO	5.79	21.85	41.88	34.11	40.14	43.15	24.87
Aurora	**12.26**	28.95	44.88	44.89	44.25	52.94	25.16
FCN3	5.72	23.97	33.81	38.03	33.41	44.46	18.01
Precipitation	Hyd.	Geos.	Kinem.	Thermo.	Vapor	Moist.	Thermo.
GFS	9.17	18.85	7.75	38.27	29.34	14.75	9.13
Pangu	19.79	34.62	27.44	51.82	39.57	30.77	16.59
GraphCast	14.09	33.32	29.71	53.60	**40.23**	33.49	19.67
FengWu	**25.22**	**38.51**	**31.31**	**55.96**	39.47	**34.74**	**20.81**
SFNO	17.31	29.88	26.72	44.50	32.04	28.69	11.91
Aurora	22.39	32.88	26.48	48.77	35.27	30.29	14.87
FCN3	14.68	28.35	19.78	41.67	28.64	22.86	10.66
Tropical Cyclone	Hyd.	Geos.	Kinem.	Thermo.	Circ.	PV	Outflow	Warm
GFS	26.03	39.26	35.48	56.65	62.84	30.05	21.18	76.08
Pangu	27.01	58.86	62.79	69.97	79.98	51.77	44.10	83.47
GraphCast	27.15	**61.83**	**65.17**	**71.60**	83.40	59.44	42.84	**86.34**
FengWu	**38.19**	57.87	62.06	69.95	69.72	49.60	**44.16**	79.35
SFNO	25.90	61.21	63.45	65.65	81.81	55.77	43.46	80.95
Aurora	36.79	59.91	64.79	69.28	**87.23**	**61.12**	44.03	83.14
FCN3	29.86	57.62	56.16	61.94	81.74	51.75	29.19	77.58

Higher is better; bold marks each physical family's maximum; Hyd.: hydrostatic balance; Geos.: geostrophic wind and shear; Kinem.: flow kinematics; Thermo.: thermodynamics; Adv.: heat and moisture advection; Stab.: static stability; Frontog.: frontogenesis; Vapor: water vapor content; Moist.: moisture transport and convergence; Circ.: inner-core circulation; PV: potential vorticity and rotation; Outflow: upper-level outflow; Warm: warm-core structure. The universal and precipitation-specific Thermo. physical families contain different diagnostics.

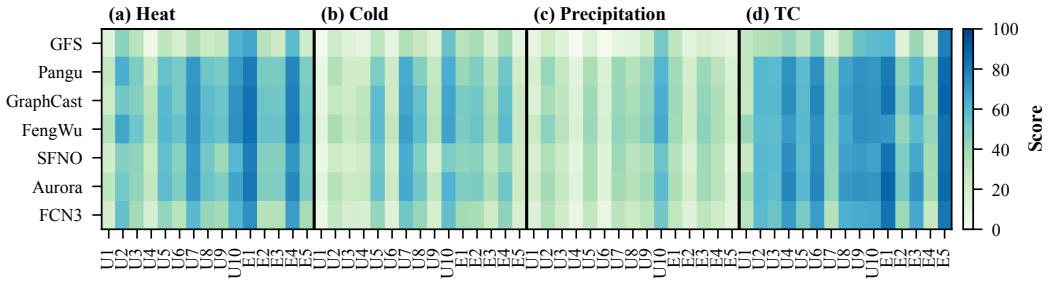

Figure 2: Diagnostic Physics Scores by event category; codes follow Table 3. Higher is better. Each diagnostic uses its applicable events, unlike the common complete support in Table 5.

the metrics from differences in spatial sampling. Appendix F.1 defines the RMSE composite. The formal Physics Score leaderboard and the RMSE summary over these 15 fields have a model rank correlation of $\rho = 0.893$ Overall. FengWu ranks first and GFS seventh under both summaries; the heat and precipitation orders agree exactly. Yet Aurora rises from fourth by RMSE to second by Overall Physics Score. TC has the largest reordering ($\rho = 0.703$): Pangu ranks first by RMSE and fourth by Physics Score, while Aurora is tied second by RMSE and first by Physics Score.

Within models, events with lower field errors also tend to have higher Physics Scores. All 28 Spearman associations between RMSE composite skill and Event Physics Score are positive, and their nominal bootstrap 95% intervals exclude zero. Median correlations across systems are 0.611 for heat, 0.762 for cold, 0.534 for precipitation, and 0.867 for TC. These comparisons use 314 heat,

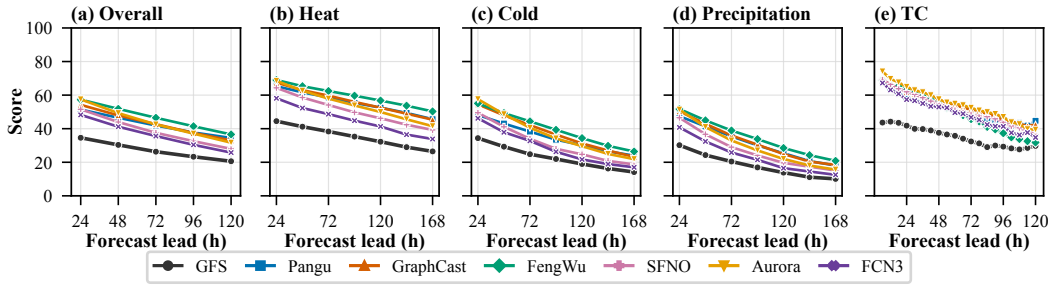

Figure 3: Physics Scores by forecast lead. Higher is better. TC cohorts shrink at longer leads; curves therefore reflect both forecast lead and event composition.

106 cold, 275 precipitation, and 55 TC events with common valid scores. Appendix F.1 also gives a supplementary rank comparison on matched event scores, whose aggregation differs from the formal Overall definition. Physical structure diagnostics complement RMSE by locating departures in transport, stratification, rotation, and other physical processes.

Variation with forecast lead. All systems score lower at 120 h than at 24 h (Figure 3); TC trends also reflect a shrinking event cohort (Appendix F.5). A smaller score decline does not necessarily indicate better physical structure agreement, because systems start from different score levels and the score mapping is nonlinear.

Sensitivity to the verifying analysis. As a supplementary reference sensitivity experiment, we rescore GFS against its archived 0-h analyses while holding tasks, operators, scales, and aggregation fixed. This change raises GFS's Overall score from 26.82 to 30.33, with the largest category change in TC ($+8.55$). This experiment quantifies the effect of the verifying analysis; the main leaderboard retains ERA5 as the common reference for all systems (Appendix F.6).

7 Discussion and Conclusion

ExtremePhysBench evaluates physical structure within extreme weather events. Positive associations with RMSE, alongside differences in model ranking, support reporting physical structure agreement together with conventional field accuracy.

Physical families connect model performance to atmospheric processes. Physical family Physics Scores give model differences a physical interpretation. Different systems lead water content and moisture transport for precipitation, and circulation and warm core structure for TCs. These distinctions connect forecast evaluation to how moisture is supplied and redistributed, and how circulation and thermal structure are represented within storms. Following an aggregate score through physical families to individual diagnostics provides a way to identify the processes associated with a model's performance and examine them in specific extreme weather events.

Complementary evaluation across extreme weather events. Shared event masks and independently diagnosed TC centers provide a common basis for comparing physical structures with ERA5. Universal diagnostics connect the evaluation across event categories, while event-specific diagnostics retain the processes characteristic of each category. This organization supports a common comparison of forecast systems together with a physical interpretation of their performance in different extreme weather events. Physics Scores summarize agreement with the reference, while dimensional errors and physical family reports retain the detail needed to interpret that agreement. The resulting view is complementary to conventional accuracy metrics and hazard verification, including TC track and intensity.

Reproducible comparisons and future work. Fixed tasks, explicit scoring rules, and saved errors and provenance make comparisons repeatable. Applying the same evaluation to successive model versions can track whether improvements extend across physical families and event categories. Future work can extend the catalog and physical families and evaluate probabilistic forecasts, bringing the representation of uncertainty into the assessment of physical structure within extreme weather events.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and polishing the manuscript, checking consistency, and preparing LaTeX, as well as literature retrieval and discovery to identify potentially relevant related work. AI outputs were not treated as scientific evidence or cited as sources. The authors retain responsibility for verifying the cited literature and for the accuracy and integrity of the manuscript, including its scientific claims, numerical results, and references.

ETHICS STATEMENT

ExtremePhysBench is intended for research evaluation. Its catalog covers four categories during 2022–2025, with land and latitude restrictions for gridded events and USA agency fields for TC selection. Polar, oceanic, and untracked hazards are consequently underrepresented. Physical structure agreement with ERA5 should be considered alongside hazard verification and operational requirements before drawing conclusions about forecast use. Upstream ERA5-HEAT, GPM IMERG, and IBTrACS archives remain subject to their providers' terms and are not redistributed in full.

REPRODUCIBILITY STATEMENT

Benchmark codes are available at `https://anonymous.4open.science/r/ExtremePhysBench-REVIEW/`. Benchmark data resources are available at `https://zenodo.org/records/22951825?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTc5MDMyMDEwNCwiZXhwIjoxNzk4Njc1MTk5fQ.eyJpZCI6IjA5NjM2ZTIwLTlmMzEtNDBkZi05MDI1LTIwODhhNWM3ODJjYyIsImRhdGEiOnt9LCJyYW5kb20iOiJhN2Q0ZjhlY2MwYWU1MmQ5MWZmNDk0YTY4NWY2NmViMCJ9.7Ml1ytsH9BZtmTE5TOrVa50YOEZFTjXCYY5RgZ9j19y10R8DJg13-q35aq0RALVFXFc9YpxWYUbefIbkZuCbzQ`. These include the event catalog, forecast task inventory, physical diagnostics and scoring implementation, and instructions for reproducing the evaluation.

REFERENCES

Peter C. Banacos and David M. Schultz. The use of moisture flux convergence in forecasting convective initiation: Historical and operational perspectives. *Weather and Forecasting*, 20(3):351–366, 2005. doi: 10.1175/WAF858.1.

George S. Benton and Mariano A. Estoque. Water-vapor transfer over the north american continent. *Journal of Meteorology*, 11(6):462–477, 1954. doi: 10.1175/1520-0469(1954)011⟨0462:WVTOTN⟩2.0.CO;2.

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. doi: 10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, et al. A foundation model for the Earth system. *Nature*, 641:1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. doi: 10.1029/2023GL107377.

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

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

Robert W. Carver and Alex Merose. ARCO-ERA5: An analysis-ready cloud-optimized reanalysis dataset. Conference abstract and oral presentation at the 22nd Conference on Artificial Intelligence for Environmental Science, 103rd AMS Annual Meeting, 2023. URL https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/415842. Denver, Colorado, 2023-01-09, Paper 4A.1.

Jule G. Charney. On the scale of atmospheric motions. *Geofysiske Publikasjoner*, 17(2):3–17, 1948. URL https://geofysikk.org/NGF/GeoPub/NGF_GP_Vol17_no2.pdf.

Kang Chen, Tao Han, Fenghua Ling, Junchao Gong, Lei Bai, Xinyu Wang, Jing-Jia Luo, Ben Fei, Wenlong Zhang, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Communications Earth & Environment*, 6(1):518, 2025. doi: 10.1038/s43247-025-02502-y.

Claudia Di Napoli. Thermal comfort indices derived from ERA5 reanalysis, 2020. Dataset, version 1.1; accessed 2026-08-24.

Claudia Di Napoli, Christopher Barnard, Christel Prudhomme, Hannah L. Cloke, and Florian Pappenberger. ERA5-HEAT: A global gridded historical dataset of human thermal comfort indices from climate reanalysis. *Geoscience Data Journal*, 8(1):2–10, 2021. doi: 10.1002/gdj3.102.

Earth2Studio Contributors. NVIDIA Earth2Studio. https://github.com/NVIDIA/earth2studio, 2024. Open-source AI weather and climate workflow toolkit; repository tag 0.14.0 used in this work.

Hans Ertel. Ein neuer hydrodynamischer wirbelsatz. *Meteorologische Zeitschrift*, 59(9):277–281, 1942. URL https://empslocal.ex.ac.uk/people/staff/gv219/classics.d/Schubert-Ertel04.pdf. English translation reproduced as paper 1 in Schubert et al. (2004).

Jennifer Gahtan, Kenneth R. Knapp, Carl J. Schreck, III, Howard J. Diamond, James P. Kossin, and Michael C. Kruk. International best track archive for climate stewardship (IBTrACS) project, version 4r01, 2024. Dataset.

Milton Gomez, Marie McGraw, Saranya Ganesh S., Frederick Iat-Hin Tam, Ilia Azizi, Samuel Darmon, Monika Feldmann, Stella Bourdin, Louis Poulain-Auzéau, Suzana J. Camargo, Jonathan Lin, Dan Chavas, Chia-Ying Lee, Ritwik Gupta, Andrea Jenney, and Tom Beucler. TCBench: A benchmark for tropical cyclone track and intensity forecasting at the global scale. arXiv preprint arXiv:2601.23268, 2026. URL https://arxiv.org/abs/2601.23268.

Google DeepMind. GraphCast: Learning skillful medium-range global weather forecasting. GitHub repository, release v0.1, 2023. URL https://github.com/google-deepmind/weathernext/tree/858301cde5de5c728f8172f782dafba1ea07ac2e. Version 0.1, released 2023-10-31; accessed 2026-08-10.

Google LLC. WeatherBench-X: A modular framework for evaluating weather forecasts. GitHub software release, 2025. URL https://github.com/google-research/weatherbenchX/releases/tag/v2025.02.1. Version v2025.02.1; successor to the WeatherBench 2 evaluation code; accessed 2026-08-18.

Gregory J. Hakim and Sanjit Masanam. Dynamical tests of a deep learning weather prediction model. *Artificial Intelligence for the Earth Systems*, 3(3):e230090, 2024. doi: 10.1175/AIES-D-23-0090.1.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Timothy O. Hodson, Thomas M. Over, and Sydney S. Foks. Mean squared error, deconstructed. *Journal of Advances in Modeling Earth Systems*, 13(12):e2021MS002681, 2021. doi: 10.1029/2021MS002681.

James R. Holton and Gregory J. Hakim. *An Introduction to Dynamic Meteorology*, volume 88 of *International Geophysics*. Academic Press, 5 edition, 2013. ISBN 978-0-12-384866-6. doi: 10.1016/C2009-0-63394-8.

Brian J. Hoskins, Michael E. McIntyre, and Andrew W. Robertson. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, 111(470):877–946, 1985. doi: 10.1002/qj.49711147002.

George J. Huffman, David T. Bolvin, Dan Braithwaite, Kuo-Lin Hsu, Robert J. Joyce, Christopher Kidd, Eric J. Nelkin, Soroosh Sorooshian, Erich F. Stocker, Jackson Tan, David B. Wolff, and Pingping Xie. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In Vincenzo Levizzani, Christopher Kidd, Dalia B. Kirschbaum, Christian D. Kummerow, Kenji Nakamura, and F. Joseph Turk (eds.), *Satellite Precipitation Measurement*, volume 67 of *Advances in Global Change Research*, pp. 343–353. Springer International Publishing, 2020. doi: 10.1007/978-3-030-24568-9_19.

George J. Huffman, Erich F. Stocker, David T. Bolvin, Eric J. Nelkin, and Jackson Tan. GPM IMERG final precipitation L3 1 day 0.1 degree x 0.1 degree V07, 2023. Dataset.

IPCC. Weather and climate extreme events in a changing climate. Climate Change 2021: The Physical Science Basis, Chapter 11, 2021. URL https://www.ipcc.ch/report/ar6/wg1/chapter/chapter-11/.

Emma Kasteleyn, Timo Maier, Axel Lauer, Veronika Eyring, Pierre Gentine, and Ana Lucic. PhysMetrics.Weather: An evaluation framework for physical consistency in ML weather models. arXiv preprint arXiv:2606.10642, 2026. URL https://arxiv.org/abs/2606.10642.

Kenneth R. Knapp, Michael C. Kruk, David H. Levinson, Howard J. Diamond, and Charles J. Neumann. The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. doi: 10.1175/2009BAMS2755.1.

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. doi: 10.1126/science.adi2336.

Ruize Li, Zhibin Wen, Tao Han, Hao Chen, Fenghua Ling, Wei Zhang, Song Guo, and Lei Bai. RealBench: Benchmarking data-driven numerical weather forecasting under operational conditions and extreme event challenges. arXiv preprint arXiv:2605.24945, 2026. URL https://arxiv.org/abs/2605.24945.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110v2.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. WeatherQA: Can multimodal language models reason about severe weather? arXiv preprint arXiv:2406.11217, 2024. URL https://arxiv.org/abs/2406.11217v2.

Amy McGovern, Taylor Mandelbaum, Daniel Rothenberg, Nicholas Loveday, Corey Potvin, Montgomery Flora, Linus Magnusson, Eric Gilleland, and John Allen. Extreme weather bench: A framework and benchmark for evaluation of high-impact weather. arXiv preprint arXiv:2605.01126, 2026. URL https://arxiv.org/abs/2605.01126.

NCEP, NOAA. The global forecast system (GFS). https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php, n.d. Accessed 2026-08-11.

J. David Neelin and Isaac M. Held. Modeling tropical convergence based on the moist static energy budget. *Monthly Weather Review*, 115(1):3–12, 1987. doi: 10.1175/1520-0493(1987)115⟨0003:MTCBOT⟩2.0.CO;2.

Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. ClimateLearn: Benchmarking machine learning for weather and climate modeling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75009–75025. Curran Associates, Inc., 2023. doi: 10.52202/075280-3279. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed73c36e771881b232ef35fa3a1dec14-Paper-Datasets_and_Benchmarks.pdf.

NVIDIA. Modulus checkpoints: Spherical Fourier neural operator. NVIDIA NGC model card, version 0.1.0, 2024. URL https://catalog.ngc.nvidia.com/orgs/nvidia/modulus/models/sfno_73ch_small. Accessed 2026-08-10.

NVIDIA. Earth-2 checkpoints: FourCastNet 3. Hugging Face model card, 2025. URL https://huggingface.co/nvidia/fourcastnet3/tree/76ef0c60237e458b33196ba027134e27f3fc4538. Accessed 2026-08-10.

NVIDIA. earth2studio.models.px: Prognostics — Earth2Studio. Versioned software documentation, version 0.14.0, 2026. URL https://nvidia.github.io/earth2studio/v/0.14.0/modules/models_px.html. Accessed 2026-08-10.

OECD, European Union, and European Commission Joint Research Centre. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris, 2008. doi: 10.1787/9789264043466-en.

OpenEarthLab. FengWu: Released models and inference code. GitHub repository, n.d. URL https://github.com/OpenEarthLab/FengWu. Accessed 2026-09-26.

Olivier C. Pasche, Jonathan Wider, Zhongwei Zhang, Jakob Zscheischler, and Sebastian Engelke. Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, 4(1):e240033, 2025. doi: 10.1175/AIES-D-24-0033.1.

Sverre Petterssen. Contribution to the theory of frontogenesis. *Geofysiske Publikasjoner*, 11(6):5–27, 1936. URL https://geofysikk.org/NGF/GeoPub/NGF_GP_Vol11_no6.pdf.

Nian Ran, Peng Xiao, Yue Wang, Wesley Shi, Jianxin Lin, Qi Meng, and Richard Allmendinger. HR-Extreme: A high-resolution dataset for extreme weather forecasting. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.18885v2.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. doi: 10.1029/2020MS002203.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. 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. doi: 10.1029/2023MS004019.

Tobias Selz and George C. Craig. Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophysical Research Letters*, 50(20):e2023GL105747, 2023. doi: 10.1029/2023GL105747.

Akshay Subramaniam, Dale Durran, David Pruitt, Nathaniel Cresswell-Clay, and William Yik. Imposing the fundamental dynamical constraint of hydrostatic balance to improve global ML weather prediction. arXiv preprint arXiv:2506.08285, 2025. URL https://arxiv.org/abs/2506.08285.

Paul A. Ullrich, Colin M. Zarzycki, Elizabeth E. McClenny, Marielle C. Pinheiro, Alyssa M. Stansfield, and Kevin A. Reed. TempestExtremes v2.1: A community framework for feature detection, tracking, and analysis in large datasets. *Geoscientific Model Development*, 14(8):5023–5048, 2021. doi: 10.5194/gmd-14-5023-2021.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. arXiv preprint arXiv:2406.19314, 2024. URL https://arxiv.org/abs/2406.19314v2. Version 2, revised April 2025.

Hugh E. Willoughby. Gradient balance in tropical cyclones. *Journal of the Atmospheric Sciences*, 47(2):265–274, 1990. doi: 10.1175/1520-0469(1990)047⟨0265:GBITC⟩2.0.CO;2.

APPENDIX CONTENTS

A Comparison with Related Evaluation Frameworks **16**

B Extreme Weather Events Dataset Details **17**

B.1 Source Data and Common Conventions 17

B.2 Gridded Event Construction 17

B.3 Landfalling Tropical Cyclones 18

B.4 Catalog and Task Records 18

B.5 Catalog Summary 18

C Benchmark Inventory, Scales, and Aggregation **19**

C.1 Diagnostic and Task Inventory 19

C.2 Fixed Physics Score Scales 20

C.3 Explicit Hierarchical Estimator 21

D Diagnostic Operators and Error Geometry **24**

D.1 Notation and Spherical Operators 24

D.2 From Governing Relations to Observable Physical Structure 24

D.3 Reference-Only Support 27

D.4 Universal Diagnostics U1–U10 27

D.5 Thermal-Stress Diagnostics E1–E5 28

D.6 Precipitation Diagnostics E1–E5 29

D.7 Tropical-Cyclone Diagnostics E1–E5 29

D.8 Diagnostic Error Geometries 30

E Forecast Systems and Inference Details **30**

E.1 Earth2Studio Execution Layer 30

E.2 Initial Conditions and Forecast Outputs 32

F Additional Experimental Results **33**

F.1 Conventional RMSE Comparison 33

F.2 Aggregate Model Comparisons 35

F.3 Event Distributions and Paired Comparisons 35

F.4 Physical Structure Profiles by Diagnostic 35

F.5 Detailed Views by Forecast Lead 40

F.6 Sensitivity to the GFS Verifying Analysis 45

A Comparison with Related Evaluation Frameworks

The main text discusses benchmarks that focus on extreme or severe weather. Table 6 places those tasks alongside broader weather evaluation frameworks. The comparison describes the published evaluation targets, rather than treating the absence of a particular experiment as a restriction on what a software package could support.

Table 6: Published evaluation targets of representative frameworks. Entries describe their focus rather than all possible extensions.

Framework	Evaluation setting	Principal comparison
HR-Extreme	Extreme weather cases	Forecast field errors within event masks.
ExtremeWeatherBench	Curated weather hazards	Observations and verification metrics specific to each phenomenon.
TCBench	Tropical cyclones	Track and intensity forecasts.
WeatherQA	Severe weather reasoning	Affected areas and severe convection potential from multimodal inputs.
WeatherBench / WeatherBench 2	Global weather forecasts	Accuracy of forecast variables under shared evaluation protocols.
WeatherBench-X	Weather evaluation software	Modular computation of forecast verification metrics.
ClimateLearn	Weather and climate tasks	Includes temperature errors conditioned on extremes.
RealBench	Operational conditions and extremes	Performance under operational forecasting constraints.
PhysMetrics.Weather	Global forecast samples	Conservation, spectral, dynamical, and thermodynamic diagnostics.
ExtremePhysBench	**Cataloged extreme weather events**	**Physical structure by diagnostic, physical family, event category, and lead.**

General forecast evaluation. WeatherBench and WeatherBench 2 establish common datasets and metrics for comparing global forecast fields (Rasp et al., 2020; 2024). WeatherBench-X provides modular evaluation infrastructure (Google LLC, 2025). These resources make conventional comparisons reproducible, but a global aggregate alone cannot locate physical structure departures inside cataloged extremes. ClimateLearn includes an Extreme-ERA5 experiment in which temperature errors are evaluated on pixels selected by percentile thresholds (Nguyen et al., 2023). This conditions field accuracy on extreme temperatures; it does not organize the evaluation around the physical families used here. RealBench examines forecast performance under operational conditions and extreme weather event challenges (Li et al., 2026), which also makes initialization and deployment conditions part of evaluation.

Evaluation within extreme weather events. HR-Extreme provides extreme weather cases from HRRR and reports errors within event masks (Ran et al., 2025). ExtremeWeatherBench standardizes cases, observations, and metrics relevant to hazards (McGovern et al., 2026). TCBench focuses on tropical cyclone track and intensity (Gomez et al., 2026). WeatherQA uses meteorological images and expert discussions to evaluate reasoning about severe weather (Ma et al., 2024). These evaluations differ in inputs, outputs, spatial scales, and targets. ExtremePhysBench adds comparisons of physical diagnostics derived from forecast fields within an independently constructed event catalog. Its four categories do not imply broader hazard coverage than those resources.

Physical diagnostics. Studies of learned dynamics and model imbalance motivate examining physical relationships alongside field accuracy (Hakim & Masanam, 2024; Selz & Craig, 2023; Bonavita, 2024; Subramaniam et al., 2025). PhysMetrics.Weather brings conservation, spectral, dynamical, and thermodynamic diagnostics into a common evaluation of global forecasts (Kasteleyn et al., 2026). Its evaluation on global samples addresses a different sampling question from repeated evaluation inside extreme weather events. ExtremePhysBench selects its physical families to describe physical structures within those events and compares their forecast values with ERA5. Neither the availability of diagnostic fields nor agreement with a reanalysis establishes closure of all prognostic equations from archived snapshots.

B EXTREME WEATHER EVENTS DATASET DETAILS

The catalog covers 2022–2025 as one evaluation period. These years are not a train, validation, or test partition for the forecast systems; model exposure dates are reported separately in Appendix Table 14. Event construction used only the stated source products, and its scripts do not read forecast outputs.

B.1 SOURCE DATA AND COMMON CONVENTIONS

Thermal events used ERA5-HEAT daily minimum and maximum Universal Thermal Climate Index (UTCI) fields (Di Napoli et al., 2021; Di Napoli, 2020). The files contain `utci_daily_min` and `utci_daily_max` on a $0.25°$ grid from $90°$N to $60°$S and $180°$W to $179.75°$E. A 1991–2020 baseline supplied the local calendar day p_{05} of daily minimum UTCI and p_{95} of daily maximum UTCI. Precipitation events used Version 07B Final Run daily precipitation from the Integrated Multi-satellitE Retrievals (IMERG) for the Global Precipitation Measurement (GPM) mission (Huffman et al., 2020; 2023), in mm day^{-1}. The source $0.1°$ product was conservatively remapped with the Climate Data Operators `remapcon` command to a $0.25°$, 1440×721 longitude–latitude grid after selecting precipitation and replacing source missing values with zero. Its local p_{95} baseline used 2001–2020.

Both baselines pooled a nominal window of 15 days centered on each calendar date across all reference years. The implementation truncated rather than wrapped the window at the start and end of the calendar, so dates near those boundaries used 8–14 days and all other dates used 15. The baselines indexed calendar days against the 366 days of year 2000; February 29 was the mean of February 28 and March 1. Event masks followed the proleptic Gregorian calendar. Event construction used complete daily source sequences, and no missing day was replaced by a neighboring day. Comparisons involving a nonfinite source value or threshold did not enter a binary event mask.

Thermal and precipitation masks used the same ERA5 land mask, aligned to each source grid by nearest grid point. A cell was land when `lsm`≥ 0.5, and gridded detection was restricted to $60°$S–$60°$N. Longitude was normalized only when required to align the source and mask grids; released centroids and task bounds use $[0, 360)$.

B.2 GRIDDED EVENT CONSTRUCTION

The heat, cold, and precipitation pipelines used the following fixed sequence.

1. For each calendar day, heat cells satisfied `utci_daily_max`$> p_{95}$, cold cells satisfied `utci_daily_min`$< p_{05}$, and precipitation cells satisfied both precipitation$\geq p_{95}$ and precipitation≥ 10 mm day^{-1}. All masks were intersected with the common land and latitude mask. A core diagnostic mask at 50 mm day^{-1} was retained in the intermediate precipitation product but did not determine event eligibility.
2. TempestExtremes `DetectBlobs` labeled spatial components with diagonal connectivity. Monthly files were listed chronologically, and `StitchBlobs` processed the complete 2022–2025 sequence so objects could cross file, month, year, and calendar boundaries. It required at least three daily steps and at least 1,600 grid cells per thermal blob or 300 grid cells per precipitation blob. Because eligibility counts cells on a regular latitude–longitude grid rather than physical area, the implied area threshold decreases with absolute latitude within the $60°$ bounds.
3. `BlobStats` recorded the event ID, day index, date, area, centroid, and latitude–longitude bounds. The released script applied a fixed start buffer of nine days and removed objects whose day 0 date preceded 2022-01-10; IDs were then renumbered consistently in the catalog and event label NetCDF. No further merging or post hoc deduplication was applied: one stitched label defined one released event.
4. A task used 00 UTC on event day 1, the second retained catalog day, as its target. Because every retained object had at least three days, this was the earliest fixed index with a retained day on both sides. It supplied one common anchor without selecting a peak or consulting forecast output. Each event produced seven tasks at 24–168 h in 24 h increments. The daily catalog mask fixed the evaluation region, while the forecast represented

the prescribed instantaneous target time. Thus, 00 UTC need not coincide with the subdaily UTCI extremum or precipitation episode; conditioning is on the event day and mask rather than confirmation that the instantaneous state is itself extreme.

TempestExtremes produced 315 heat, 107 cold, and 278 precipitation objects after its duration and size rules. The start boundary filter removed 1, 1, and 3 objects, respectively, leaving 314, 106, and 275 events. Every retained object contained days 0, 1, and 2, permitting day 1 to serve as target. Seven lead times per event then yielded 2,198, 742, and 1,925 tasks. The buffer of nine days reduces the risk of left boundary truncation but does not prove that every retained object begins at its complete physical onset. No right boundary filter was applied. Two retained heat objects remained active on 2025-12-31, the last available source day, so their reported durations may be right censored; no retained cold or precipitation object reached that boundary. Counts before the TempestExtremes duration and size rules depend on transient daily components and are not treated as candidate events.

B.3 Landfalling Tropical Cyclones

Tropical cyclone (TC) construction used version 04r01 of the International Best Track Archive for Climate Stewardship (IBTrACS) and did not use TempestExtremes (Knapp et al., 2010; Gahtan et al., 2024). We retained storms whose complete recorded lifecycle fell within 2022–2025, selected 00, 06, 12, and 18 UTC fixes with valid positions, and identified every consecutive transition from an offshore fix with `DIST2LAND`>0 km to a landed fix 6 h later with `DIST2LAND`≤ 0 km. The offshore fix required `USA_WIND`≥ 34 kt. When a storm had multiple eligible transitions, the deterministic order was strongest USA wind, lowest USA pressure, then earliest time, all evaluated at the offshore fix. The target was 24 h before that offshore fix, or 30 h before the following landed fix; the nearest best track fix to the target had to be within 3 h. Task leads ran from 6 to 120 h in 6 h increments. TC initializations were retained only from the first synoptic fix with `USA_WIND`≥ 25 kt onward, excluding weaker precursor states.

The time subset contained 417 storm IDs, of which 414 had complete lifecycles inside the period. Among these, 176 had a transition from offshore to land, 165 also had a retained synoptic fix with `USA_WIND`≥ 25 kt, and 125 satisfied the remaining transition, intensity, target, and lifecycle rules. The 125 events would yield 2,500 tasks on a complete grid of 20 leads; the event definition excludes 1,048 leads preceding that threshold and retains 1,452 tasks, or 1–20 tasks per storm. These excluded leads are not cases of missing forecast output and do not remove any of the 125 TC target times from the catalog. The final basin counts were 45 western North Pacific, 24 North Atlantic, 23 South Indian, 15 North Indian, 11 eastern North Pacific, and 7 South Pacific events. At the selected offshore fix, USA wind ranged from 35 to 165 kt (median 65 kt); counts in the ranges 34–63, 64–82, 83–95, 96–112, 113–136, and at least 137 kt were 60, 26, 13, 13, 12, and 1.

B.4 Catalog and Task Records

The gridded event catalogs store `eventid`, day index, date, centroid, area in m^2, and geographic bounds. Their event label NetCDF files store the renumbered integer label for every day and grid cell. Gridded task tables add a stable `task_id`, category, target and initialization times, lead, integration step, and the day 1 geometry. The TC event table stores storm identifiers, basin, target and transition fields, search coordinates, intensity, source variable names, and retained synoptic lifecycle fields; its task table adds one row for every retained lead. Thus, every reported result can be joined back to an event, valid time, initialization, lead, and either the gridded catalog mask geometry or, for TC, the IBTrACS search seed without consulting a forecast value.

Large upstream ERA5-HEAT, GPM IMERG, and full IBTrACS files are not redistributed in the benchmark package. The package instead retains versioned subsets or derived products, the exact land mask, construction scripts, intermediate masks and labels, final catalogs, and task tables. Provider citations and source version strings are preserved.

B.5 Catalog Summary

Table 7 reports annual and monthly counts using each event's fixed target time. Thermal events occur in every calendar month because the catalog spans both hemispheres. The largest monthly

Table 7: Temporal distribution of retained events by target time.

Year	Heat	Cold	Precipitation	TC
2022	52	38	70	32
2023	82	29	60	31
2024	106	20	68	33
2025	74	19	77	29

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Heat	17	24	22	21	24	30	30	39	38	29	17	23
Cold	13	13	9	9	14	7	5	7	4	11	4	10
Precipitation	13	16	33	27	19	19	21	31	26	22	29	19
TC	6	7	7	3	6	6	13	15	22	22	11	7

Table 8: Duration and target region area of retained gridded events.

	Duration (days)				Target area (10^6 km^2)		
Category	Min.	Median	95th pct.	Max.	5th pct.	Median	95th pct.
Heat	3	5	15	53	1.06	1.82	4.11
Cold	3	4	8	12	1.07	2.21	4.22
Precipitation	3	3	5	8	0.23	0.45	1.14

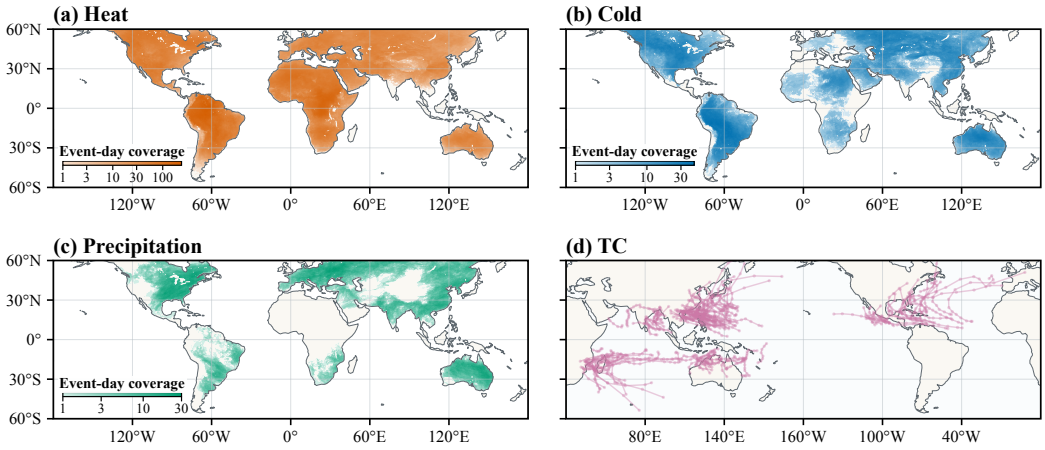

Figure 4: Global coverage of the Extreme Weather Events Dataset. Panels (a)–(c) count retained event days at each grid cell. Panel (d) shows one retained IBTrACS fix per storm day along the TC tracks, for 1,239 points during 2022–2025.

counts are August for heat (39), May for cold (14), March for precipitation (33), and September and October for TC (22 each). Table 8 summarizes the tracked duration and the catalog mask area used at the fixed target for gridded events. Areas describe the task geometry rather than the union of an event over its lifetime. Figure 4 summarizes the geographic coverage of the retained event catalog.

C Benchmark Inventory, Scales, and Aggregation

C.1 Diagnostic and Task Inventory

Table 3 introduces the diagnostics in the main text. Tables 9 and 10 give their units, pressure levels, and precise scored objects. Table 11 reports event, task, and initialization counts.

Table 9: Universal physical structure diagnostics.

Diagnostic	Scored physical structure	Level(s) (hPa)	Units
U1: Normalized signed hydrostatic components	Three sublayer thickness residuals with symmetric normalization	850, 700, 600, 500	1
U2: Geostrophic-wind components	Geopotential-derived eastward and northward geostrophic wind	500	$\mathrm{m\,s^{-1}}$
U3: Geostrophic-wind shear components	Eastward and northward 850-minus-500 hPa geostrophic shear	850, 500	$\mathrm{m\,s^{-1}}$
U4: Relative vorticity	Spherical rotational structure	850	$\mathrm{s^{-1}}$
U5: Horizontal divergence	Spherical divergent structure	850	$\mathrm{s^{-1}}$
U6: Relative vorticity	Midtropospheric rotational structure	500	$\mathrm{s^{-1}}$
U7: Horizontal divergence	Midtropospheric divergent structure	500	$\mathrm{s^{-1}}$
U8: Horizontal deformation components	Signed spherical stretching and shearing components	850	$\mathrm{s^{-1}}$
U9: Dry Brunt–Väisälä frequency squared	Dry-potential-temperature layer stability	850–500	$\mathrm{s^{-2}}$
U10: Moist static energy	Thermodynamic energy state	850	$\mathrm{J\,kg^{-1}}$

Each diagnostic is derived independently from forecast and ERA5 fields. U2 and U3 exclude $|\phi| < 15^\circ$. Horizontal derivatives use spherical operators with a cyclic longitude stencil on a complete global grid. Joint-component errors preserve signed layer or vector components.

Table 10: Event-specific physical structure diagnostics.

Block	Diagnostic	Level(s) (hPa)	Scored physical structure
Heat/cold	E1: Temperature advection	850	Horizontal dry-temperature transport
	E2: Moisture advection	850	Horizontal specific-humidity transport
	E3: Partial-layer horizontal MSE advection	850–500	Gauge-invariant $-\int \mathbf{V}\cdot\nabla h\, dp/g$ contribution
	E4: Dry Brunt–Väisälä frequency squared	850–700	Lower-tropospheric dry stability
	E5: Horizontal kinematic temperature frontogenesis	850	Horizontal change of temperature-gradient magnitude
Precipitation	E1: Free-tropospheric water vapor	850–300	Layer water amount
	E2: Free-tropospheric vapor-transport components	850–300	Eastward and northward vapor transport
	E3: Free-tropospheric vapor-flux convergence	850–300	Layer-integrated vapor convergence
	E4: Total moisture-flux convergence	850	Low-level $-\nabla\cdot(q\mathbf{V})$
	E5: Total MSE layer difference	850–500	Signed lower-minus-midlevel thermodynamic contrast
TC	E1: Axisymmetric tangential wind	850; 0–150 km	Inner-core circulation profile
	E2: Total Ertel potential vorticity	700	Rotation, stratification, and tilting
	E3: Radial relative vorticity	500	Midlevel vortex profile
	E4: Inner-outflow radial divergence	200; 0–500 km	Upper-level outflow profile
	E5: Warm-core thickness anomaly	300–700	Inner-core thickness relative to an ERA5-defined environment

E1–E5 are local codes within each block. MSE denotes moist static energy. Heat and cold share one diagnostic configuration but use distinct event masks. TC radial support and the warm-core environment are frozen from ERA5 and reused for the forecast.

Table 11: Final task inventory.

Event category	Events	Tasks	Leads (h)	Unique initializations
Extreme heat stress	314	2198	24:24:168	1160
Extreme cold stress	106	742	24:24:168	602
Extreme precipitation	275	1925	24:24:168	1093
Landfalling TC	125	1452	6:6:120	1296
Total	**820**	**6317**	–	–

C.2 Fixed Physics Score Scales

Table 12 lists the prescribed regime parameters used to construct C_i; Table 13 gives the resulting characteristic magnitudes, tolerances, and score scales.

Table 12: Fixed regime parameters used to construct characteristic magnitudes C_i.

Regime or operator	Pre-specified characteristic parameters
Synoptic universal	$L = 250$ km, $U = 15$ m s^{-1}, $f_0 = 10^{-4}$ s^{-1}, $\theta_0 = 300$ K, $H = 5$ km, $\Delta\theta = 20$ K, $\Delta T = 10$ K, $\Delta q = 0.005$ kg kg^{-1}, and $\Delta H = 1$ km. These values apply to the universal block for stress and precipitation events.
Thermal stress	Universal synoptic values; the 850–500-hPa MSE-advection scale additionally uses $\Delta p = 350$ hPa and $\Delta h = 3$ kJ kg^{-1}; the 850–700-hPa stability scale uses $H = 1.5$ km and $\Delta\theta = 10$ K.
Precipitation	Universal synoptic values; the fixed free-tropospheric layer uses $\Delta p = 550$ hPa and $\bar{q} = 0.005$ kg kg^{-1}. The 850-hPa flux-convergence scale uses $q_0 = 0.01$ kg kg^{-1} and $L = 500$ km. The MSE layer-difference scale uses $\Delta T = 30$ K, $\Delta q = 0.008$ kg kg^{-1}, and $\bar{T} = 270$ K.
Tropical cyclone	$R = L = 200$ km, $U = 40$ m s^{-1}, $f_0 = 5 \times 10^{-5}$ s^{-1}, $\theta_0 = 300$ K, $H = 5$ km, $\Delta\theta = 20$ K, $\Delta T = 8$ K, $\Delta q = 0.008$ kg kg^{-1}, and $\Delta H = 1.5$ km. The event block uses $U_{850} = 40$, $U_{500} = 30$, and $U_{200} = 20$ m s^{-1}; PV uses $\Delta p = 350$ hPa.

Constants are $g = 9.80665$ m s^{-2}, $R_d = 287.05$ J kg^{-1} K^{-1}, $c_p = 1004.0$ J kg^{-1} K^{-1}, and $L_v = 2.5 \times 10^6$ J kg^{-1}. These are dimensional reference magnitudes, not statistics of the evaluated events or systems.

For every domain–diagnostic row, the benchmark declares a characteristic magnitude C_i and a dimensionless tolerance τ_i, with dimensional score scale

$$S_i = C_i \tau_i. \tag{3}$$

This product sets the diagnostic error corresponding to a score of $100/e$. The complete inventory in CSV format contains 45 rows: 15 for thermal stress, 15 for precipitation, and 15 for TC. Heat and cold share the thermal stress rows. The printed table collapses only identical universal synoptic values; it does not introduce another scoring regime. All C_i and τ_i values are fixed across systems, events, and leads. The tolerances specify the benchmark's reference error magnitude and are neither observational uncertainties nor thresholds of physical correctness.

Figure 5 shows the transformation of each task score. With $r = E_{i,k}/S_i$, the score is 100 at exact agreement, 50 at $r = \sqrt{\ln 2} \simeq 0.833$, $100/e \simeq 36.79$ at $r = 1$, 1.83 at $r = 2$, and 0.012 at $r = 3$.

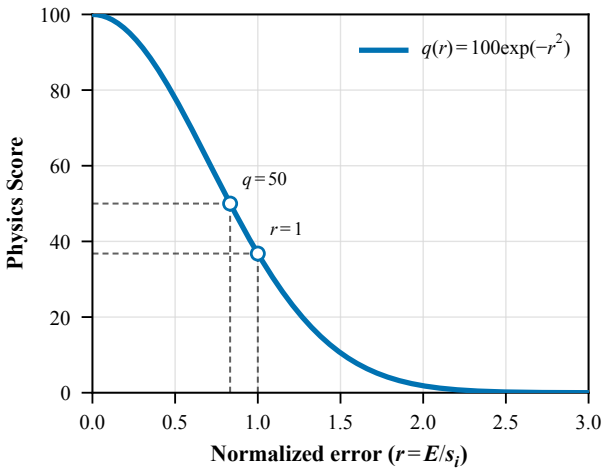

Figure 5: Fixed mapping for each task $q_{i,k} = 100\exp(-r^2)$ for normalized diagnostic RMSE $r = E_{i,k}/(C_i\tau_i)$. The map is prespecified and is applied before every averaging step.

C.3 Explicit Hierarchical Estimator

The estimator controls how much influence each event and physical family has on a summary. Mapping before aggregation is essential because the exponential map is nonlinear: mapping a mean squared error would define a different score. Let $q_{me\ell i}$ be the mapped score for model m, event e, retained task or lead ℓ, and diagnostic i. Let L_{ei} contain the tasks on which diagnostic i is applicable according to the reference. For a complete model record, the event and diagnostic mean is

$$\bar{q}_{mei} = \frac{1}{|L_{ei}|} \sum_{\ell \in L_{ei}} q_{me\ell i}. \tag{4}$$

Table 13: Fixed characteristic magnitude C_i, tolerance fraction τ_i, and score scale $S_i = C_i\tau_i$. Identical universal rows for the stress and precipitation domains are collapsed into one synoptic block; the released `physical_scales.csv` records all 45 domain–diagnostic rows without collapsing.

Regime	Code	Diagnostic	C_i construction	C_i	τ_i	S_i	Units
Synoptic universal	U1	Hydrostatic normalized residual components	1	1	0.001	0.001	1
	U2	Geostrophic-wind components, 500	$U/\sqrt{2}$	10.6066	0.50	5.3033	$\mathrm{m\,s^{-1}}$
	U3	Geostrophic-wind shear components, 850–500	$U/\sqrt{2}$	10.6066	0.50	5.3033	$\mathrm{m\,s^{-1}}$
	U4/U6	Relative vorticity, 850/500	U/L	6.0000×10^{-5}	0.50	3.0000×10^{-5}	$\mathrm{s^{-1}}$
	U5/U7	Divergence, 850/500	$U^2/(f_0L^2)$	3.6000×10^{-5}	1.00	3.6000×10^{-5}	$\mathrm{s^{-1}}$
	U8	Deformation components, 850	$U/(\sqrt{2}L)$	4.2426×10^{-5}	1.00	4.2426×10^{-5}	$\mathrm{s^{-1}}$
	U9	Dry N^2, 850–500	$g\Delta\theta/(\theta_0 H)$	1.3076×10^{-4}	0.10	1.3076×10^{-5}	$\mathrm{s^{-2}}$
	U10	MSE, 850	$c_p\Delta T + L_v\Delta q + g\Delta H$	32346.65	0.15	4851.9975	$\mathrm{J\,kg^{-1}}$
Thermal stress	E1	Temperature advection, 850	$U\Delta T/L$	6.0000×10^{-4}	0.25	1.5000×10^{-4}	$\mathrm{K\,s^{-1}}$
	E2	Moisture advection, 850	$U\Delta q/L$	3.0000×10^{-7}	0.25	7.5000×10^{-8}	$\mathrm{kg\,kg^{-1}\,s^{-1}}$
	E3	Partial-layer horizontal MSE advection, 850–500	$(\Delta p/g)U\Delta h/L$	642.4212	0.50	321.2106	$\mathrm{W\,m^{-2}}$
	E4	Dry N^2, 850–700	$g\Delta\theta/(\theta_0 H)$	2.1793×10^{-4}	0.25	5.4481×10^{-5}	$\mathrm{s^{-2}}$
	E5	Horizontal kinematic temperature frontogenesis, 850	$U\Delta T/L^2$	2.4000×10^{-9}	0.25	6.0000×10^{-10}	$\mathrm{K\,m^{-1}\,s^{-1}}$
Precipitation	E1	Free-tropospheric water vapor, 850–300	$\Delta p\bar{q}/g$	28.0422	0.10	2.8042	$\mathrm{kg\,m^{-2}}$
	E2	Vapor-transport components, 850–300	$\Delta p\bar{q}U/g$	420.6329	0.10	42.0633	$\mathrm{kg\,m^{-1}\,s^{-1}}$
	E3	Vapor-flux convergence, 850–300	$\Delta p\bar{q}U/(gL)$	1.6825×10^{-3}	0.25	4.2063×10^{-4}	$\mathrm{kg\,m^{-2}\,s^{-1}}$
	E4	Moisture-flux convergence, 850	q_0U/L	3.0000×10^{-7}	1.00	3.0000×10^{-7}	$\mathrm{kg\,kg^{-1}\,s^{-1}}$
	E5	MSE layer difference, 850–500	$\lvert c_p\Delta T + L_v\Delta q - R_d\bar{T}\ln(850/500)\rvert$	8994.4533	0.25	2248.6133	$\mathrm{J\,kg^{-1}}$
TC universal	U1	Hydrostatic normalized residual components	1	1	0.001	0.001	1
	U2	Geostrophic-wind components, 500	$U/\sqrt{2}$	28.2843	0.50	14.1421	$\mathrm{m\,s^{-1}}$
	U3	Geostrophic-wind shear components, 850–500	$U/\sqrt{2}$	28.2843	0.50	14.1421	$\mathrm{m\,s^{-1}}$
	U4/U6	Relative vorticity, 850/500	U/L	2.0000×10^{-4}	0.50	1.0000×10^{-4}	$\mathrm{s^{-1}}$
	U5/U7	Divergence, 850/500	U/L	2.0000×10^{-4}	0.20	4.0000×10^{-5}	$\mathrm{s^{-1}}$
	U8	Deformation components, 850	$U/(\sqrt{2}L)$	1.4142×10^{-4}	0.50	7.0711×10^{-5}	$\mathrm{s^{-1}}$
	U9	Dry N^2, 850–500	$g\Delta\theta/(\theta_0 H)$	1.3076×10^{-4}	0.10	1.3076×10^{-5}	$\mathrm{s^{-2}}$
	U10	MSE, 850	$c_p\Delta T + L_v\Delta q + g\Delta H$	42741.975	0.10	4274.1975	$\mathrm{J\,kg^{-1}}$
TC event	E1	Tangential-wind profile, 850, 0–150 km	U_{850}	40.0	0.25	10.0	$\mathrm{m\,s^{-1}}$
	E2	Total Ertel PV, 700	$g(f_0 + U/L)\Delta\theta/\Delta p$	1.4010×10^{-6}	0.25	3.5024×10^{-7}	$\mathrm{K\,m^2\,kg^{-1}\,s^{-1}}$
	E3	Radial relative-vorticity profile, 500	U_{500}/R	1.5000×10^{-4}	0.50	7.5000×10^{-5}	$\mathrm{s^{-1}}$
	E4	Radial divergence profile, 200, 0–500 km	U_{200}/R	1.0000×10^{-4}	0.20	2.0000×10^{-5}	$\mathrm{s^{-1}}$
	E5	Warm-core thickness anomaly, 300–700	$R_d\Delta T\ln(700/300)/g$	198.4097	0.10	19.8410	m

An empty applicability set does not receive a score. Missing or invalid model output on a reference applicable task makes the corresponding record incomplete rather than reducing its denominator. Model comparisons retain the same complete events across all compared systems within each category.

Applicability also determines which diagnostics and physical families contribute within an event. Write I_{ef} for the applicable diagnostics in physical family f, F_{eg} for the nonempty physical families in group $g \in \{U, E\}$, and $\mathcal{E}_c$ for the common complete events in category c. The hierarchy is

$$Q_{mef} = \frac{1}{|I_{ef}|} \sum_{i \in I_{ef}} \bar{q}_{mei}, \tag{5}$$

$$Q_{meg} = \frac{1}{|F_{eg}|} \sum_{f \in F_{eg}} Q_{mef}, \tag{6}$$

$$Q_{me}^{\text{event}} = \tfrac{1}{2}(Q_{meU} + Q_{meE}), \tag{7}$$

$$Q_{mc}^{\text{category}} = \frac{1}{|\mathcal{E}_c|} \sum_{e \in \mathcal{E}_c} Q_{me}^{\text{event}}. \tag{8}$$

Thus each applicable diagnostic has equal weight within its physical family, each nonempty physical family has equal weight within its group, and the two groups have equal weight within an event. Each complete event then has equal weight within its category. Inapplicability, which is determined by the reference, is distinct from invalid forecast output. For example, an equatorial event can omit geostrophic diagnostics for every model without assigning them zero scores.

Overall retains the common universal group and equal category weights:

$$Q_m^{\text{overall}} = \frac{1}{4} \sum_{c \in \{\text{heat,cold,precipitation,TC}\}} \frac{1}{|\mathcal{E}_c|} \sum_{e \in \mathcal{E}_c} Q_{meU}. \tag{9}$$

This avoids averaging physically different event-specific diagnostics across categories. Lead reports retain the lead coordinate and apply their declared event cohort and physical family grouping; they are reported separately from the mean across native leads.

Physical family reports. For the main physical family table, let $\mathcal{E}_{cf}$ be the events in $\mathcal{E}_c$ with nonempty physical family f. The reported mean is

$$Q_{mcf}^{\text{family}} = \frac{1}{|\mathcal{E}_{cf}|} \sum_{e \in \mathcal{E}_{cf}} Q_{mef}. \tag{10}$$

Its event count is $|\mathcal{E}_{cf}|$. A physical family with no supported events is reported as missing. When physical family applicability differs across events, averaging these final means does not reproduce Equation 8; reconstruction must start from the physical family scores within each event. The report retains separate universal and event-specific physical families even where their display names are both thermodynamics.

Weighting rationale. Hierarchical averaging specifies the contribution of each level rather than allowing groups with more observations or more measures to dominate automatically. The OECD/JRC guidance on composite indicators treats normalization, weighting, and aggregation as choices that require justification and sensitivity assessment (OECD et al., 2008). Equal weights here express the benchmark's reporting priorities; they do not imply that the physical families have identical causal importance. Correlated diagnostics and compensation between physical families remain relevant limitations. No new weighting sensitivity experiment is claimed.

Reporting beyond a leaderboard. LiveBench presents category results beside an overall score, while HELM evaluates multiple dimensions under shared scenarios (White et al., 2024; Liang et al., 2023). We use the corresponding reporting principle to keep the event category, physical family, and diagnostic views visible. The overall ranking is an entry point to the evaluation; the disaggregated results identify the particular physical structures responsible for it. This motivates the result layout without importing another benchmark's task definitions, units, or scoring rule.

D Diagnostic Operators and Error Geometry

This appendix specifies the observable physical structures compared by ExtremePhysBench. It does not define residuals of the complete prognostic equations. Terms requiring time tendencies, vertical motion, diabatic heating, friction, surface exchange, or subgrid tendencies are unavailable from the archived instantaneous fields on pressure levels. Every primary operator below is therefore evaluated independently on the forecast and ERA5 state, and the two resulting diagnostic fields are compared on support determined without the forecast. Classical definitions follow standard dynamic meteorology (Holton & Hakim, 2013); pressure levels, discrete geometry, support, and error aggregation are benchmark choices.

D.1 Notation and Spherical Operators

Let longitude, latitude, and Earth radius be (λ, ϕ, a); let (u, v) denote eastward and northward wind, Φ geopotential, T dry-air temperature, q specific humidity, p pressure, $f = 2\Omega \sin\phi$, and

$$T_v = T(1 + 0.608q), \qquad \theta = T(p_0/p)^{R_d/c_p}, \qquad h = c_p T + \Phi + L_v q. \tag{11}$$

For a scalar A, the horizontal gradient in the local orthonormal east–north basis is

$$\nabla_h A = \left(\frac{1}{a\cos\phi} \frac{\partial A}{\partial \lambda}, \frac{1}{a} \frac{\partial A}{\partial \phi} \right). \tag{12}$$

For horizontal wind, the covariant gradient components are

$$u_e = \frac{1}{a\cos\phi} u_\lambda - \frac{v\tan\phi}{a}, \qquad u_n = \frac{1}{a} u_\phi, \tag{13}$$

$$v_e = \frac{1}{a\cos\phi} v_\lambda + \frac{u\tan\phi}{a}, \qquad v_n = \frac{1}{a} v_\phi. \tag{14}$$

Hence divergence, vertical relative vorticity, stretching deformation, and shearing deformation are

$$\delta = u_e + v_n, \qquad \zeta = v_e - u_n, \qquad D_s = u_e - v_n, \qquad D_h = v_e + u_n. \tag{15}$$

Derivatives use centered finite differences where possible. Longitude is cyclic only when the input is a complete uniform global grid; the geographic poles are excluded from the singular eastward basis. The spherical connection terms are retained throughout the diagnostic suite.

D.2 From Governing Relations to Observable Physical Structure

The governing equations organize the diagnostic suite, but they are not themselves the scored object. For atmospheric state X and diagnostic operator $\mathcal{D}_i$, forecast and ERA5 are processed independently,

$$D_i^{\mathrm{F}} = \mathcal{D}_i(X^{\mathrm{F}}), \qquad D_i^{\mathrm{E}} = \mathcal{D}_i(X^{\mathrm{E}}), \qquad \Delta D_i = D_i^{\mathrm{F}} - D_i^{\mathrm{E}}. \tag{16}$$

The derivations below therefore end at an observable field D_i whose forecast–reference departure can be evaluated. Exact identities and coordinate transformations are written with equality signs. Geostrophic balance, source-free horizontal scalar evolution, and axisymmetric gradient-wind balance are controlled approximations; they motivate diagnostic fields or audit terms but do not assert closure of a forecast snapshot.

D.2.1 Hydrostatic Structure, Thickness, and Stability

Let ρ be density and z geometric height. The moist-air equation of state and hydrostatic balance are

$$p = \rho R_d T_v, \qquad \frac{\partial p}{\partial z} = -\rho g, \qquad \mathrm{d}\Phi = g\,\mathrm{d}z. \tag{17}$$

Eliminating density gives the pressure-coordinate differential identity

$$\mathrm{d}\Phi = -R_d T_v \frac{\mathrm{d}p}{p} = -R_d T_v\,\mathrm{d}\ln p. \tag{18}$$

For lower pressure surface p_l and upper surface $p_u < p_l$, integration yields the hypsometric relation

$$\Phi_u - \Phi_l = R_d \int_{p_u}^{p_l} T_v \, \mathrm{d} \ln p. \tag{19}$$

Endpoint trapezoidal quadrature of Equation 19 produces the two independently represented layer-thickness terms A_{lu} and B_{lu} in U1. Their normalized signed difference is retained because the benchmark compares the forecast residual with the ERA5 residual; it does not require either residual to vanish.

Potential temperature follows from dry adiabatic compression or expansion,

$$\theta = T \left(\frac{p_0}{p} \right)^{\kappa}, \qquad \kappa = \frac{R_d}{c_p}. \tag{20}$$

The dry squared Brunt–Väisälä frequency is the restoring-acceleration coefficient for a small vertical parcel displacement,

$$N^2 = \frac{g}{\theta} \frac{\partial \theta}{\partial z}. \tag{21}$$

Replacing the derivative by a finite layer difference and using $\Delta z = (\Phi_u - \Phi_l)/g$ gives Equation 47. Thus U9 and thermal-stress E4 are finite-layer dry-stability structures, not moist parcel-instability or convective available potential energy diagnostics.

Equation 19 also explains the TC warm-core operator. The thickness

$$Z_{300:700} = \frac{\Phi_{300} - \Phi_{700}}{g} = \frac{R_d}{g} \int_{300}^{700} T_v \, \mathrm{d} \ln p \tag{22}$$

increases with layer-mean virtual temperature. Subtracting the ERA5-defined outer environment therefore isolates a warm-core thickness anomaly while removing the regional background.

D.2.2 Rotating Momentum, Geostrophic Wind, and Thermal Wind

On a pressure surface, horizontal momentum may be written

$$\frac{D\boldsymbol{V}}{Dt} + f\boldsymbol{k} \times \boldsymbol{V} = -\nabla_h \Phi + \boldsymbol{F}, \tag{23}$$

where $\boldsymbol{k}$ is the local vertical unit vector and $\boldsymbol{F}$ collects frictional and unresolved momentum tendencies. Away from the equator, neglecting acceleration and $\boldsymbol{F}$ relative to the pressure-gradient and Coriolis terms gives the geostrophic approximation

$$f\boldsymbol{k} \times \boldsymbol{V}_g \approx -\nabla_h \Phi, \qquad (u_g, v_g) = \left(-\frac{\Phi_y}{f}, \frac{\Phi_x}{f} \right). \tag{24}$$

The current U2 operator scores these two geopotential-derived components at 500 hPa. U3 scores their 850-minus-500-hPa endpoint difference, so neither U2 nor U3 mixes forecast wind with reference geopotential. The approximation is used to define an observable balanced-flow structure rather than to label ageostrophic motion as forecast error (Charney, 1948; Holton & Hakim, 2013).

Differentiating Equation 24 with respect to $\ln p$ and substituting Equation 18 gives

$$\frac{\partial u_g}{\partial \ln p} = \frac{R_d}{f} \frac{\partial T_v}{\partial y}, \qquad \frac{\partial v_g}{\partial \ln p} = -\frac{R_d}{f} \frac{\partial T_v}{\partial x}. \tag{25}$$

Integrating from 500 to 850 hPa yields

$$\boldsymbol{V}_{g,850} - \boldsymbol{V}_{g,500} = \frac{R_d}{f} (\partial_y, -\partial_x) \int_{500}^{850} T_v \, \mathrm{d} \ln p. \tag{26}$$

The implementation evaluates the log-pressure integral by trapezoidal quadrature over 500, 600, 700, and 850 hPa. Its difference from the geopotential-derived endpoint shear is exported as a signed audit residual. It is not the primary U3 score, and both sides retain the $|\phi| \geq 15^\circ$ applicability restriction.

D.2.3 Horizontal Budgets, Layer Mass, and Moist Transport

Let $\omega = Dp/Dt$ be pressure vertical velocity, $\dot{Q}$ diabatic heating per unit mass, and S_q the net moisture source. The temperature and moisture equations in pressure coordinates contain

$$\frac{\partial T}{\partial t} + \boldsymbol{V} \cdot \nabla_h T + \omega \frac{\partial T}{\partial p} - \frac{\kappa T}{p}\omega = \frac{\dot{Q}}{c_p}, \tag{27}$$

$$\frac{\partial q}{\partial t} + \boldsymbol{V} \cdot \nabla_h q + \omega \frac{\partial q}{\partial p} = S_q. \tag{28}$$

Instantaneous archived fields do not identify every term in these equations. The benchmark therefore extracts only the observable horizontal contributions $-\boldsymbol{V} \cdot \nabla_h T$ and $-\boldsymbol{V} \cdot \nabla_h q$ for thermal-stress E1 and E2. These operators diagnose resolved redistribution at one time; they are not estimates of the complete temperature or moisture tendency.

Hydrostatic balance also supplies the exact layer-mass conversion

$$\rho \, \mathrm{d}z = -\frac{\mathrm{d}p}{g}. \tag{29}$$

Consequently, for an upper bound p_u and lower bound $p_l > p_u$,

$$\int_{z_l}^{z_u} \rho A \, \mathrm{d}z = \frac{1}{g} \int_{p_u}^{p_l} A \, \mathrm{d}p. \tag{30}$$

Choosing $A = q$ gives partial-layer water vapor; choosing $A = q\boldsymbol{V}$ gives partial-layer horizontal vapor transport. Applying negative horizontal divergence gives

$$-\nabla_h \cdot \left(\frac{1}{g} \int_{300}^{850} q\boldsymbol{V} \, \mathrm{d}p \right), \tag{31}$$

which is precipitation E3. The fixed 850–300-hPa limits make these free-tropospheric rather than total-column quantities.

For the instantaneous moist static energy $h = c_p T + \Phi + L_v q$, horizontal differentiation gives the exact decomposition

$$\boldsymbol{V} \cdot \nabla_h h = c_p \boldsymbol{V} \cdot \nabla_h T + \boldsymbol{V} \cdot \nabla_h \Phi + L_v \boldsymbol{V} \cdot \nabla_h q. \tag{32}$$

Pressure integration of its negative over 500–850 hPa gives thermal-stress E3. An additive constant in h disappears under ∇_h, making this advective form gauge invariant. The conservative identity

$$-\nabla_h \cdot (h\boldsymbol{V}) = -\boldsymbol{V} \cdot \nabla_h h - h \nabla_h \cdot \boldsymbol{V} \tag{33}$$

provides audit terms but not a closed energy budget. Likewise, precipitation E4 follows from the moisture-flux product rule

$$-\nabla_h \cdot (q\boldsymbol{V}) = -\boldsymbol{V} \cdot \nabla_h q - q \nabla_h \cdot \boldsymbol{V}. \tag{34}$$

D.2.4 Kinematic Frontogenesis

Let $\boldsymbol{g}_T = \nabla_h T$. Under the explicitly restricted, source-free horizontal evolution $\partial_t T + \boldsymbol{V} \cdot \nabla_h T = 0$, covariant differentiation gives

$$\frac{D_h \boldsymbol{g}_T}{Dt} = -(\nabla_h \boldsymbol{V})^{\mathsf{T}} \boldsymbol{g}_T. \tag{35}$$

Projecting this tendency onto $\boldsymbol{g}_T/|\boldsymbol{g}_T|$ yields

$$\frac{D_h |\boldsymbol{g}_T|}{Dt} = -\frac{\boldsymbol{g}_T^{\mathsf{T}} (\nabla_h \boldsymbol{V}) \boldsymbol{g}_T}{|\boldsymbol{g}_T|}, \tag{36}$$

which is the thermal-stress E5 field (Petterssen, 1936). The derivation explains its kinematic meaning while making clear that vertical motion, tilting, diabatic heating, and turbulent mixing are omitted.

D.2.5 ERTEL POTENTIAL VORTICITY AND STORM-RELATIVE BALANCE

Ertel potential vorticity begins with the coordinate-independent scalar

$$P = \rho^{-1}\boldsymbol{\omega}_a \cdot \nabla\theta. \tag{37}$$

Using pressure as the vertical coordinate, hydrostatic balance converts this identity to

$$P = -g\left[(f+\zeta)\theta_p - v_p\theta_x + u_p\theta_y\right]. \tag{38}$$

The first term couples absolute vertical rotation to static stability; the remaining two are tilting contributions. TC E2 evaluates all terms at 700 hPa, with p derivatives obtained from the Lagrange polynomial through 850, 700, and 500 hPa. This is an instantaneous PV-structure diagnostic, not a test of material PV conservation, which would additionally require adiabatic and frictionless evolution (Ertel, 1942; Hoskins et al., 1985).

For TC radial structure, let $\boldsymbol{e}_r$ be the local outward great-circle direction and $\boldsymbol{e}_\theta$ its counterclockwise tangent. Then

$$V_\theta = \boldsymbol{V} \cdot \boldsymbol{e}_\theta \tag{39}$$

defines the primitive quantity whose complete-ring mean is scored by E1. Under the additional audit assumptions of axisymmetry, quasi-steady tangential flow, and negligible radial acceleration and friction, the radial momentum equation reduces to

$$\frac{\partial\Phi}{\partial r} \approx \frac{V_\theta^2}{r} + fV_\theta, \qquad R_{\mathrm{GW}} = \frac{\partial\Phi}{\partial r} - \frac{V_\theta^2}{r} - fV_\theta. \tag{40}$$

The signed and normalized R_{GW} profiles are audit outputs rather than the E1 score (Willoughby, 1990). TC E3 and E4 instead apply the exact spherical covariant vorticity and divergence definitions in Equation 15 before complete-ring averaging. Together with the hypsometric warm-core construction in Equation 22, these chains connect the TC block to rotation, outflow, and thermal structure without claiming a closed cyclone model.

D.3 REFERENCE-ONLY SUPPORT

For metric i, let $A_e(x)$ be the catalog event mask, $p_s^{\mathrm{E}}(x)$ ERA5 surface pressure, $p_i^{\max}$ the deepest required pressure level, and H_i the operator-stencil validity mask. Before reference-diagnostic finiteness is checked, the requested support is

$$\widetilde{\Omega}_{ie} = A_e \cap \{p_s^{\mathrm{E}} \geq p_i^{\max}\} \cap H_i. \tag{41}$$

The pressure inequality has no added terrain margin. U2 and U3 additionally require $|\phi| \geq 15^\circ$. The final support is

$$\Omega_{ie} = \widetilde{\Omega}_{ie} \cap \{\text{all primary ERA5 diagnostic components are finite}\}. \tag{42}$$

There is no arbitrary minimum event-area or valid-fraction cutoff. A joint-component sample is retained only if all reference components are finite. Forecast values do not alter Ω_{ie}; a nonfinite forecast on this support is recorded as invalid model output. Area fields use $\cos\phi$ weights on the common 0.25° grid.

In the released TC evaluation, all 2,102 rows carrying `model_output_invalid` belong to U2 or U3: they are the two geostrophic diagnostics for 1,051 model–task pairs. They do not denote absent forecast tasks or missing leads. In every case, the independently diagnosed forecast center lies closer to the equator than the ERA5 center, so the forecast-side $|\phi| < 15^\circ$ geostrophic mask enters part of the ERA5-frozen storm-relative support. The status therefore records a nonfinite derived diagnostic on frozen support; the other diagnostics for the same model–task pair remain separately evaluable. Events affected by this condition are not averaged over a model-selected subset in complete paired comparisons.

D.4 UNIVERSAL DIAGNOSTICS U1–U10

The same ten universal diagnostics are used for all four event categories.

U1: normalized signed hydrostatic sublayer residuals. For each adjacent pair (p_l, p_u) in $(850, 700)$, $(700, 600)$, and $(600, 500)$ hPa,

$$A_{lu} = \Phi_u - \Phi_l, \tag{43}$$

$$B_{lu} = \frac{R_d}{2}(T_{v,l} + T_{v,u}) \ln(p_l/p_u), \tag{44}$$

$$R^H_{lu} = \frac{A_{lu} - B_{lu}}{\frac{1}{2}(|A_{lu}| + |B_{lu}|)}. \tag{45}$$

The three signed R^H_{lu} fields are scored jointly with pressure-thickness component weights. The comparison is forecast residual versus ERA5 residual, not residual versus zero.

U2–U3: geopotential-derived balanced-flow structure. At pressure p,

$$(u_g, v_g)_p = \left(-\frac{\Phi_y}{f}, \frac{\Phi_x}{f}\right)_p. \tag{46}$$

U2 jointly scores the eastward and northward components at 500 hPa. U3 jointly scores the endpoint shear $(u_{g,850} - u_{g,500}, v_{g,850} - v_{g,500})$. A log-pressure virtual-temperature thermal-wind residual is exported for audit, but it is not the primary U3 field. The 15° latitude exclusion avoids division by small f.

U4–U8: spherical kinematics. U4 and U6 are ζ at 850 and 500 hPa; U5 and U7 are δ at those levels. U8 jointly scores the signed (D_s, D_h) components at 850 hPa. The unsigned deformation magnitude is retained only as an audit summary.

U9: dry layer stability. For lower and upper levels $(l, u) = (850, 500)$ hPa,

$$N^2_{lu} = g\frac{\theta_u - \theta_l}{\bar{\theta}_{lu}\,\Delta z_{lu}}, \qquad \bar{\theta}_{lu} = \tfrac{1}{2}(\theta_l + \theta_u), \quad \Delta z_{lu} = \frac{\Phi_u - \Phi_l}{g}. \tag{47}$$

Samples with nonpositive $\bar{\theta}_{lu}$ or Δz_{lu} are not diagnostically valid in ERA5 and therefore do not enter reference support.

U10: moist static energy. U10 is h_{850} from Equation 11. It compares the observable thermodynamic state and is not an MSE budget-closure test (Neelin & Held, 1987).

D.5 Thermal-Stress Diagnostics E1–E5

Heat and cold events share the same thermal stress operator block.

E1–E2: horizontal scalar advection. At 850 hPa,

$$A_T = -\boldsymbol{V} \cdot \nabla_h T, \qquad A_q = -\boldsymbol{V} \cdot \nabla_h q. \tag{48}$$

E3: partial-layer horizontal MSE advection. The primary field is the gauge-invariant advective contribution

$$A_h = -\frac{1}{g}\int_{500\,\mathrm{hPa}}^{850\,\mathrm{hPa}} \boldsymbol{V} \cdot \nabla_h h \, dp, \tag{49}$$

evaluated by trapezoidal pressure integration over 500, 600, 700, and 850 hPa. The conservative $-\nabla_h \cdot (h\boldsymbol{V})$ form, its $-h\nabla_h \cdot \boldsymbol{V}$ term, and their discrete product-rule residual are retained for audit only. E3 excludes vertical transport, storage, sources, sinks, and the atmosphere outside 850–500 hPa; it is not a column budget.

E4: lower-tropospheric dry stability. E4 uses Equation 47 for 850–700 hPa.

E5: horizontal kinematic temperature frontogenesis. With $\boldsymbol{g}_T = \nabla_h T$ and the covariant wind-gradient tensor $\nabla_h \boldsymbol{V}$,

$$F_T = -\frac{\boldsymbol{g}_T^{\mathsf{T}}(\nabla_h \boldsymbol{V})\boldsymbol{g}_T}{|\boldsymbol{g}_T|}, \tag{50}$$

with zero returned where $|\boldsymbol{g}_T| = 0$. This is the two-dimensional kinematic contribution to the material change of temperature-gradient magnitude (Petterssen, 1936); vertical-motion, tilting, and diabatic terms are outside its scope.

D.6 Precipitation Diagnostics E1–E5

The free-tropospheric operators use the fixed levels 850, 700, 600, 500, 400, and 300 hPa. Defining $\mathcal{I}[A] = -g^{-1}\int_{850}^{300} A\, dp$,

$$W_{850:300} = \mathcal{I}[q], \tag{51}$$
$$\boldsymbol{Q}_{850:300} = \mathcal{I}[q\boldsymbol{V}], \tag{52}$$
$$C_{850:300} = -\nabla_h \cdot \boldsymbol{Q}_{850:300}. \tag{53}$$

E1 scores W, E2 jointly scores the signed eastward and northward components of $\boldsymbol{Q}$, and E3 scores C. The fixed 850–300-hPa bounds mean that these are neither total-column water nor total-column transport quantities (Benton & Estoque, 1954).

E4 is the conservative flux-form 850-hPa moisture-flux convergence

$$C_q = -\nabla_h \cdot (q\boldsymbol{V}). \tag{54}$$

The advective and divergent product-rule components and their discrete residual are exported for audit (Banacos & Schultz, 2005). E5 is the signed layer difference

$$\Delta h_{850:500} = h_{850} - h_{500} = c_p(T_{850} - T_{500}) + (\Phi_{850} - \Phi_{500}) + L_v(q_{850} - q_{500}). \tag{55}$$

D.7 Tropical-Cyclone Diagnostics E1–E5

At every task valid time, the IBTrACS position seeds separate searches for the ERA5 and forecast minimum mean-sea-level pressure within $\pm 8°$. Each state is then sampled relative to its own center. The spherical azimuthal-equidistant patch extends to 500 km at 25-km spacing. Exact-radius great-circle rings contain 360 azimuths and a radial value is valid only when the complete ring is available. This construction isolates storm-relative organization; it deliberately does not score center displacement. Figure 6 summarizes the sampling and scoring geometry.

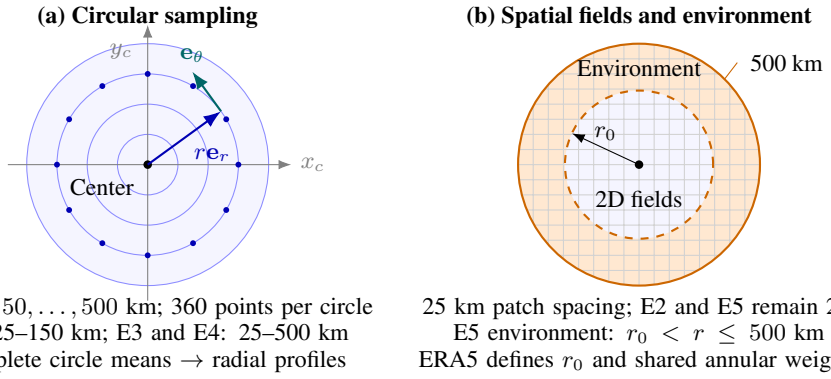

Figure 6: Storm-relative evaluation geometry about independently diagnosed centers. The E5 environmental boundary r_0 and annular weights are determined from ERA5 and reused for the forecast. Circles and grid spacing are schematic.

E1: inner-core tangential wind. At each ring, wind is projected onto the local great-circle tangential direction and azimuthally averaged. E1 scores the complete-ring profile $\overline{V_\theta}(r)$ for $0 < r \leq 150$ km at 850 hPa. The associated gradient-wind terms are audit outputs, not the primary score.

E2: total Ertel potential vorticity. At 700 hPa,

$$P = -g\left[(f+\zeta)\theta_p - v_p\theta_x + u_p\theta_y\right]. \tag{56}$$

The vertical derivatives use three-level Lagrange differentiation at 850, 700, and 500 hPa. The stretching term and both tilting terms are exported separately, while their total is scored (Ertel, 1942; Hoskins et al., 1985).

E3–E4: radial rotation and outflow structure. E3 is the complete-ring azimuthal-mean profile of full covariant ζ_{500} over $0 < r \leq 500$ km. E4 is the corresponding δ_{200} profile over the same inner-outflow radius. Their profile MSE gives equal weight to each retained radius bin.

E5: warm-core thickness anomaly. Let $Z_{300:700} = (\Phi_{300} - \Phi_{700})/g$. ERA5 first defines a storm outer radius as the first zero crossing after the positive maximum of the hemisphere-signed, azimuthal-mean ζ_{500} profile. The environmental mean is the spherical-annulus-weighted mean of complete $Z_{300:700}$ rings outside that radius and within 500 km. That ERA5-derived annulus and its weights are frozen and reused for the forecast. E5 scores

$$Z'_{300:700}(x) = Z_{300:700}(x) - \langle Z_{300:700}\rangle_{\mathrm{env}}. \tag{57}$$

If ERA5 has no valid outer zero crossing or complete environment, the diagnostic is inapplicable for every model on that task rather than receiving a substituted climatological radius.

D.8 Diagnostic Error Geometries

A diagnostic field must be reduced to one task error before the score map can be applied. For a scalar field, grid cell area weights w_x keep a regular longitude and latitude grid from overrepresenting small cells at higher latitudes. On the frozen diagnostic support $\Omega_{i,k}$, the mean squared error is

$$M_{i,k} = \frac{\sum_{x\in\Omega_{i,k}} w_x \left(D^{\mathrm{F}}_{i,k}(x) - D^{\mathrm{E}}_{i,k}(x)\right)^2}{\sum_{x\in\Omega_{i,k}} w_x}, \qquad E_{i,k} = \sqrt{M_{i,k}}. \tag{58}$$

Squaring follows differencing, so signs and directions are retained in the comparison rather than discarded by comparing component magnitudes. To give multiple components the declared relative influence, joint component diagnostics use

$$M_i = \frac{\sum_j \gamma_j \sum_{x\in\Omega_i} w_x (D^{\mathrm{F}}_{ij}(x) - D^{\mathrm{E}}_{ij}(x))^2}{\sum_j \gamma_j \sum_{x\in\Omega_i} w_x}, \tag{59}$$

where j indexes components and γ_j is equal for vector/deformation components and proportional to sublayer pressure thickness for U1. Storm-patch fields use spherical patch-cell areas, while E1, E3, and E4 use equal radius-bin weights after complete-ring averaging. Exact spherical-annulus weights are used for the E5 environmental mean, not for radial-profile MSE. These choices preserve signed direction, component, layer, and radial structure until after forecast–ERA5 differencing. The resulting RMSE is normalized and mapped as specified in Section 4.3, then aggregated according to Appendix C.3.

E Forecast Systems and Inference Details

E.1 Earth2Studio Execution Layer

All six DWP systems were run through NVIDIA Earth2Studio 0.14.0 (Earth2Studio Contributors, 2024), using the released model weights and implementations. DWP initial conditions came from ARCO ERA5, while operational Global Forecast System (GFS) forecasts were retrieved from the National Oceanic and Atmospheric Administration's National Centers for Environmental Prediction (NOAA/NCEP). Forecast outputs were standardized to common coordinates and selected at the prescribed lead times, while GFS retained its operational initial conditions.

Tables 14 and 15 record the components and default artifact identifiers selected by the benchmark code together with the complete variable interfaces resolved from the versioned Earth2Studio documentation and implementation (NVIDIA, 2026). Architecture and original model descriptions

Table 14: Forecast systems and checkpoint data provenance used in ExtremePhysBench.

System	Architecture	Earth2Studio component	Checkpoint	Training data
GFS	NWP	`GFS_FX`	Operational forecast archive	N/A
Pangu	3D transformer	`Pangu6`	Original-publication ONNX checkpoints: `pangu_weather_6.onnx` and `pangu_weather_24.onnx`	*Train:* ERA5, 1979–2017 *Val.:* 2019 *FT.:* none reported
GraphCast	graph network	`GraphCast Operational`	Operational checkpoint at 0.25° with 13 pressure levels; ERA5 pretraining and HRES fine tuning	*Train:* ERA5, 1979–2017 *Val.:* 2016–2017 *FT.:* HRES, 2016–2021
FengWu	multimodal transformer	`FengWu`	ERA5-evaluation ONNX checkpoint without operational-analysis transfer learning: `fengwu_v1.onnx`; 69 variables	*Train:* ERA5, 1979–2015 *Val.:* 2016–2017 *FT.:* ERA5, 1979–2015
SFNO	spherical neural operator	`SFNO`	Modulus-Makani checkpoint: `sfno_73ch_small`, package version 0.1.0	*Train:* ERA5, 1979–2015 *Val.:* 2016–2017 *FT.:* two-step autoregressive training; no separate date split reported
Aurora	foundation transformer	`Aurora`	Original-publication checkpoint; default loader selects `aurora-0.25-pretrained.ckpt` without LoRA	*Pretrain:* mixture of ten sources, 1950–2020; coverage varies by source *Val.:* IFS HRES 0.25°, 2020 *FT.:* none; pretrained checkpoint without LoRA
FCN3	geometric neural operator	`FCN3`	NVIDIA FourCastNet 3 v1 checkpoint	*Train:* ERA5, 1980–2016 *Val.:* 2018–2021; reported metrics use 2020 *FT.:* ERA5, 2012–2016

Checkpoint descriptions follow the Earth2Studio 0.14.0 API documentation (NVIDIA, 2026); the selected checkpoints correspond to the default models in that release. Training dates and protocols follow the sources discussed below. GraphCast validation dates refer to the development split. Paper training protocols are not independent verification of the released weights; source differences are detailed below. N/A: not applicable; Val.: validation; FT.: fine tuning. GFS uses an operational forecast archive rather than a checkpoint.

follow the corresponding publications (Bi et al., 2023; Bodnar et al., 2025; Bonev et al., 2023; Chen et al., 2025; Bonev et al., 2025; Lam et al., 2023), whereas supplementary training data provenance follows the released Spherical Fourier Neural Operator (SFNO) and FourCastNet 3 (FCN3) model cards and the GraphCast repository documentation (NVIDIA, 2024; 2025; Google DeepMind, 2023). We distinguish the training protocols reported in papers from metadata specific to the selected checkpoints; differences between these sources are recorded below. Table 15 reports the full state variable coordinates exposed by each pinned wrapper and separately identifies internally supplied static fields and derived forcings where applicable. Its principal comparison quantity is the common output pool $\mathcal{C} = \bigcap_s \mathcal{O}_s = \mathcal{A}_5 \cup \{\texttt{msl}, \texttt{u10m}, \texttt{v10m}, \texttt{t2m}\}$, which contains 69 variables across all seven systems. Every raw forecast field used by a diagnostic was selected from $\mathcal{C}$; the benchmark did not score every field in $\mathcal{C}$, and no system received credit for an output unavailable from another system. Every DWP model was advanced autoregressively in 6-h increments and supplied global fields on the common nominal $0.25°$ grid used by the diagnostics.

Training dates and model variants. Pangu's paper reports ERA5 training in 1979–2017 and validation in 2019. GraphCast's paper uses 2016–2017 for development validation, while the selected Operational checkpoint is documented as pretrained on ERA5 in 1979–2017 and fine-tuned on HRES in 2016–2021; the development validation split is not a separately documented validation split for that checkpoint. SFNO's paper (Section 5.2) reports training in 1979–2015, validation in 2016–2017, and two-step autoregressive fine-tuning; its NGC package does not independently specify these dates. FengWu's Supplementary Notes 2 and 4 report ERA5 training and fine-tuning in 1979–2015 and validation in 2016–2017; the main text describes the broader ERA5 period as 1979–2017. The original repository distinguishes `fengwu_v1.onnx`, intended for ERA5 evaluation without operational-analysis transfer learning, from `fengwu_v2.onnx`, which was fine-tuned on operational analysis data through 2021 (OpenEarthLab, n.d.). The benchmark inference scripts call the default FengWu loader, which selects v1; the local cached ONNX file also matches the SHA-256 of the v1 artifact in the Earth2Studio package source. The 2021 cutoff therefore describes v2 rather than the selected checkpoint. Aurora's Supplementary Table C3 and Section C.4 give the source-dependent 1950–2020 pretraining coverage and HRES 2020 validation; the selected pretrained checkpoint does not use the subsequent task-specific fine-tuning (Bi et al., 2023; Lam et al.,

Table 15: Earth2Studio model interfaces and the common output pool from which diagnostic inputs were selected.

System	Complete input interface	Complete output interface	Context / forecast setting
GFS	N/A: operational source; model inputs are not exposed	246 requestable fields: $\{u, v, w, z, t, r, q\} \times \mathcal{L}_{\mathrm{GFS}}$ and $\mathcal{S}_{\mathrm{GFS}}$	Operational cycle; deterministic archived forecast
Pangu	$\mathcal{A}_5 \cup \{\texttt{msl}, \texttt{u10m}, \texttt{v10m}, \texttt{t2m}\}$ (69)	Same 69 state variables	t_0; 6-h deterministic rollout
GraphCast	$\mathcal{A}_6 \cup \{\texttt{t2m}, \texttt{msl}, \texttt{u10m}, \texttt{v10m}\}$ (82); derived $\mathcal{F}$; static `land_sea_mask` and `geopotential_at_surface`	Same 82 state variables plus `tp06` (83)	$t_0 - 6$ h, t_0; 6-h deterministic rollout
FengWu	$\mathcal{A}_5 \cup \{\texttt{u10m}, \texttt{v10m}, \texttt{t2m}, \texttt{msl}\}$ (69)	Same 69 state variables	$t_0 - 6$ h, t_0; 6-h deterministic rollout
SFNO	$\mathcal{A}_5$ plus `u10m`, `v10m`, `u100m`, `v100m`, `t2m`, `sp`, `msl`, and `tcwv` (73)	Same 73 state variables	t_0; 6-h deterministic rollout
Aurora	Same 69 state variables as Pangu; packaged static `z`, `slt`, and `lsm`	Same 69 state variables	$t_0 - 6$ h, t_0; 6-h deterministic rollout
FCN3	$\mathcal{A}_5$ plus `u10m`, `v10m`, `u100m`, `v100m`, `t2m`, `msl`, and `tcwv` (72)	Same 72 state variables	t_0; one realization, seed 333
Common intersection	N/A	$\mathcal{C} = \mathcal{A}_5 \cup \{\texttt{msl}, \texttt{u10m}, \texttt{v10m}, \texttt{t2m}\}$ (69)	Admissible pool of evaluation variables; outputs specific to individual systems excluded

Variable codes follow Earth2Studio 0.14.0. The set $\mathcal{L}_{13}$ of 13 levels contains 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa; $\mathcal{A}_5 = \{z, q, t, u, v\} \times \mathcal{L}_{13}$ and $\mathcal{A}_6 = \{z, q, t, u, v, w\} \times \mathcal{L}_{13}$. The GraphCast forcing set $\mathcal{F}$ contains `toa_incident_solar_radiation`, `year_progress_sin`, `year_progress_cos`, `day_progress_sin`, and `day_progress_cos`. The set $\mathcal{L}_{\mathrm{GFS}}$ of GFS pressure levels contains 1, 2, 3, 5, 7, 10, 15, 20, 30, 40, 50, 70, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 925, 950, 975, and 1000 hPa. Its set $\mathcal{S}_{\mathrm{GFS}}$ of surface and single level variables contains `u10m`, `v10m`, `u100m`, `v100m`, `t2m`, `d2m`, `r2m`, `q2m`, `sp`, `msl`, `tcwv`, `tp`, `2d`, `fg10m`, and `refc`; `2d` is a second Earth2Studio alias for 2-m dew point temperature, and some GFS fields are available only beyond lead zero. Here u and v are horizontal wind, w is pressure vertical velocity, z is geopotential, t is temperature, q is specific humidity, and r is relative humidity; Aurora's static `z`, `slt`, and `lsm` denote surface geopotential, soil type, and land–sea mask. Parenthetical counts include state variables only.

2023; Google DeepMind, 2023; Bonev et al., 2023; NVIDIA, 2024; Chen et al., 2025; Bodnar et al., 2025).

FCN3 source differences. Table 14 reports the FCN3 paper's training split from its Table 3 and Appendix E.4 (1980–2016), fine-tuning split from its Appendix E.3 (2012–2016), and out-of-sample validation period (2018–2021, with reported metrics in 2020). The same paper's Appendix E.2 instead starts pretraining in 1979, an internal difference in that source. The pinned artifact model card lists training in 1980–2015, testing in 2016–2017, and evaluation in 2018–2019; its configuration contains a run-directory label indicating fine-tuning in 2013–2016. These records differ, so the paper dates describe its reported experiments and do not establish the exact training history of the selected weights (Bonev et al., 2025; NVIDIA, 2025).

Table 14 identifies the evaluated checkpoints and reports the paper training protocols with explicit source labels. For FCN3, the selected v1 checkpoint metadata and the paper report different splits, as detailed above.

E.2 Initial Conditions and Forecast Outputs

For the DWP models, the local ARCO adapter served 6-hourly ERA5 fields in the variable vocabulary and coordinate order requested by the corresponding Earth2Studio wrapper. Models with one initial state consumed the analysis at t_0; Aurora, FengWu, and GraphCast additionally consumed $t_0 - 6$ h. Static fields required by Aurora and GraphCast were loaded from their released packages. Although the benchmark stored only variables needed by the downstream diagnostics, the inference call always received the complete state required by the checkpoint. Every scored lead came from the corresponding model integration output.

Earth2Studio's default model packages may be hosted by NVIDIA NGC, Hugging Face, Google Cloud, or the original model provider; Earth2Studio supplies the uniform software interface but does not change ownership or licensing of these artifacts from third parties. Pangu and FengWu ex-

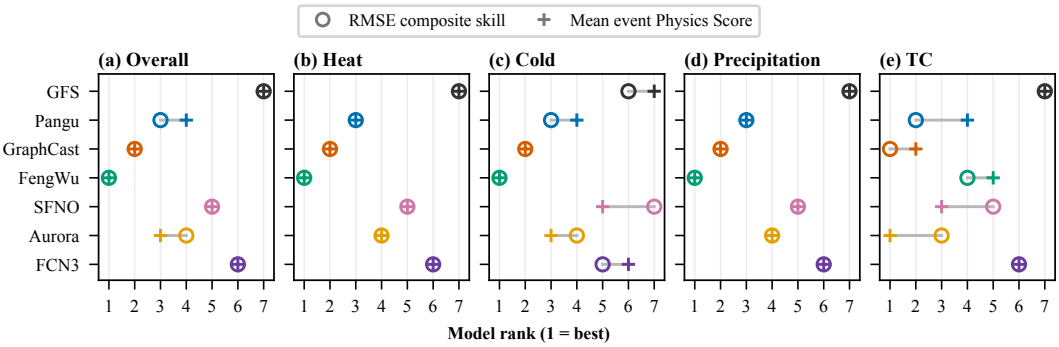

Figure 7: Model ranks on matched event support from conventional RMSE composite skill and mean Event Physics Score. Both summaries use the same events for every model; Overall weights the four event categories equally. This plotted mean of Event Physics Scores is a comparison diagnostic, not a redefinition of the formal Overall score based only on universal diagnostics.

ecuted their released Open Neural Network Exchange (ONNX) graphs, GraphCast retained its JAX implementation, and the remaining DWP systems used their packaged PyTorch implementations.

For FCN3, the random number generator was reset to seed 333 when the model was loaded, and only one realization was evaluated.

GFS was treated differently because it is an operational numerical forecast, not a checkpoint initialized from ERA5. The benchmark used Earth2Studio's `GFS_FX` source to retrieve the NOAA/NCEP forecast initialized at the task cycle. Consequently, GFS retained its operational data assimilation state while sharing the same valid time, requested variables, and verification operators as the DWP models.

F Additional Experimental Results

F.1 Conventional RMSE Comparison

The conventional comparison uses RMSE for the 15 shared fields on pressure levels $\{z, t, q, u, v\} \times \{850, 700, 500\}$ on the same event support defined by the reference. For each event category and variable, task RMSE is averaged within event and then log transformed and standardized across the model–event panel. The RMSE composite skill for each event is the negative mean of these standardized log RMSE values, so larger values mean lower field error. This composite is an internal summary for comparison with the Physics Score; it is not an externally calibrated accuracy scale.

For heat, cold, and precipitation, the conventional and physical structure evaluations use the catalog mask. For TC, primitive fields are sampled on the common grid with 25 km spacing relative to each storm center after the forecast and ERA5 centers are diagnosed independently, matching the centering convention of the TC physical structure diagnostics. Correlation calculations across events require common finite support across all seven systems. This retains 314 heat, 106 cold, 275 precipitation, and 55 TC events. The TC benchmark itself still contains and evaluates 125 target times; the 55-event intersection is used only where a finite paired Physics Score is required and is unrelated to early leads excluded by the catalog.

Across seven systems and four event categories, all 28 correlations between RMSE composite skill and Event Physics Score are positive, and every nominal percentile bootstrap 95% interval excludes zero. The median correlations across systems are 0.611 for heat, 0.762 for cold, 0.534 for precipitation, and 0.867 for TC. Figure 8 shows the coefficients and their intervals; Figure 9 summarizes the coefficients in a heatmap. These associations support relatedness, not equivalence: a monotone relationship across events does not imply that RMSE identifies which balance, kinematic, thermodynamic, transport, or vortex structure differs.

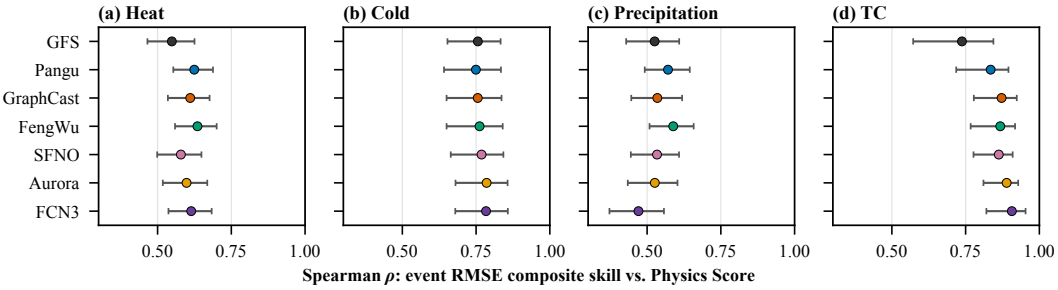

Figure 8: Association across events within each model between conventional RMSE composite skill and Event Physics Score. Points show Spearman's ρ and bars show percentile bootstrap 95% intervals; all models use common events within each category.

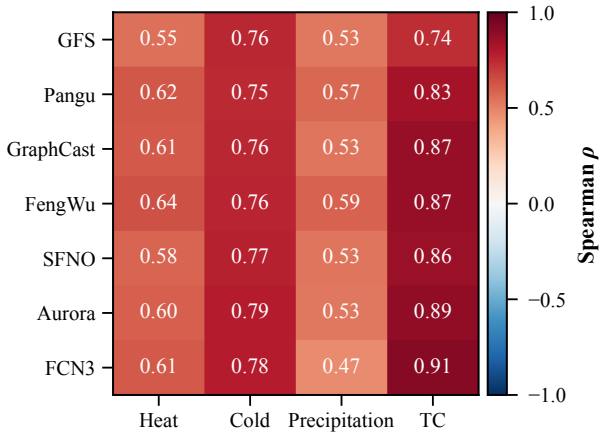

Figure 9: Spearman correlation across the full common event sample between conventional RMSE composite skill and Event Physics Score for every model and event category, using common event support for all models.

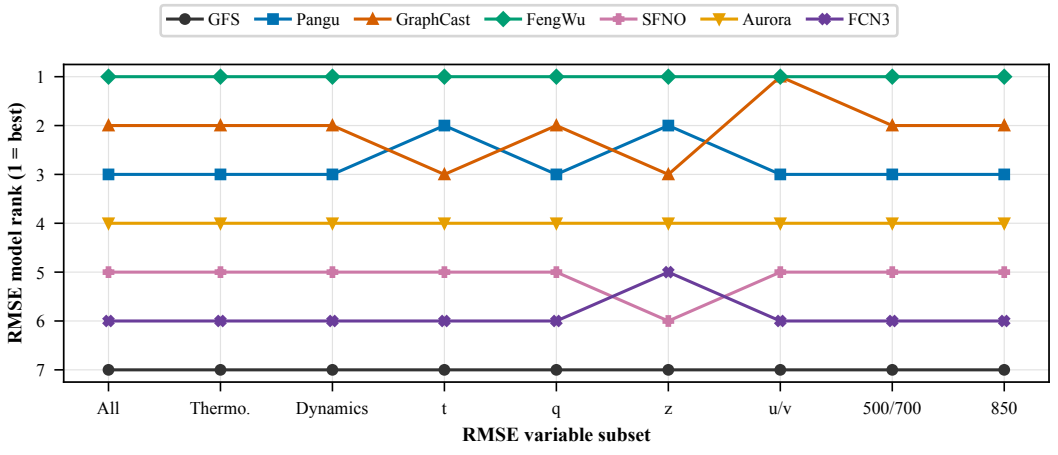

Figure 10: Sensitivity of model ranks from conventional RMSE to nine subsets of variables and pressure levels. For each subset, ranks of mean RMSE for each variable are averaged over variables and then equally over event categories. Rank 1 is best.

Figure 10 shows that the conventional model order also depends on which primitive variables are summarized. The main comparison uses all 15 fields; the subsets are sensitivity analyses, not alternative benchmark leaderboards.

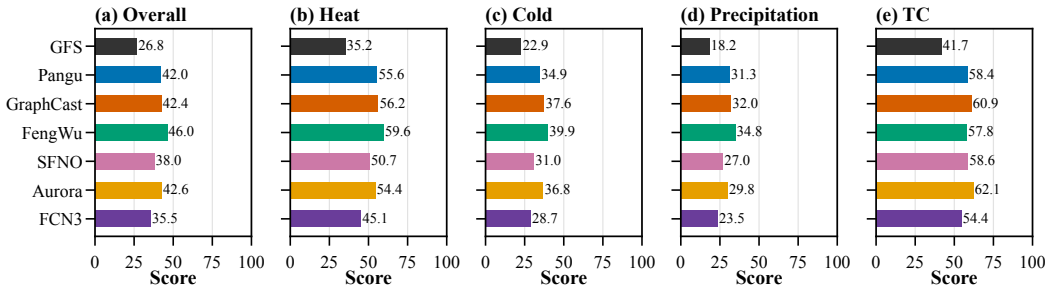

Figure 11: Overall Physics Score and Category Physics Scores from the formal scoreboard. Overall uses only the universal diagnostic group with equal weights across the four event categories; Category Physics Scores combine universal and applicable groups of event-specific diagnostics. Higher values indicate closer agreement with ERA5 under the declared diagnostic scales and tolerances.

The formal leaderboard uses the complete declared support for each summary: its correlation between model ranks is 0.893 Overall, 1.000 for heat, 0.857 for cold, 1.000 for precipitation, and 0.703 for TC. The view of Event Physics Scores on matched support drawn in Figure 7 is slightly different because its plotted Overall coordinate averages Event Physics Scores on common support; its correlations are 0.964 Overall and 0.750 for TC. In either construction, TC exhibits the largest rank disagreement. This estimator distinction is why the main text does not treat a rank plot as another definition of the formal Overall Physics Score.

F.2 Aggregate Model Comparisons

Figures 11–13 complement the formal scoreboard with category, contribution, and lower tail views. FengWu has the highest Overall Physics Score and leads heat, cold, and precipitation, while Aurora has the highest TC Category Physics Score. All six DWP systems have positive contributions from the universal group relative to GFS in all four event categories. The lower tail view in Figure 12 also places FengWu highest in the mean of the category fifth percentiles. These are descriptive summaries of physical structure fidelity, not additional hypothesis tests or measures of hazard skill.

F.3 Event Distributions and Paired Comparisons

Figure 14 shows the distributions across events behind the category means. Their overlap makes clear that an ordering of model averages is not a statement about every event. Figure 15 therefore compares each DWP system with FengWu on paired events. All five Overall differences are negative, with nominal 95% intervals excluding zero. By category, every comparison system is lower for heat, cold, and precipitation. TC differs: Aurora is $+4.33$ points $[1.94, 6.88]$ and GraphCast is $+3.11$ $[0.70, 5.82]$ relative to FengWu; intervals for Pangu and SFNO include zero; FCN3 is lower by -3.43 $[-6.23, -0.73]$. These are nominal intervals without multiplicity adjustment. For Overall, paired events are resampled within each category before averaging universal scores with equal category weights.

Pairwise win rates in Figure 16 answer a different question from mean differences. They count the fraction of paired events on which the row system has the larger score, with ties contributing one half. Consequently, they need not reproduce the order of average score differences.

F.4 Physical Structure Profiles by Diagnostic

Figure 17 reports the Diagnostic Physics Scores from which the physical family and group summaries are constructed. Every column is mapped using its own fixed dimensional $C_i\tau_i$ scale before averaging, so color differences between columns are score differences under the declared tolerances rather than comparisons of raw units. Figure 18 provides paired descriptive differences from GFS, and Figure 19 separates pooled correlations from correlations after removing additive model and event effects. None of these auxiliary views changes the aggregation weights in the formal scoreboard.

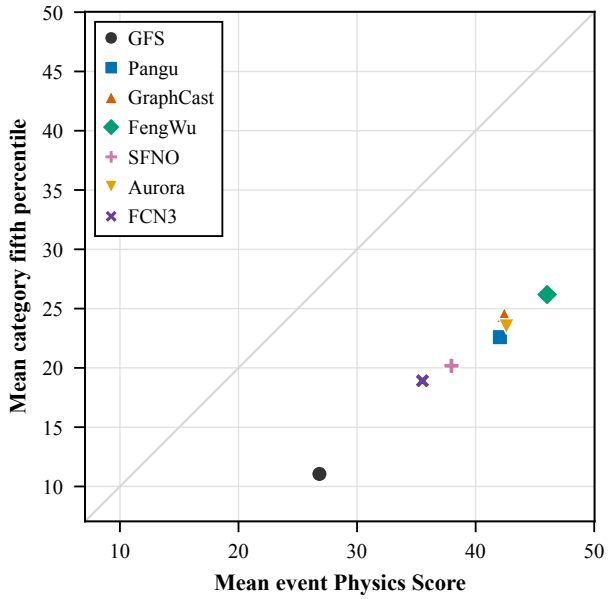

Figure 12: Mean versus lower tail performance of the universal diagnostic group on common finite event support across all seven systems. The horizontal coordinate averages event means equally across the four event categories; the vertical coordinate averages their fifth percentiles with the same category weights. Both axes use scores for the universal group, not the combined Event Physics Score. The diagonal marks equality, and distance below it is not a confidence interval.

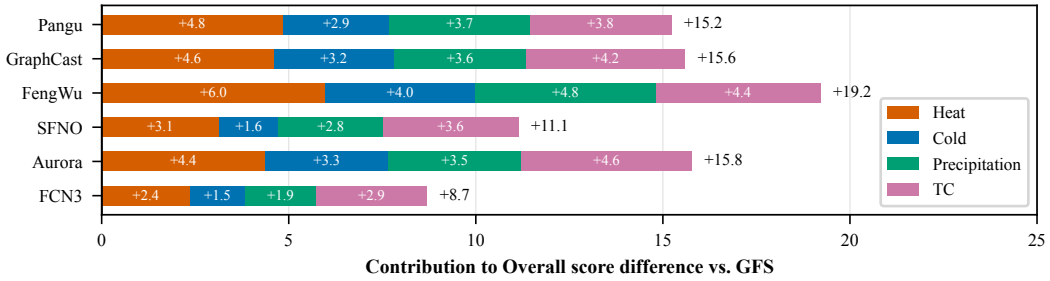

Figure 13: Contributions to the Overall Physics Score difference from GFS. Each segment is one quarter of the difference in mean score for the universal group for the named event category; the printed total is their sum. Diagnostics specific to an event category do not enter this decomposition. Positive values favor the named system; values are point estimates without uncertainty intervals.

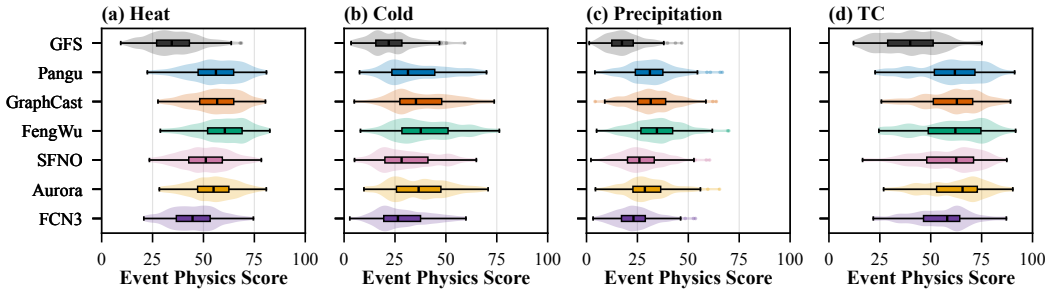

Figure 14: Distributions of Event Physics Scores by model and event category. Each event first averages mapped task scores within diagnostic, then diagnostics within physical family, physical families within group, and the universal and event-specific diagnostic groups equally.

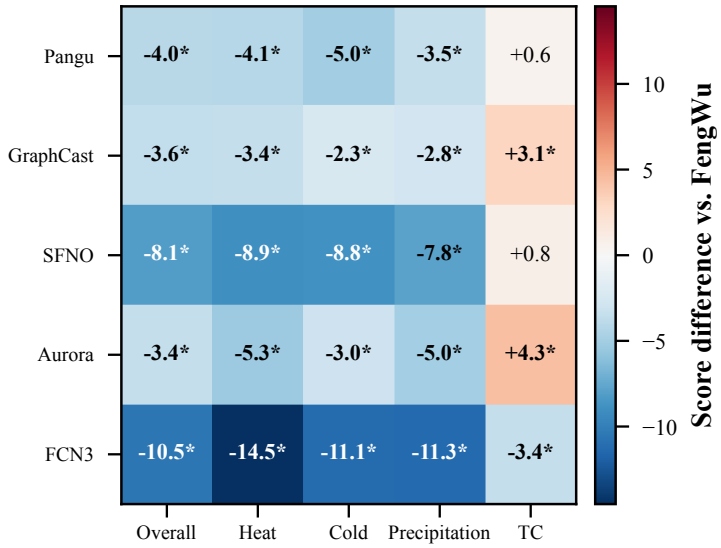

Figure 15: Physics Score differences from FengWu for the other five DWP systems. Category intervals resample paired events; Overall gives the differences in universal scores across the four event categories equal weight. Positive values favor the named comparison system; asterisks mark nominal paired bootstrap 95% intervals that exclude zero, without multiplicity adjustment.

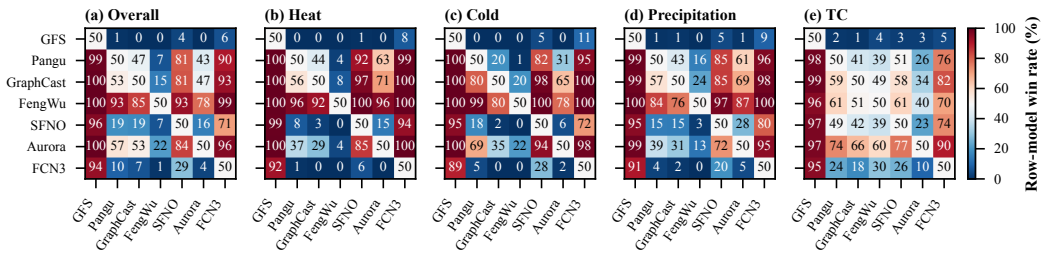

Figure 16: Pairwise Event Physics Score win rates. The diagonal is 50%; Overall is the unweighted mean of the four event category matrices. The matrix reports frequency, not difference magnitude or uncertainty.

Diagnostic contrasts within an extreme weather event category. The heatmap in Figure 17 shows that the leading system can change even between closely related diagnostics. For extreme precipitation events, GraphCast has the highest water vapor score (E1), while FengWu leads vapor-flux convergence (E3), low-level moisture-flux convergence (E4), and the MSE layer difference (E5). Thus, agreement in layer water amount does not by itself establish agreement in its transport or vertical thermodynamic context. For TC events, Aurora leads axisymmetric tangential wind (E1) and the two event-specific rotation diagnostics (E2–E3), whereas GraphCast leads warm-core thickness (E5) and FengWu leads upper-level outflow (E4). These diagnostic comparisons complement the physical family comparisons on common event support in the main text.

Figure 18 makes the GFS comparison local to each diagnostic. Broad improvements can coexist with small or negative differences in individual columns; for example, the TC hydrostatic scores (U1) of Pangu, GraphCast, and SFNO are close to the GFS value in the absolute heatmap. The paired differences are the appropriate view for assessing those contrasts because they restrict each comparison to shared available events. A large positive cell identifies a diagnostic contribution worth examining, but its size alone does not establish statistical significance or determine the category ranking, which also depends on the physical family and group weights.

Shared variation among diagnostics. In Figure 19, the pooled lower triangles contain widespread positive associations: model–event combinations with high scores for one physical structure often

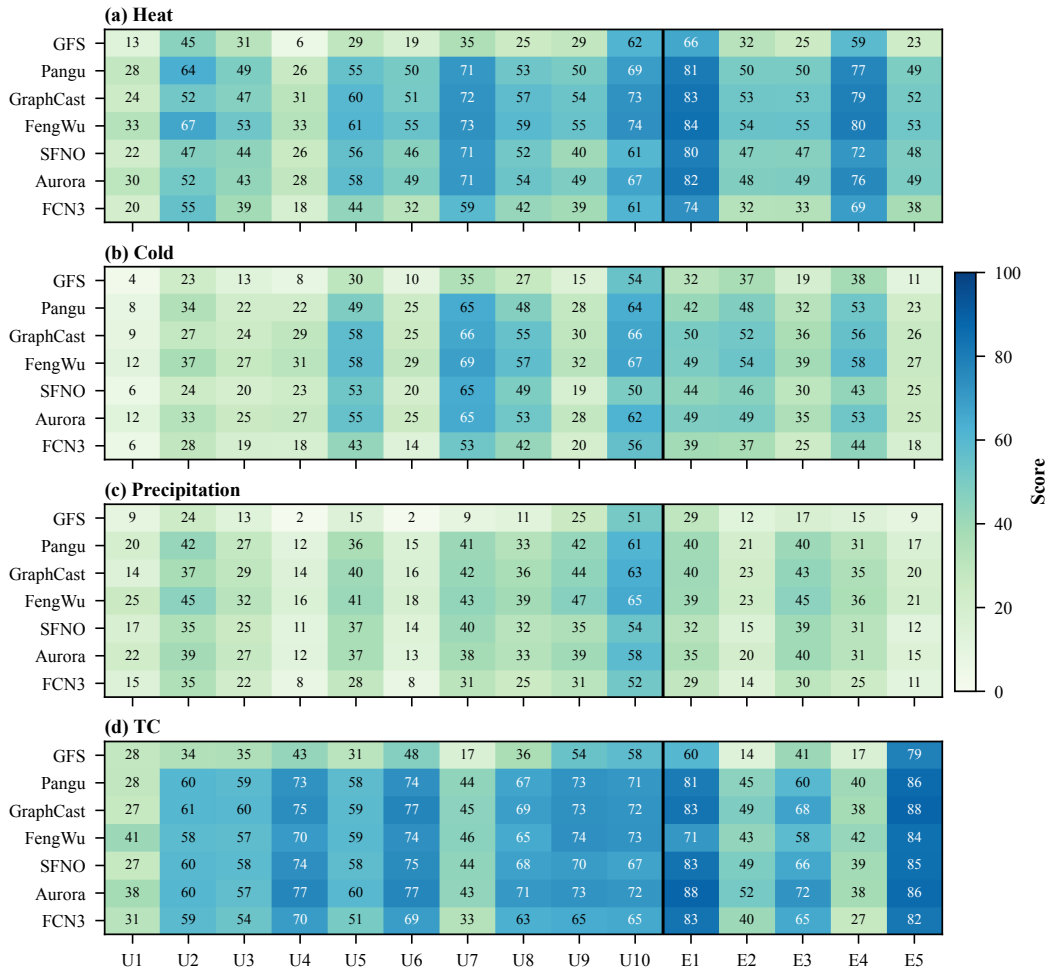

Figure 17: Mean Diagnostic Physics Scores by event category. Rows are forecast systems; columns U1–U10 and E1–E5 follow Tables 9 and 10. Mapped scores are averaged over leads within each event and then over applicable events. Diagnostic support differs from the common complete events used in Table 5; higher scores indicate closer agreement with ERA5 under fixed diagnostic scales.

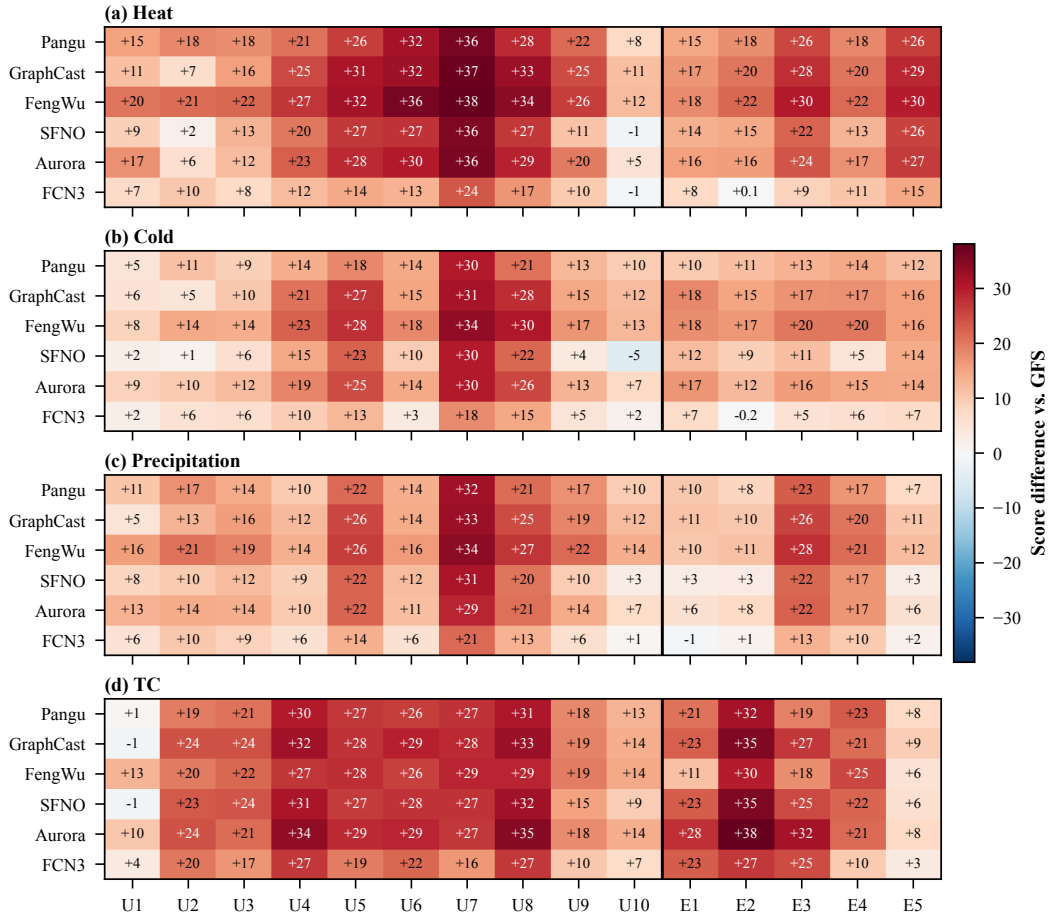

Figure 18: Diagnostic Physics Score differences from GFS. Each non-GFS system is paired with GFS on common available events; positive values favor the named system. Values are point estimates without intervals for individual diagnostics.

Table 16: Descriptive Physics Score change from 24 to 120 h, expressed per 24 h.

Model	Overall	Heat	Cold	Precipitation	TC
FengWu	−5.18	−3.06	−5.17	−5.80	−7.79
Aurora	−6.45	−4.64	−7.00	−7.35	−6.39
GraphCast	−5.30	−3.74	−5.82	−6.37	−4.89
Pangu	−4.23	−3.18	−4.49	−5.51	−3.51
SFNO	−5.89	−4.54	−6.19	−6.72	−6.35
FCN3	−5.61	−4.20	−6.13	−6.07	−5.64
GFS	**−3.50**	**−3.06**	**−3.85**	**−4.09**	**−3.01**

Each entry is $[S(120) - S(24)]/4$; bold marks the largest (least negative) value in each column. Endpoints summarize the events available at each lead, and the TC cohort varies with lead. Values are descriptive, not fitted slopes for a fixed cohort.

also score well for others. Many associations weaken in the upper triangles after additive model and event effects are removed. This reduction indicates that part of the pooled association reflects differences in average model performance and event difficulty. Some links remain strong, including the relationship between geostrophic wind and geostrophic shear (U2–U3), consistent with their shared dependence on geopotential gradients. The residual correlations describe co-variation beyond the fitted additive effects; they do not isolate causal relationships or demonstrate that diagnostics are interchangeable.

Figure 20 summarizes similarities among the diagnostic profiles, while Figures 21–24 retain individual diagnostic identities within each event category. The radar views compare each system with the median across the seven systems, helping locate relative strengths and weaknesses that a category mean can hide. Neither geometric proximity nor radar area defines an additional Physics Score.

Similarity of complete score profiles. The first two principal components in Figure 20 explain 84.1% and 8.5% of the variance, respectively, retaining 92.6% of the variance in this standardized representation. The projection therefore captures much of the variation across the seven systems, while the Ward tree provides a view using all retained features. Each category–diagnostic feature is standardized across systems, so a diagnostic with a narrow absolute score range can still influence the geometry. These equal feature weights differ from the benchmark's physical family and group aggregation; proximity in this figure should be read as similarity of relative diagnostic profiles.

Category-specific strengths in the radar views. For extreme heat stress events (Figure 21), FengWu's profile is broadly strong across both the universal and event-specific blocks, consistent with its advantage across all seven applicable physical families. The profiles for extreme cold stress events (Figure 22) are less uniform: Aurora slightly exceeds FengWu in hydrostatic structure (U1), and GraphCast leads temperature advection (E1), while FengWu retains advantages in moisture and MSE advection (E2–E3) and lower-layer stability (E4). These local departures explain why the best category mean need not imply the best value on every axis. For extreme precipitation events (Figure 23), the contrast between water amount (E1), vapor supply and convergence (E2–E4), and vertical energy contrast (E5) is particularly useful: the axes separate processes that are combined in the category summary. The profiles for TC events (Figure 24) instead separate circulation, rotation, outflow, and warm-core structure. Aurora's strong circulation profile and GraphCast's strong warm-core score therefore represent different physical structures within TC events. In all four panels, the median is an empirical reference across the seven evaluated systems, not a physical acceptance threshold.

F.5 Detailed Views by Forecast Lead

Figure 3 summarizes score evolution. The common Overall window is 24–120 h because the gridded events use daily leads and TC uses 6-hourly leads. Figure 25 annotates the absolute values. Aurora is 0.15 points above FengWu Overall at 24 h; FengWu leads from 48 h onward. Between 24 and 120 h, all displayed model–scope series decline. Table 16 and Figure 26 report the endpoint changes per 24 h. The Overall change ranges from -6.45 for Aurora to -3.50 for GFS. The smaller GFS decline occurs from a substantially lower starting score and is not evidence of better absolute structure fidelity.

Heat scores decline less rapidly than cold or precipitation for most systems. The bounded score map compresses large errors, so score changes must be read with their starting values and are not dimensional error growth rates. The TC cohort contracts from 104 events at 24 h to 32 at 120 h.

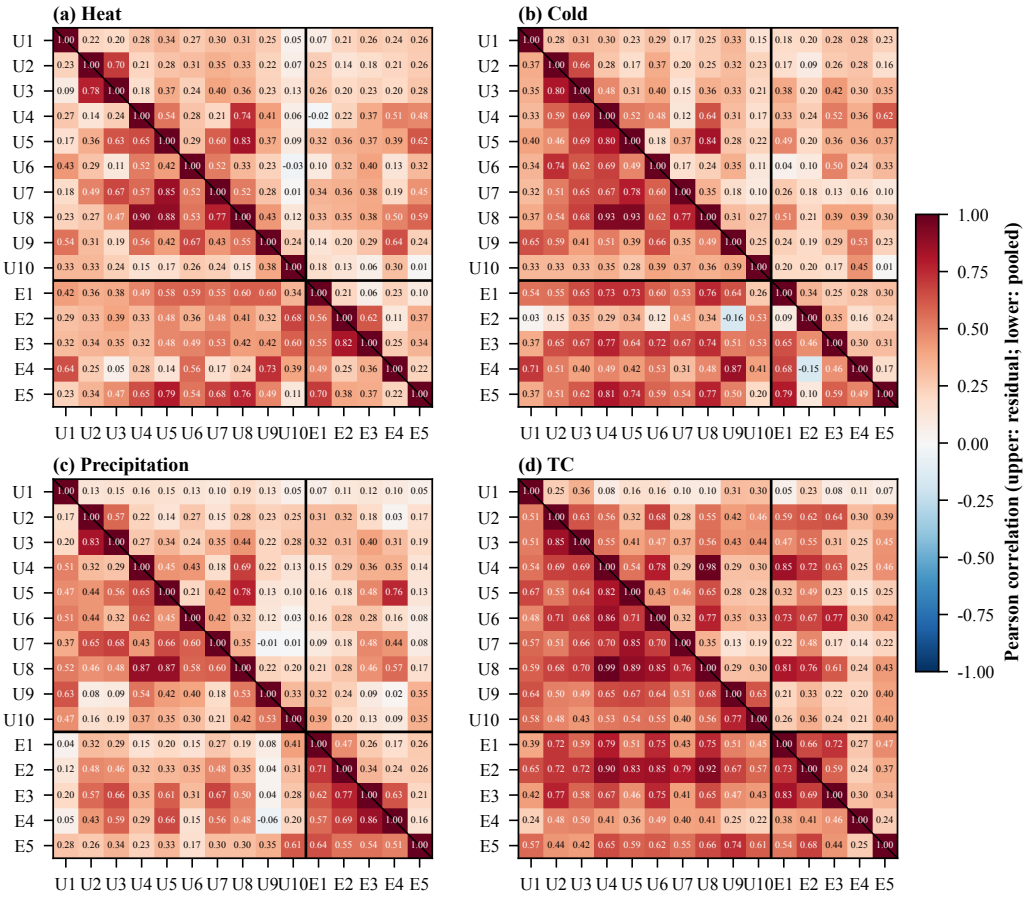

Figure 19: Pearson correlations among Diagnostic Physics Scores within events. The lower triangle pools model–event observations; the upper triangle correlates residuals after removing additive model and event effects. Heavy rules separate U1–U10 from E1–E5.

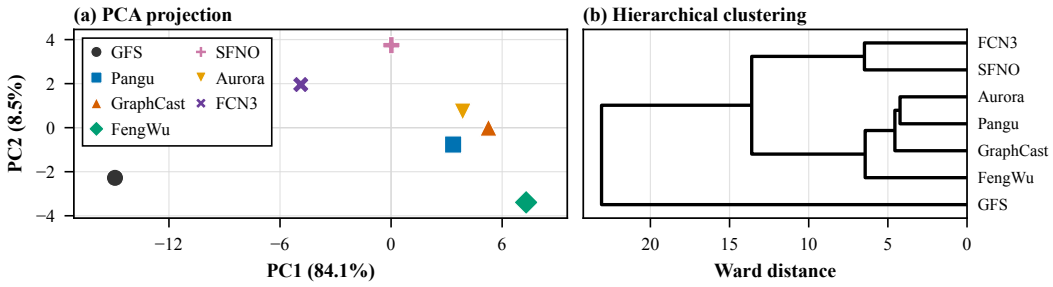

Figure 20: Similarity among systems from 60 Diagnostic Physics Score features (four event categories times 15 applicable diagnostics), each standardized across the seven systems. Panel (a) projects these features onto two principal components; panel (b) applies Ward clustering. Distances and clusters describe this diagnostic representation, not shared architecture or statistically distinct model families.

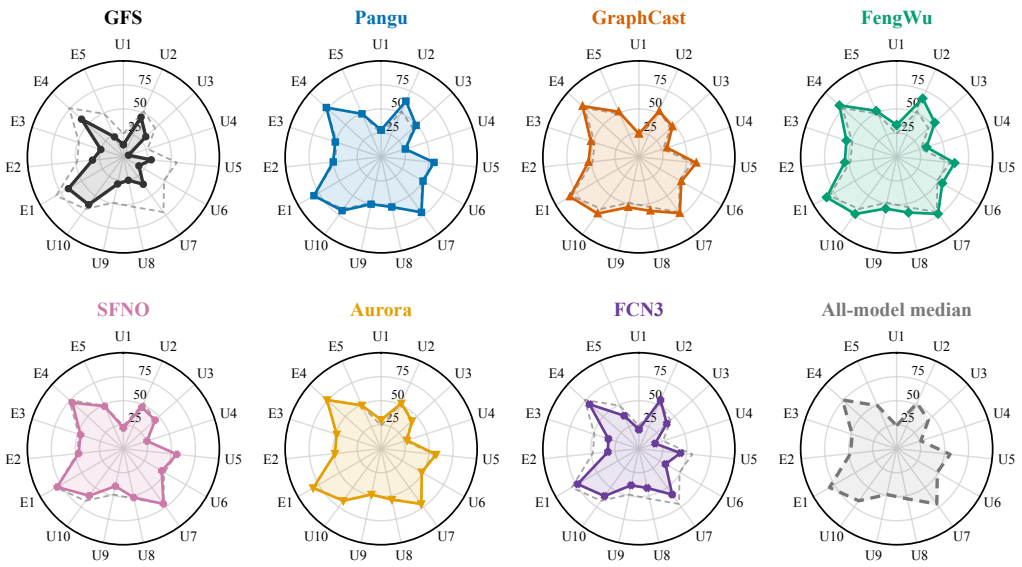

Figure 21: Diagnostic Physics Score profiles for heat events. Each system is compared with the median across all seven systems on each diagnostic. U1–U10 denote universal diagnostics and E1–E5 the applicable block of event-specific diagnostics; codes follow Tables 9 and 10. Radial values use fixed diagnostic scales and tolerances; enclosed area is not an aggregate score.

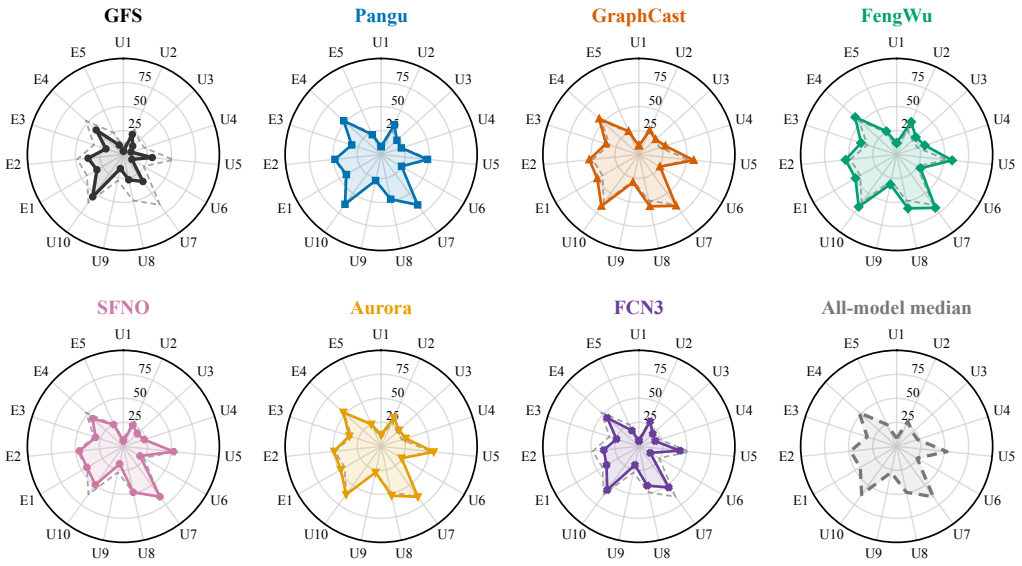

Figure 22: Diagnostic Physics Score profiles for cold events. Each system is compared with the median across all seven systems on each diagnostic. U1–U10 denote universal diagnostics and E1–E5 the applicable block of event-specific diagnostics; codes follow Tables 9 and 10. Radial values use fixed diagnostic scales and tolerances; enclosed area is not an aggregate score.

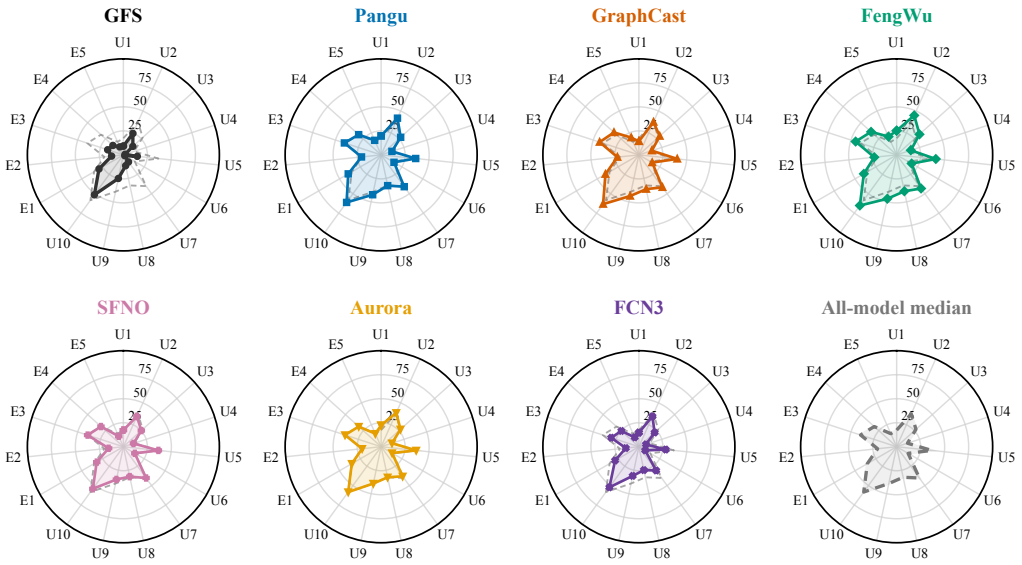

Figure 23: Diagnostic Physics Score profiles for precipitation events. Each system is compared with the median across all seven systems on each diagnostic. U1–U10 denote universal diagnostics and E1–E5 the applicable block of event-specific diagnostics; codes follow Tables 9 and 10. Radial values use fixed diagnostic scales and tolerances; enclosed area is not an aggregate score.

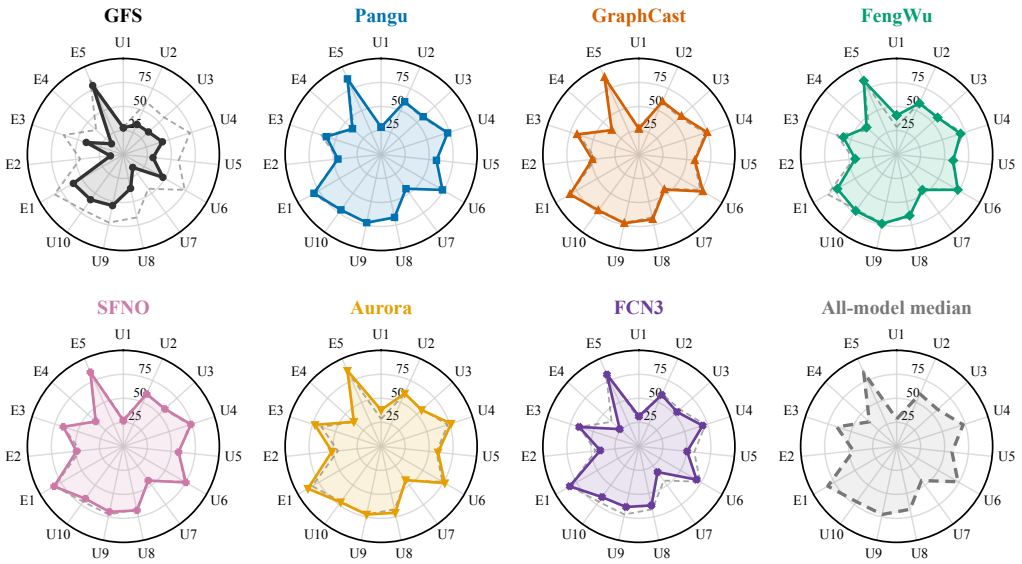

Figure 24: Diagnostic Physics Score profiles for TC events. Each system is compared with the median across all seven systems on each diagnostic. U1–U10 denote universal diagnostics and E1–E5 the applicable block of event-specific diagnostics; codes follow Tables 9 and 10. Radial values use fixed diagnostic scales and tolerances; enclosed area is not an aggregate score.

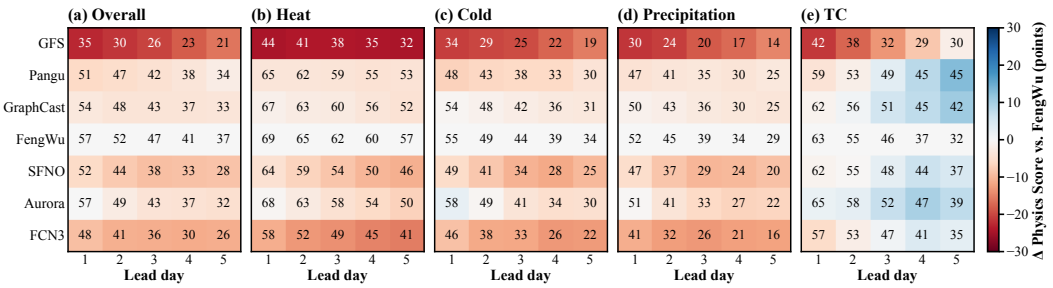

Figure 25: Physics Score scorecard over 24–120 h. Labels are absolute scores; shading is the difference from FengWu. Overall uses only universal scores and gives the four event categories equal weight; category panels combine universal and event-specific diagnostic groups.

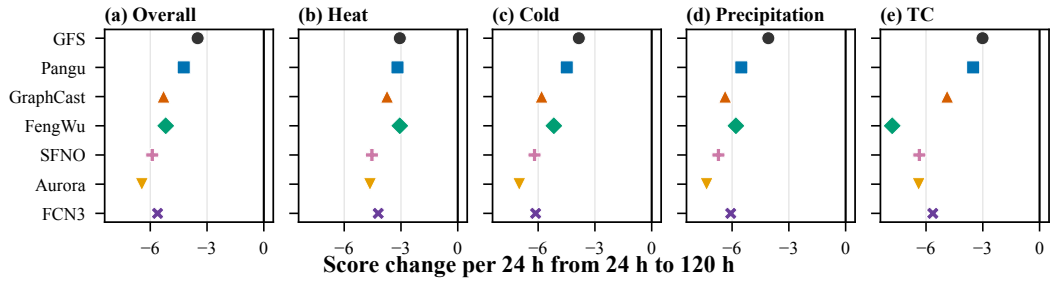

Figure 26: Endpoint change $[S(120) - S(24)]/4$ in Physics Score points per 24 h. Negative values denote a lower score at 120 h. Endpoints use the events available at each lead; markers are descriptive changes, not fitted rates.

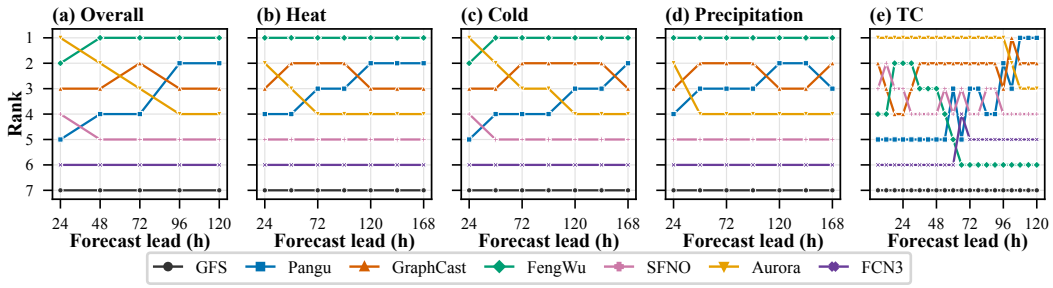

Figure 27: System rank by forecast lead; rank 1 has the highest Physics Score. Ranks describe the available event cohort at each lead and have no uncertainty intervals.

TC changes therefore combine forecast lead and cohort composition. In particular, FengWu's TC Category score falls from 62.94 to 31.79, while Pangu falls from 58.72 to 44.69 and becomes 12.90 points higher at the last displayed lead. The comparison uses the events eligible at each lead and does not isolate error growth in a fixed set of storms.

Figure 27 gives the corresponding rank paths, and Figure 28 shows absolute score differences from GFS. Figures 29–32 reproduce the diagnostic lead curves for all four event categories, showing how individual physical structures evolve rather than reducing them to one category mean.

Rank changes and absolute margins. Figure 27 emphasizes crossings between systems, whereas Figure 28 retains the score margins relative to a common baseline. Reading them together distinguishes changes in ordering from changes in the size of an advantage: a rank crossing can occur with a small score difference, and an unchanged rank can accompany a substantial decline.

Heat events: differences in score decline across physical structures. Figure 29 shows diagnostic scores for extreme heat stress events through 168 h. Geostrophic wind and shear (U2–U3), dry stability (U9), and MSE (U10) show pronounced score declines. By comparison, horizontal divergence (U5 and U7), temperature advection (E1), and frontogenesis (E5) retain flatter score tra-

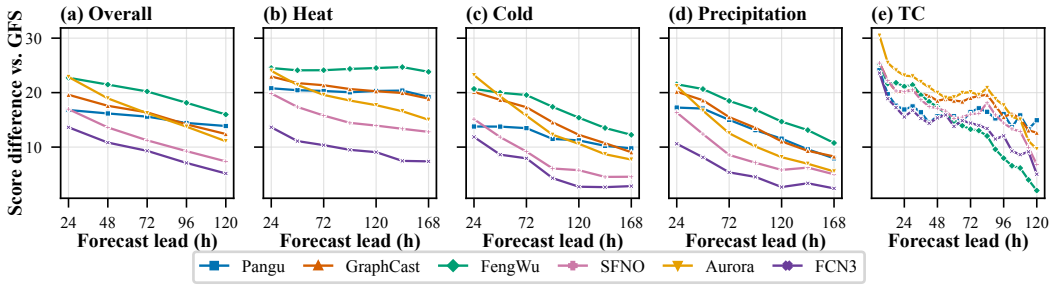

Figure 28: Physics Score differences from GFS by lead. Positive values favor the named system. Overall uses the common universal block; categories use their universal and event-specific diagnostic groups.

jectories. The Category Physics Score for heat events therefore summarizes agreement in physical structures with different lead dependence. FengWu remains strong across many diagnostics, but the separation among systems and the amount of score loss vary by axis.

Cold events: declining scores for stability and transport. For extreme cold stress events (Figure 30), dry stability scores fall markedly in both the universal diagnostic for the 850–500 hPa layer (U9) and the event-specific diagnostic for the 850–700 hPa layer (E4). Scores for geostrophic wind and shear (U2–U3) and temperature and MSE advection (E1 and E3) also decline, while divergence scores (U5 and U7) change more gradually. Several low-scoring diagnostics become more closely grouped at long leads. Such convergence near the lower part of the bounded score scale should not be interpreted as convergence of the underlying dimensional errors.

Precipitation events: different score trends for water supply and convergence. Figure 31 shows steep score declines for layer water amount (E1), vapor-transport components (E2), and vertical MSE contrast (E5). The vapor-flux and low-level moisture-flux convergence scores (E3–E4) vary more gradually, despite their connection to the same moisture field. The universal diagnostics likewise distinguish declining geostrophic and thermodynamic scores from relatively flatter divergence scores. These curves concern the physical structures within the selected extreme precipitation events; they do not directly measure accumulated precipitation error.

TC events: different score trends for circulation and outflow. In Figure 32, FengWu's tangential-wind score (E1) declines especially strongly and falls below the other DWP systems at later leads. Its event-specific PV and midlevel-vorticity scores (E2–E3) also lose ground, whereas its outflow score (E4) remains comparatively strong. Scores for warm-core thickness (E5) decline across systems but remain relatively high. The late-lead reversals and fluctuations should be interpreted with particular care because the eligible storm sample changes with lead and diagnostic applicability.

F.6 Sensitivity to the GFS Verifying Analysis

Replacing ERA5 with archived GFS 0-h analyses changes both the reference diagnostic and the support defined by that reference. The formal GFS Overall score rises from 26.82 to 30.33, a $+3.51$ point change. The changes are $+2.83$ for heat, $+3.23$ for cold, $+1.39$ for precipitation, and $+8.55$ for TC. Table 17 reports the endpoints. These are differences between formal aggregates, each computed on its reference's own support; they therefore do not receive paired confidence intervals. Figure 33 resolves the reference sensitivity into individual diagnostics. Figure 34 summarizes the aggregate and lead comparisons, paired event-score changes, and benchmark context; only its event-distribution panel uses explicitly paired events.

The reference effect is diagnostic dependent. Figure 33 shows that the aggregate increase is not a uniform upward shift. For extreme heat and cold stress events, the lower-layer stability score (E4) increases by about 10.5 and 8.9 points, respectively, while the midlevel divergence score (U7) decreases by about 8.8 and 9.4 points. Changes for extreme precipitation events also have mixed signs: the MSE score (U10) increases by about 6.5 points, whereas the layer vapor-flux convergence score (E3) decreases by about 6.1 points. All displayed TC diagnostic changes are positive, with particularly large increases in hydrostatic structure (U1; 18.2 points), tangential wind (E1; 17.2

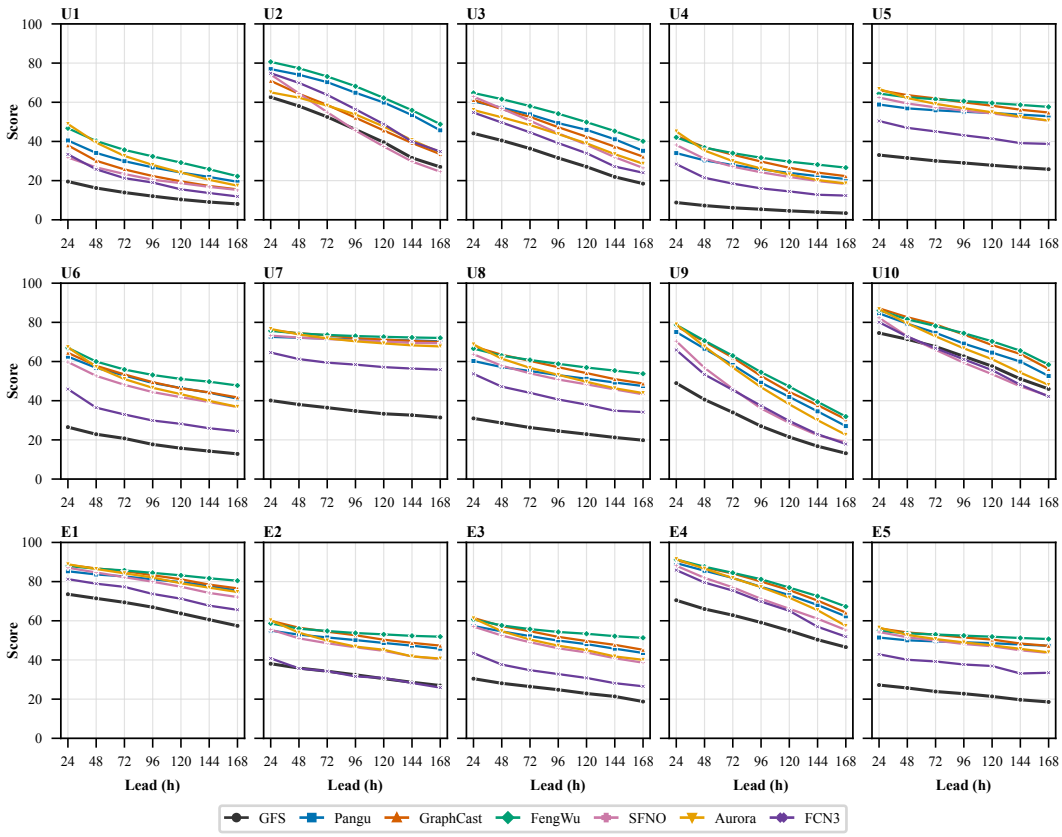

Figure 29: Diagnostic Physics Scores by forecast lead for heat events. Curves retain the U1–U10 and E1–E5 identities, allowing changes in individual physical structures to be examined alongside the aggregate lead curves. Available events receive equal weight at each lead; codes follow Tables 9 and 10.

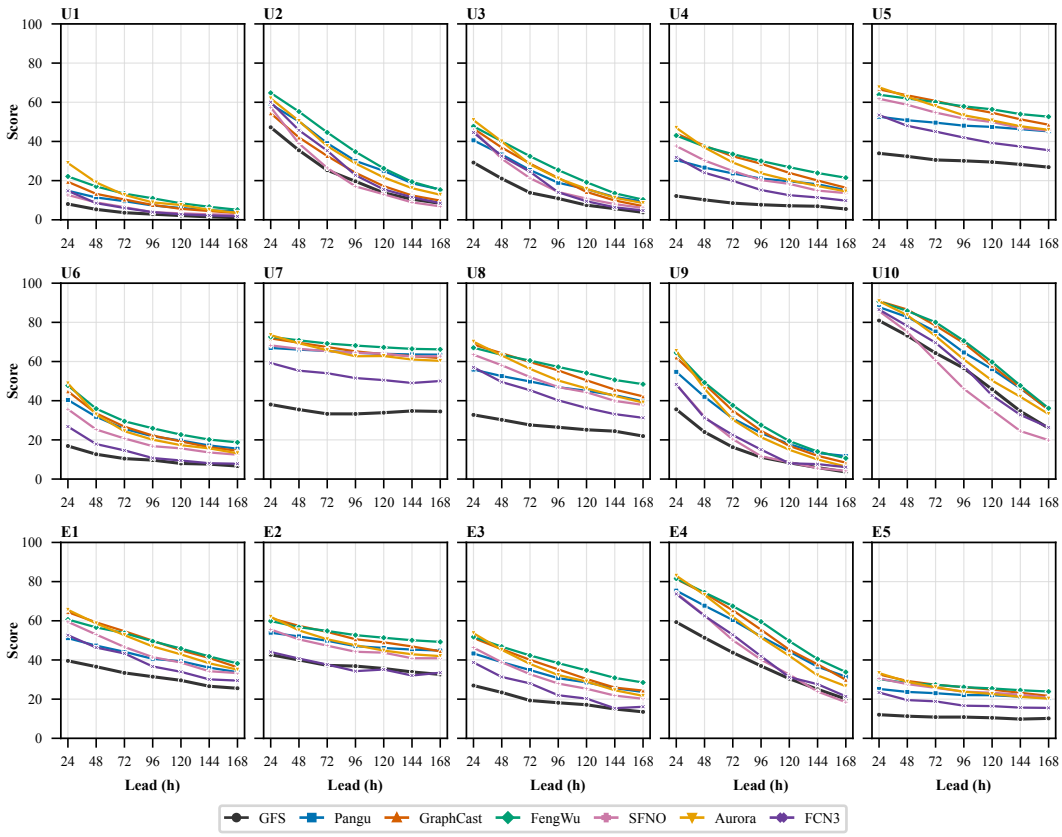

Figure 30: Diagnostic Physics Scores by forecast lead for cold events. Curves retain the U1–U10 and E1–E5 identities, allowing changes in individual physical structures to be examined alongside the aggregate lead curves. Available events receive equal weight at each lead; codes follow Tables 9 and 10.

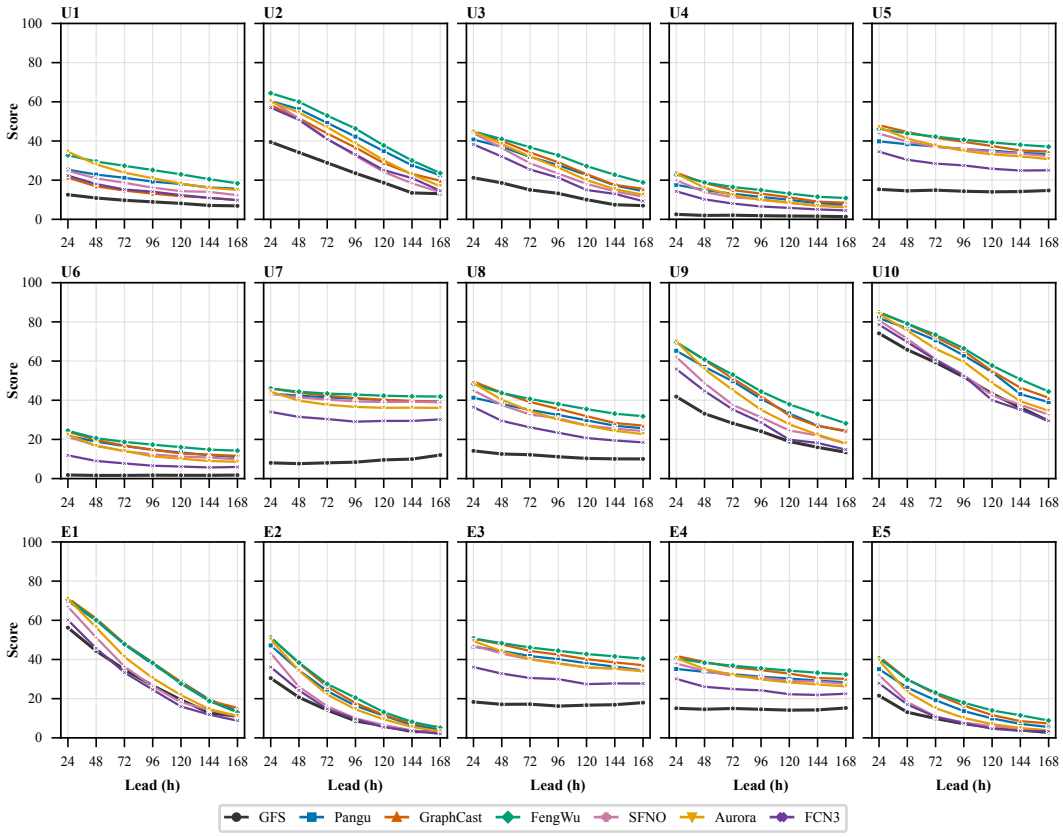

Figure 31: Diagnostic Physics Scores by forecast lead for precipitation events. Curves retain the U1–U10 and E1–E5 identities rather than collapsing processes into one category mean.

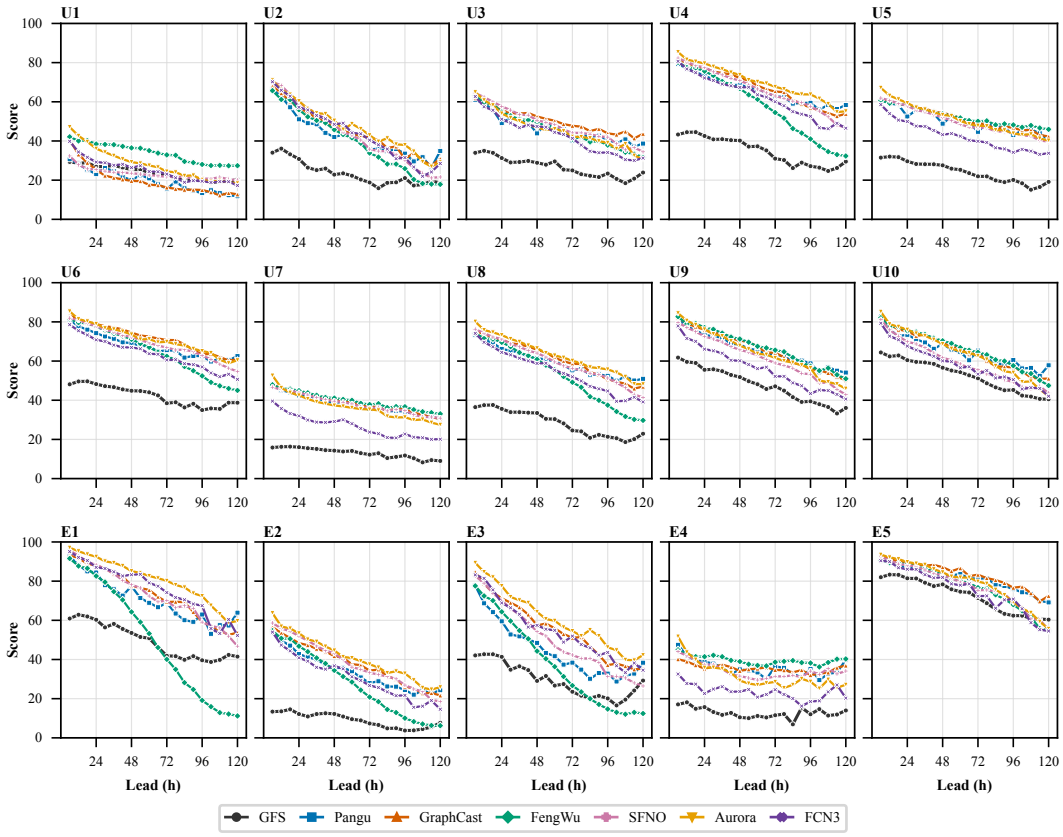

Figure 32: Diagnostic Physics Scores by forecast lead for TC events. Later leads use a smaller event cohort, so apparent trends combine forecast evolution and sample composition.

Table 17: Sensitivity of formal GFS Physics Scores to the verifying analysis.

Scope	ERA5	GFS 0-h	Difference
Overall	26.82	30.33	+3.51
Heat	35.22	38.05	+2.83
Cold	22.86	26.08	+3.23
Precipitation	18.18	19.56	+1.39
TC	41.66	50.21	+8.55

Difference is GFS 0-h minus ERA5. Each formal aggregate uses the diagnostic support admissible under its own reference; values are therefore sensitivity endpoints rather than strictly paired estimators. The GFS 0-h scores must not be inserted into the ERA5 leaderboard.

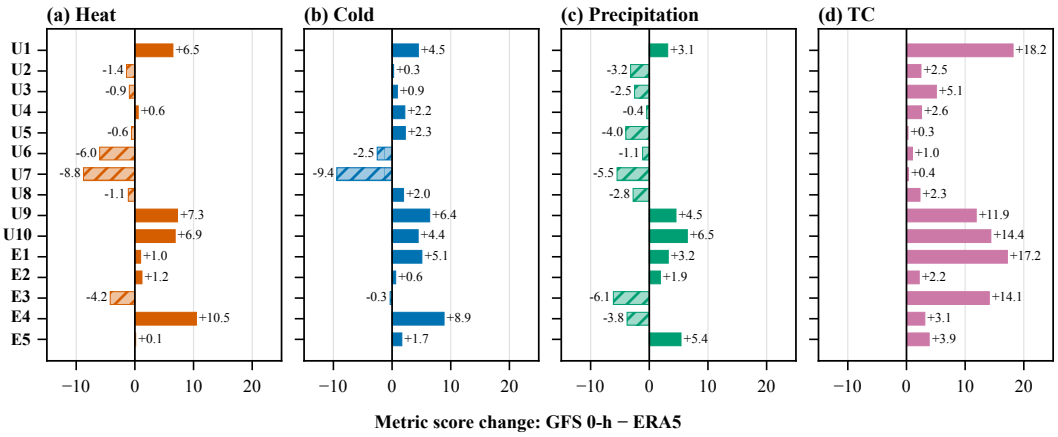

Figure 33: Change in each GFS Diagnostic Physics Score when GFS 0-h replaces ERA5 as verifying analysis. Hatched bars indicate a lower score under GFS 0-h verification. Each formal metric aggregate uses the named reference's own admissible support.

points), and midlevel radial vorticity (E3; 14.1 points). The sensitivity therefore depends on which physical structure is evaluated, in addition to the event category.

Aggregate, lead, and event views answer different questions. Panels (a) and (b) of Figure 34 show how the formal GFS score and its lead dependence change when the verifying analysis changes. Panel (c) instead restricts attention to paired events and plots the fraction whose score increase exceeds a given threshold. Its value at zero is the fraction that improves under GFS 0-h verification; the remainder of the curve describes how widespread larger increases are. This event-frequency view complements an aggregate difference, which can be influenced by a smaller number of large changes. Panel (d) places the alternative Overall value in the numerical range of the main benchmark. Because only GFS is rescored against the alternative analysis, that panel does not establish a revised cross-system ranking or separate reference-field differences from changes in admissible support.

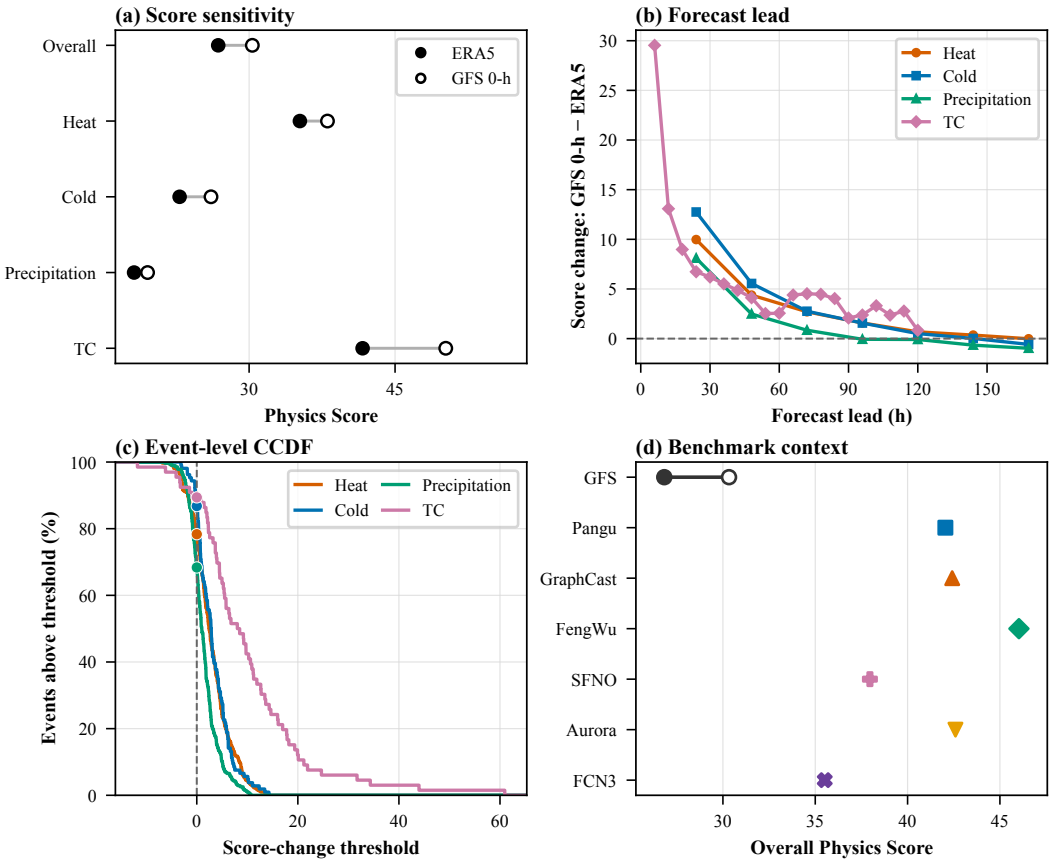

Figure 34: Sensitivity of GFS Physics Scores to the verifying analysis. Panels (a) and (b) compare formal scores on each reference's own support; panel (c) shows paired event-score changes; panel (d) places the alternative GFS endpoint beside the ERA5-referenced benchmark only as context, not as a leaderboard combining different references.