Blind-Window Forecasting: Real-Time Benchmarking and Multimodal Reconstruction for Tropical Cyclones

Zhaoran Feng[1,2] Xuanhong Chen[3,*] Zeng-Bing Chen[1]
Shengjun Wu[1,*] Bin He[4,*] Kairui Feng[4,2,*]

[1]National Laboratory of Solid State Microstructures, School of Physics and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing 210093, China

[2]Shanghai Innovation Institute, Shanghai 200231, China

[3]USC-SJTU Institute of Cultural and Creative Industry, Shanghai Jiao Tong University, Shanghai 200241, China

[4]Shanghai Research Institute for Intelligent Autonomous Systems & State Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai 201210, China

[*]Corresponding authors: chenxuanhong@sjtu.edu.cn;hebin@tongji.edu.cn; sjwu@nju.edu.cn; Kelvinfkr@tongji.edu.cn

Abstract

Tropical cyclone forecasting models are often trained and evaluated with reanalysis inputs. Reanalysis provides physically rich environmental context, but using analysis-time fields can make research evaluation less faithful to the information setting at forecast issuance. This can introduce a hindsight-style advantage in benchmark evaluation and lead to optimistic estimates of deployable forecast skill. Satellite observations, in contrast, are available closer to real time but provide incomplete and noisy views of the storm system. We formulate this data-availability mismatch as forecasting through a reanalysis blind window. To study this setting, we introduce a real-time-faithful benchmark spanning 1980–2023 that combines lagged reanalysis with visible, infrared, water-vapor, and passive microwave satellite imagery. We then develop RECAST-TC, a multimodal reconstruction framework that fuses time-lagged environmental histories with real-time satellite observations to estimate a forecast-sufficient latent storm state from delayed, partial, and noisy inputs. We use idealized analysis-available forecasting and delayed-reanalysis-only forecasting as reference regimes. RECAST-TC consistently improves track and intensity forecasts over delayed-reanalysis baselines and moves noticeably closer to the idealized reference. These results highlight data availability and observation timing as central design variables for realistic evaluation and deployment-oriented AI forecasting of tropical cyclones.

1 Introduction

Tropical cyclones (TCs) are among the most destructive weather systems on Earth, making accurate forecasts of storm track and intensity both scientifically important and operationally critical [Knapp et al., 2010, Cangialosi et al., 2020]. Recent advances in machine learning have substantially reshaped numerical weather prediction: large-scale data-driven models such as Pangu-Weather, GraphCast, FuXi, FengWu, and FourCastNet have shown that neural forecasting systems can learn forecast-relevant atmospheric structure directly from gridded reanalysis fields [Bi et al., 2023, Lam et al., 2023, Chen et al., 2023b,a, Pathak et al., 2022]. In parallel, TC-focused learning methods have begun to combine storm history, environmental predictors, and satellite observations for track and intensity forecasting [Giffard-Roisin et al., 2020, Rüttgers et al., 2019, Wimmers et al., 2019, Su et al., 2020].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

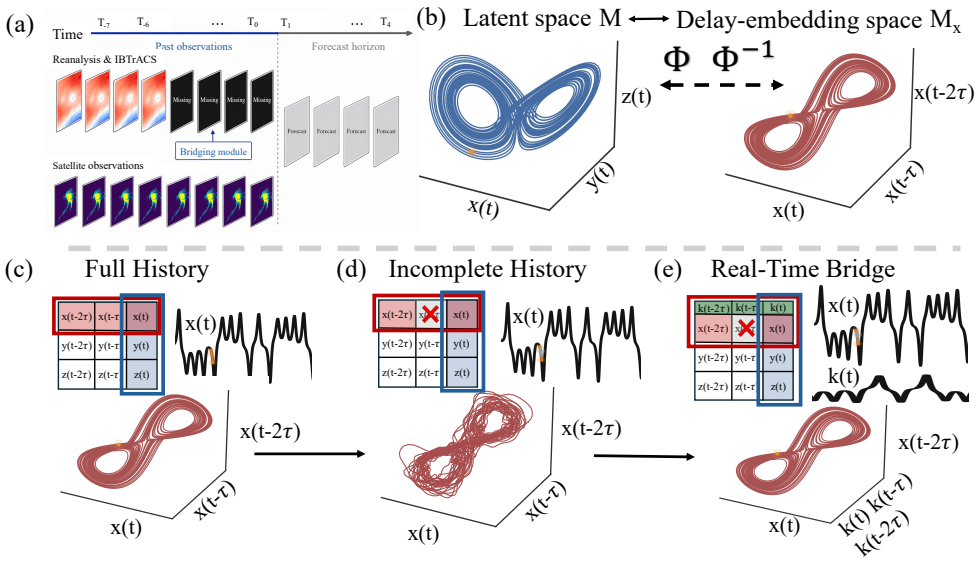

Figure 1: Real-time multimodal bridging under reanalysis latency. (a) Reanalysis latency creates a blind window, while satellites remain real-time. (b) Finite partial histories can constrain latent storm states. (c–d) Full recent observations give an oracle reference, whereas removing recent reanalysis causes information loss. (e) Auxiliary real-time modalities restore missing constraints through complementary observations.

Together, these developments suggest that data-driven models can exploit structured atmospheric and storm-specific information beyond hand-crafted predictors.

A central mismatch, however, remains under-addressed. Many TC forecasting models use reanalysis variables as standard inputs, including reanalysis-derived wind, pressure, steering-flow, sea-surface-temperature, and geopotential-height fields [Giffard-Roisin et al., 2020, Wang et al., 2023, Lam et al., 2023, Chen et al., 2023b,a, Bi et al., 2023, Huang et al., 2025]. Although highly informative, these fields are not strictly realtime products: reanalyses are produced through delayed data assimilation and retrospective processing, and ERA5 products are released with non-negligible latency before final quality-controlled versions become available [Hersbach et al., 2020]. Evaluation protocols that assume access to these latest fields therefore risk overstating deployable forecasting skill.

Table 1 illustrates that this optimism can be substantial. For global weather baselines, a nominal 6–24h TC forecast under analysis-available evaluation corresponds to an effective 30–48h forecast once the blind window is enforced, increasing ADE by about 2.5–2.7×. This shift motivates evaluating models under the actual information available at forecast issuance.

Table 1: Nominal 6–24h vs. effective 30–48h under the blind window.

Model	ADE at 6–24h	ADE at 30–48h	Shift in ADE
GraphCast	133.72	361.55	↑ 2.70×
FuXi	138.95	367.60	↑ 2.65×
Pangu-Weather	125.37	316.24	↑ 2.52×

The mismatch becomes explicit once the information available at forecast issuance is specified. Let T_0 denote the forecast initialization time. Because reanalysis products arrive with latency, the most recent environmental frames before T_0 are unavailable to an operational model, creating a reanalysis blind window. Figure 1 illustrates the three regimes studied in this work. The idealized full-history setting has access to the complete recent reanalysis trajectory and serves as an optimistic hindsight reference; the delayed-only setting removes the blind-window frames and quantifies the degradation caused by missing recent context; and the real-time bridge setting keeps those frames unavailable but adds near-real-time satellite observations as partial and noisy constraints on the evolving storm system. This recasts TC forecasting under reanalysis latency as a reduced-state, information-constrained inference problem, where delayed analyses and real-time observations are fused to recover a forecast-sufficient latent storm state.

The key question is whether real-time satellite observations carry enough latent storm-state information to bridge the blind window. We approach this question from a time-lagged mixed-observation perspective that is common in numerical dynamical systems. Classical delay-embedding results suggest that finite histories of partial observations can constrain latent state information under suitable

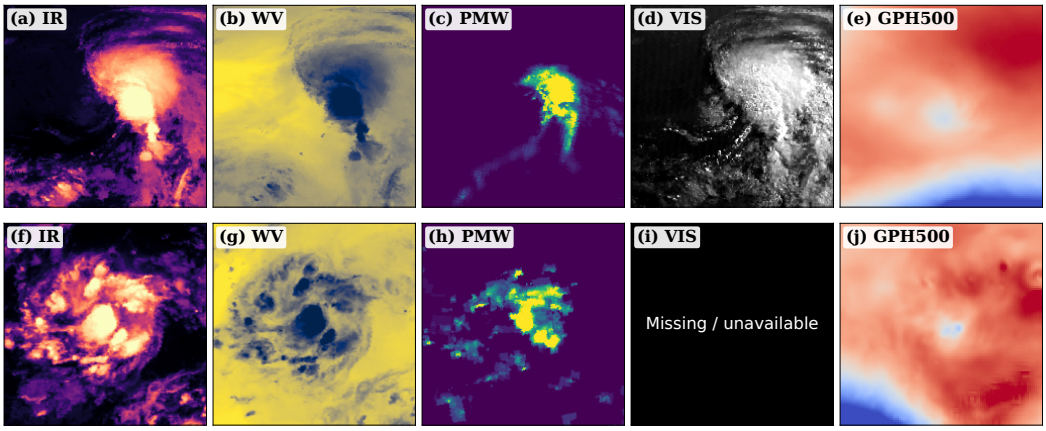

Figure 2: Dataset examples. Two storms, *Claudette* in the North Atlantic and *Amanda* in the East Pacific, are shown with IR, WV, PMW, VIS, and GPH500 fields. The examples highlight cross-basin storm diversity and heterogeneous modality availability, motivating real-time blind-window forecasting with delayed environmental analyses and partial satellite observations.

conditions [Packard et al., 1980, Takens, 1981, Sauer et al., 1991], while later extensions show that related intuitions can remain useful in noisier and more general dynamical systems [Stark et al., 1997, 2003, Hamilton et al., 2017]. To make this intuition concrete, let $z(g, T_0)$ denote the forecast-sufficient latent storm state around location g at initialization time T_0, and let $\mathcal{A}(g, T_0)$ be the set of modality–time pairs available at that location and time. We collect these available local observation directions into a mixed observation matrix

$$\mathbf{y}_{\mathcal{A}}(g, T_0) = \mathcal{O}_{\mathcal{A}}(g, T_0) z(g, T_0) + \varepsilon, \qquad \mathcal{O}_{\mathcal{A}}(g, T_0) = \left[C_m(g, t) \Phi_{t \leftarrow T_0}(g) \right]_{(m,t) \in \mathcal{A}(g, T_0)}.$$

Here, $C_m(g, t)$ is a modality-, location-, and time-dependent observation operator, and $\Phi_{t \leftarrow T_0}(g)$ relates the latent state at initialization to its observable signatures at time t. This formulation makes clear that the available information is not uniform across space or time, since different storm regions and forecast targets are observed through different modality–time directions in $\mathcal{O}_{\mathcal{A}}$. In a minimal mixed-observation model, removing the most recent reanalysis frame deletes rows from $\mathcal{O}_{\mathcal{A}}$, reducing the local observation system from full rank to rank deficient, while a real-time satellite modality can restore identifiability when it provides a non-collinear observation direction. Under noise, the recovery error is controlled by the conditioning of $\mathcal{O}_{\mathcal{A}}$. This gives the main intuition behind our architecture. The model should estimate a forecast-sufficient latent state, not reconstruct the missing reanalysis field in pixel space. The full derivation is deferred to Appendix A.

Building on this idea, we introduce a real-time-faithful benchmark and propose **RE**altime **C**ontext-**A**ware **S**tate Recons**T**ruction for Tropical Cyclones (RECAST-TC), a blind-window forecasting framework that masks recent reanalysis and uses auxiliary satellite observations to bridge the missing interval. RECAST-TC estimates a forecast-sufficient latent context for downstream prediction instead of performing full physical-field reconstruction. The observability perspective motivates Target-Adaptive Lag Routing (TALR) for target-specific temporal context, Latent Bridge Rollout for latent recovery through the blind window, and reliability-aware multimodal fusion for availability- and quality-dependent use of VIS, IR, WV, and PMW. Architectural details are provided in Sec. 3.1.

Our contributions are as follows:

- **Realtime mismatch.** We highlight a potential hindsight-style advantage in TC forecasting benchmarks caused by reanalysis inputs that may be unavailable at operational issuance time.
- **Blind-window benchmark.** We introduce a realtime-faithful benchmark that pairs lagged reanalysis with VIS, IR, WV, and PMW satellite observations.
- **Latent state reconstruction.** We propose RECAST-TC, which recovers forecast-sufficient storm states from delayed, partial, noisy, and asynchronous observations.
- **Empirical gains.** We show that RECAST-TC improves over lagged-reanalysis baselines, closes much of the gap to hindsight-style references, and yields complementary intensity gains.

2 Dataset and Benchmark Protocol

To evaluate forecasting through the blind window, we construct a realtime-aware tropical cyclone dataset and benchmark protocol that explicitly separates delayed environmental analyses from auxiliary realtime observations. The goal is not merely to assemble more data sources, but to reproduce the actual information structure at forecast issuance time: recent environmental analyses may be unavailable because of latency, whereas satellite observations remain accessible and can potentially bridge the missing recent context.

Forecasts are initialized at T_0 using eight historical frames sampled every 6 h from $T_0 - 42\,\mathrm{h}$ to T_0. The prediction lead times are 6, 12, 18, and 24 h. Under the default 24-h blind window, ERA/GPH inputs are available only through $T_0 - 24\,\mathrm{h}$, with the four more recent frames masked. Satellite inputs are restricted to observations acquired by T_0, with missing observations explicitly masked. Timestamp filtering is applied before preprocessing and fusion. The complete historical ERA sequence is used only by a stop-gradient EMA teacher during training and never enters the student forecast path or inference.

Table 2: Comparison with representative TC and weather forecasting datasets.

Criterion	Dig.Typh.	IBTrACS	WB2	SHIPS	TCND	Ours
Best-track	✓	✓	×	✓	✓	✓
Gridded env.	×	×	✓	✓	✓	✓
Physics diag.	×	×	×	✓	×	✓
Cross-scale bench.	×	×	×	×	×	✓
Global coverage	×	✓	✓	×	✓	✓
Multimodal	×	×	×	×	✓	✓
Forecast utility	×	×	✓	✓	✓	✓
Leakage-free split	×	×	✓	×	✓	✓
TC protocol	×	×	×	✓	✓	✓
Realtime	×	×	×	×	×	✓

Accordingly, our benchmark distinguishes three settings. In the oracle-analysis setting, the full recent environmental history is available and serves as an optimistic hindsight-style reference. In the delayed-only setting, the most recent analysis frames inside the blind window are removed, exposing the performance degradation caused by analysis latency. In the realtime-bridge setting, those recent analysis frames remain unavailable, but auxiliary satellite observations are provided so that the model can attempt to recover a forecast-sufficient latent state from delayed, partial, noisy, and heterogeneous observations. This protocol directly matches the problem formulation introduced in Sec. 1 and derived in Appendix A.

The dataset contains three complementary levels of information: best-track targets for TC position and intensity, gridded environmental variables with physics-aware diagnostics, and aligned multimodal satellite observations that remain available in realtime. Together, these components support both conventional TC prediction and direct evaluation of whether missing recent environmental context can be bridged in a forecast-useful manner.

To ensure leakage-free evaluation, we split the dataset by storm identity rather than by individual samples. Training and validation use disjoint storms from 1980–2016, whereas testing uses only storms from 2017–2023. No tropical cyclone appears in more than one split, making the test set both temporally separated and storm-disjoint from training and validation. This protocol corresponds to the leakage-free split reported in Table 2.

Table 3: Boundary references under the blind window.

Condition	Position (km)↓	Intensity (MAE)↓	
	ADE	Press (hPa)	Wind (m/s)
Idealized reanalysis w/ SAT	93.34	5.56	3.95
Idealized reanalysis w/o SAT	94.33	5.43	3.87
Delayed reanalysis only	378.83	10.87	6.50

Note. Reanalysis is idealized as real-time available, whereas SAT is genuinely real-time available; thus, the first two rows serve as upper-bound references and delayed reanalysis only serves as a lower-bound reference.

A key design choice is the inclusion of four satellite modalities: visible (VIS), infrared (IR), water vapor (WV), and passive microwave (PMW). These modalities are chosen because they occupy different positions in the trade-off between temporal availability, observation quality, and structural complementarity. IR and WV provide the temporally homogeneous backbone of the observation stream, whereas VIS and PMW provide complementary but more heterogeneous information. In particular, IR imagery offers a stable, day-and-night view of cloud-top organization and convective structure, while WV imagery is especially informative about the surrounding moisture field and upper-level environmental flow. VIS imagery contributes fine-scale cloud morphology and eye-region texture, but its utility is strongly conditioned on illumination and image quality. In terms of temporal coverage, IR, WV, and VIS span 1980–2023 in our dataset, whereas PMW is available from 1998–2023. PMW nevertheless provides complementary information about precipitation and

inner-core storm structure that is difficult to obtain from VIS, IR, and WV alone. Taken together, these four modalities form a physically complementary and operationally meaningful observation set for blind-window recovery.

This multimodal design reflects the structured nature of the blind-window problem. The missing information is specifically recent environmental analysis, while the available realtime observations differ in frequency, noise, and physical sensitivity. The benchmark therefore tests whether complementary realtime observations can reduce this causal information gap without relying on hindsight access to unavailable fields.

Figure 2 and Table 2 summarize the dataset and benchmark design. The dataset couples delayed environmental history, multimodal auxiliary observations, and downstream TC targets under the oracle/full-analysis, delayed-only, and realtime-bridge protocols. Compared with representative TC and weather forecasting datasets, its defining feature is the joint presence of a realtime-aware protocol, multimodal observations, cross-scale benchmarking, forecast utility, and leakage-free evaluation, which enables direct assessment of blind-window recovery rather than hindsight evaluation.

Table 3 illustrates the upper- and lower-bound references for this protocol using results from a single random seed. Idealized analysis-available settings serve as optimistic upper bounds, while delayed reanalysis without a real-time satellite bridge gives a lower-bound reference. Although our default blind window assumes only a conservative 24-hour analysis delay, most existing TC learning benchmarks do not explicitly model this data-availability delay, which can lead to optimistic estimates of deployable forecast skill.

3 Method and Main Results

3.1 From Theory to Architecture

The observability argument summarized in Sec. 1 and derived in Appendix A provides a direct blueprint for RECAST-TC. Component-dependent observability suggests that track, pressure, and wind should not share a fixed temporal context, motivating **T**arget-**A**daptive **L**ag **R**outing (TALR) to select target-specific effective lag depths. Blind-window missingness requires the model to distinguish observed environmental tokens from causally unavailable ones. The mixed-observation argument motivates the use of satellite modalities, whose IR, WV, VIS, and PMW channels provide complementary constraints during the blind window.

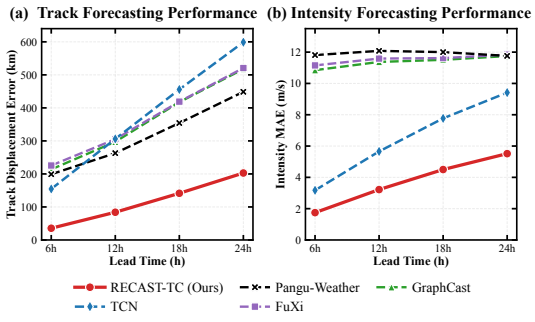

Figure 3: Baseline comparison under the blind-window protocol. RECAST-TC consistently reduces both track displacement error and intensity MAE across lead times.

Noisy stable recovery further suggests reliability-aware multimodal fusion. Together, these observations motivate a state-reconstruction view: the objective is not to reconstruct the full ERA field, but to estimate a forecast-sufficient latent state for future track and intensity prediction.

As illustrated in Fig. 4, RECAST-TC consists of four coupled modules. Given delayed ERA fields, real-time satellite observations, storm history, and an ERA availability mask, the ERA branch first encodes the available environmental tokens into an availability-aware memory H_E. In parallel, the satellite branch embeds IR, WV, VIS, and PMW streams and produces a reliability-weighted satellite memory H_S. TALR then routes information across multiple temporal lags to form a target-specific context C_{TALR}. The Latent Bridge Rollout module reconstructs the missing recent state B_{bridge} through the blind window. The final multitask decoder predicts future position, pressure, and wind:

$$\hat{Y}_{1:H} = D_\theta\Big(\mathrm{Fuse}\big(H_E,\, H_S,\, B_{\mathrm{bridge}},\, C_{\mathrm{TALR}}\big)\Big).$$

Here, H_E encodes delayed but physically structured environmental information, H_S supplies real-time satellite constraints, B_{bridge} represents the recovered blind-window latent state, and C_{TALR}

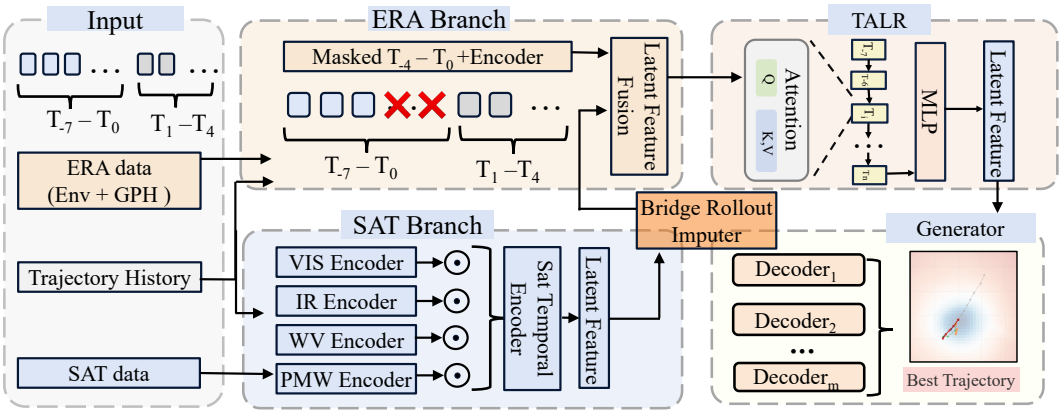

Figure 4: RECAST-TC reconstructs a forecast-sufficient latent state through the blind window. Availability-aware ERA/history encoding and quality-aware satellite encoding feed TALR and Latent Bridge Rollout, which select target-specific context and recover missing recent latent states before multitask prediction of track, pressure, and wind. The pipeline is latent state reconstruction, not generic multimodal fusion.

provides target-adaptive multi-lag context. This formulation separates three roles that are often entangled in standard forecasting models: environmental memory, real-time observation, and latent state recovery.

3.2 Main Results

Table 4 reports the main comparison under the blind-window protocol, with RECAST-TC averaged over 20 random seeds. This setting matches issuance-time availability: recent ERA/GPH analysis frames are hidden, while real-time satellite observations remain available. RECAST-TC achieves the best global performance, with ADE=109.06 $\pm$ 4.28 km, Pressure MAE=5.08 $\pm$ 0.09 hPa, and Wind MAE=3.47 $\pm$ 0.08 m/s, improving track and intensity simultaneously.

Table 4: Global performance comparison. RECAST-TC: mean$\pm$std over 20 seeds. Red numbers denote percentage degradation relative to the RECAST-TC mean. Smaller is better.

Baseline	Position (km)↓	Intensity (MAE)↓	
	ADE	Press (hPa)	Wind (m/s)
GraphCast	361.55 +231.5%	11.63 +128.9%	11.37 +227.7%
FuXi	367.60 +237.1%	11.42 +124.8%	11.55 +232.9%
Pangu-Weather	316.24 +190.0%	13.37 +163.2%	11.91 +243.2%
TCN	378.83 +247.4%	10.87 +114.0%	6.50 +87.3%
HWRF/HWFI	140.11 +28.5%	/	5.12 +47.6%
RECAST-TC (Ours)	**109.06$\pm$4.28**	**5.08$\pm$0.09**	**3.47$\pm$0.08**

Compared with Fixed-Lag Bridge, RECAST-TC reduces track ADE from 124.97 to 109.06 km, pressure MAE from 10.26 to 5.08 hPa, and wind MAE from 6.87 to 3.47 m/s. The larger intensity gain is consistent with the fact that pressure and wind depend more directly on recent storm-scale structure and inner-core evolution, whereas track is more strongly constrained by large-scale steering flow. This supports target-specific lag selection and latent reconstruction of the missing recent state.

The capacity-matched comparison is reported in Table 5. The 20-seed statistics are complemented by a 10,000-replicate storm-level cluster bootstrap (see Supplementary Material). Frozen-latent probes recover steering, motion, and storm-structure information and yield RI AUROC/AUPRC of 0.808/0.194 (see Supplementary Material).

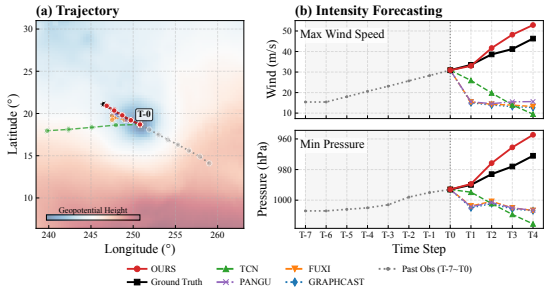

Figure 5: EP case study. RECAST-TC better follows the best-track trajectory and captures the coupled wind–pressure evolution under the blind-window protocol.

RECAST-TC also outperforms global data-driven weather models under the same blind-window protocol. This highlights a mismatch between their usual complete-analysis input assumption and the operational constraint considered here, where forecast skill depends on inferring the storm-relevant latent state from partial, delayed, and heterogeneous observations.

Figure 3 breaks down performance across lead times. RECAST-TC maintains lower track displacement error and lower intensity MAE at

all horizons. The trend suggests better initialization after the blind window: stale or incomplete latent states amplify errors during rollout, while satellite constraints reduce this initial mismatch and stabilize downstream forecasts.

Table 5: Capacity-matched comparisons under the blind-window protocol. Results are mean±std over 20 random seeds. Smaller is better.

Method	Params. (M)	ADE (km)↓	Press (hPa)↓	Wind (m/s)↓
Late Fusion	5.23	114.88 ± 4.07	5.16 ± 0.13	3.53 ± 0.06
Cross-Attention	5.13	116.79 ± 4.25	5.35 ± 0.12	3.63 ± 0.12
Assimilation-Style	5.25	115.25 ± 4.08	5.21 ± 0.05	3.50 ± 0.06
RECAST-TC	**5.19**	$\mathbf{109.06 \pm 4.28}$	$\mathbf{5.08 \pm 0.09}$	$\mathbf{3.47 \pm 0.08}$

blind window.

Table 6: Compact ablation under the blind-window protocol. Results are from a single seed. Red numbers denote percentage degradation relative to RECAST-TC. Smaller is better.

Variant	ADE (km)↓	Press (hPa)↓	Wind (m/s)↓
Model ablations			
w/o delay	141.26 +33.0%	7.42 +44.6%	5.36 +55.0%
naive fusion	132.23 +24.5%	7.67 +49.5%	5.14 +48.6%
w/o bridge rollout	134.22 +26.3%	7.48 +45.8%	5.16 +49.2%
w/o TALR	124.97 +17.6%	10.26 +100.0%	6.87 +98.6%
Data ablations			
w/o SAT	378.83 +256.6%	10.87 +111.9%	6.50 +87.9%
w/o VIS	126.60 +19.2%	6.12 +19.3%	4.26 +23.2%
w/o PMW	123.39 +16.1%	5.68 +10.7%	3.95 +14.2%
w/o VIS+PMW	260.45 +145.1%	14.43 +181.3%	8.95 +158.8%
w/o WV+VIS+PMW	124.33 +17.0%	6.31 +23.0%	4.25 +22.9%
RECAST-TC	**106.25**	**5.13**	**3.46**

The lead-time curves also explain why the gain is especially strong for pressure and wind. Track can remain acceptable when large-scale steering is roughly captured, but intensity is more sensitive to recent convective organization, moisture asymmetry, and inner-core signals. Thus, RECAST-TC recovers forecast-relevant information that is otherwise absent during the

Figure 5 gives a representative East Pacific storm-level example. RECAST-TC follows the best-track trajectory more closely and better preserves the coupled wind–pressure evolution, while baseline forecasts tend to drift after initialization. This supports the state-reconstruction view that the bridge captures information beyond large-scale steering.

We next evaluate two high-impact regimes: tail cases and rapid-intensification (RI) cases. These settings are difficult because average-history extrapolation is weakest, storm motion or structure can be atypical, and intensity may change rapidly within a short time window. As shown in Fig. 6, RECAST-TC achieves the lowest track displacement error and intensity MAE across 6–24 h lead times in both subsets. This indicates that the bridge recovers useful latent context under rare and dynamically sensitive conditions, where IR/WV provide frequent convective and moisture information and VIS/PMW add complementary structural cues when available.

Tables 7 and 8 evaluate basin-level generalization using RECAST-TC results from a single seed. For this run, RECAST-TC achieves the best global track performance, with $106.2/183.5$ km in ADE/FDE, while GraphCast, FuXi, and Pangu-Weather obtain $361.6/517.8$, $367.6/520.8$, and $316.2/448.7$ km. For intensity, RECAST-TC reaches 5.13 hPa in pressure MAE and 3.46 m/s in wind MAE, whereas the foundation-model baselines remain around 11–13 hPa and 11–12 m/s. The consistency across basins, also illustrated by Fig. 7, suggests that the gain is not limited to one metric, target, or ocean basin.

Figure 7 further compares storm evolution with background fields, geopotential height, wave speed, and intensity evolution. Baseline models may capture broad environmental patterns but miss the storm-specific latent state needed for accurate track and intensity prediction. The

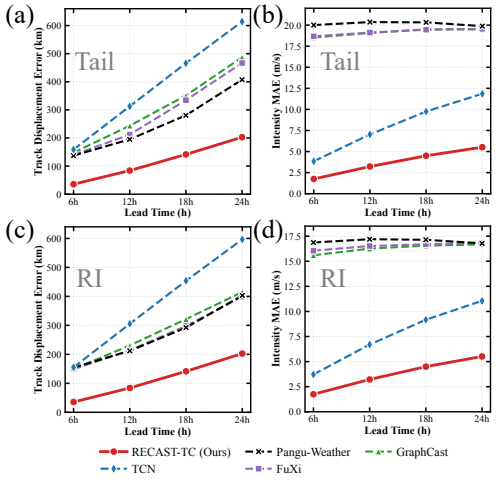

Figure 6: Tail and RI forecasting performance. RECAST-TC consistently achieves lower track displacement errors and intensity MAEs across 6–24 h lead times on both tail and rapid-intensification cases.

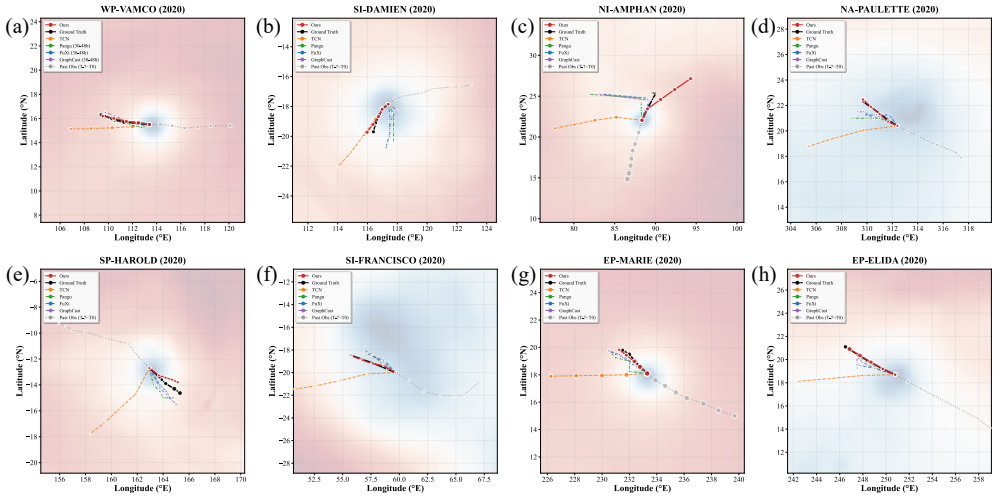

Figure 7: Storm-level case studies across multiple basins. RECAST-TC stays closer to the best-track trajectories, while baseline forecasts tend to drift after initialization under the blind-window protocol.

qualitative cases therefore complement the aggregate tables by showing how lower numerical error corresponds to more coherent storm evolution.

Table 6 verifies the theory-to-method chain. The largest degradation occurs without satellite inputs, where ADE increases from 106.25 to 378.83 km, showing that stale environmental history alone is insufficient under the blind-window protocol. Removing both VIS and PMW also causes a severe drop to 260.45 km, 14.43 hPa, and 8.95 m/s, confirming that less frequent or conditionally available modalities still provide critical complementary information.

The architectural ablations show how these observations should be used. Removing TALR mainly damages intensity, increasing pressure MAE from 5.13 to 10.26 hPa and wind MAE from 3.46 to 6.87 m/s, supporting component-dependent temporal context. Removing Latent Bridge Rollout degrades all targets, with ADE 134.22 km, pressure MAE 7.48 hPa, and wind MAE 5.16 m/s. Naive fusion also performs worse, reaching 132.23 km, 7.67 hPa, and 5.14 m/s, indicating that multimodal information requires reliability-aware fusion.

Table 5 further compares capacity-matched alternatives. All methods share the inputs, data splits, applicable encoders, decoder, losses, optimization settings, training budget, and checkpoint-selection criterion. With comparable model sizes (5.13–5.25M parameters), RECAST-TC achieves the lowest mean errors across all three metrics over 20 random seeds. These gains persist after matching model capacity.

Overall, the results support a consistent mechanism: the blind window removes recent analysis constraints, real-time satellite modalities help restore them, and RECAST-TC improves by combining availability handling, target-specific temporal dependence, latent bridge rollout, and modality reliability into a state-reconstruction framework. This yields better track forecasts, stronger intensity forecasts, robust behavior on high-impact subsets, and consistent generalization across basins.

4 Related Work

Tropical cyclone (TC) forecasting has traditionally relied on dynamical models and statistical–dynamical systems that combine storm history, climatology, persistence, and large-scale environmental predictors such as vertical shear, moisture, and sea-surface temperature [DeMaria et al., 2005, Kaplan et al., 2010, Tallapragada et al., 2014, Cangialosi et al., 2020, Gall et al., 2013]. These methods remain important in operations, but TC intensity, rapid intensification, and structural change remain challenging because they depend on multi-scale processes that are only partially observed and imperfectly represented in coarse environmental fields [Cangialosi et al., 2020, Wang and Wu, 2004, Rogers et al., 2013, Emanuel, 1987, Kieper and Jiang, 2012].

Recent machine learning methods improve TC prediction by learning from best-track records, reanalysis fields, satellite imagery, and other heterogeneous predictors [Chen et al., 2020, Wimmers

Table 7: Track errors across basins under the effective 30–48h blind-window protocol. ADE/FDE in km. RECAST-TC results are from a single seed.

Model	EP	NA	NI	SI	SP	WP	Global
GraphCast	503.5/777.2	500.2/689.5	73.3/128.2	213.6/295.9	181.5/240.4	333.9/493.0	361.6/517.8
FuXi	495.9/767.7	477.4/641.0	63.0/110.0	172.6/220.2	124.6/172.3	350.0/522.8	367.6/520.8
Pangu-Weather	431.9/644.5	458.4/623.8	113.0/146.6	149.3/192.0	140.2/170.0	292.6/433.6	316.2/448.7
HWRF/HWFI	129.1/155.6	151.2/190.9	/	/	/	/	140.1/173.2
RECAST-TC (Ours)	**67.9/118.5**	**122.0/217.2**	**100.4/167.0**	**107.4/181.5**	**138.7/234.0**	**113.7/194.1**	**106.2/183.5**

Table 8: Intensity errors across basins under the effective 30–48h blind-window protocol. Each entry reports Pressure MAE (hPa) / Wind MAE (m/s). RECAST-TC results are from a single seed.

Model	EP	NA	NI	SI	SP	WP	Global
GraphCast	9.18/12.89	11.09/12.59	11.92/13.43	9.14/8.25	20.23/17.47	13.96/10.87	11.63/11.37
FuXi	9.01/12.76	10.72/12.44	13.36/15.70	8.43/7.51	27.74/21.94	13.22/10.65	11.42/11.55
Pangu-Weather	12.07/14.00	12.84/12.92	12.48/13.41	11.65/9.10	13.43/11.01	15.98/11.51	13.37/11.91
HWRF/HWFI	//5.10	//5.13	/	/	/	/	//5.12
RECAST-TC (Ours)	**4.04/3.21**	**4.62/3.40**	**5.64/4.57**	**5.81/3.76**	**6.40/4.07**	**5.65/3.25**	**5.13/3.46**

et al., 2019, Su et al., 2020, Huang et al., 2025], while global data-driven weather models trained on large reanalysis archives have shown strong medium-range forecasting skill [Rasp et al., 2024, Price et al., 2025, Bodnar et al., 2025, Allen et al., 2025]. However, many such approaches are evaluated with recent reanalysis or analysis fields available at initialization, even though these products often rely on delayed assimilation and retrospective processing. This can introduce a hidden hindsight advantage and overestimate deployable realtime forecast skill [Bauer et al., 2015, Rasp et al., 2024].

Data assimilation combines prior model states with heterogeneous observations in operational weather prediction [Evensen, 2003, Bannister, 2017, Carrassi et al., 2018], and operational TC prediction also relies on physics-based NWP systems such as WRF, COAMPS-TC, and HWRF [Skamarock et al., 2019, Doyle et al., 2014, Alaka Jr et al., 2024]. RECAST-TC is related but deliberately reduced in scope, learning a storm-centered latent reconstruction from delayed environmental analyses and real-time satellite observations rather than performing full-field assimilation or evolving a prognostic atmospheric state. It is therefore complementary to operational DA/NWP pipelines and serves as a benchmark for reduced-state reconstruction under realistic input availability.

Satellite observations provide a natural realtime source of storm-state information. Geostationary visible, near-infrared, infrared, and water-vapor imagery supports monitoring of cloud organization, upper-tropospheric moisture, and storm-scale flow patterns [Schmit et al., 2008, Velden et al., 1997, Soden and Bretherton, 1993], infrared and Dvorak-type methods support satellite-based intensity estimation [Velden et al., 2006, Olander and Velden, 2019], and passive microwave or precipitation-sensitive sensors provide complementary information about convective bursts, eyewall structure, rainbands, and inner-core precipitation [Kummerow et al., 1998, Cecil and Zipser, 1999, Wimmers and Velden, 2007, Razin et al., 2023]. Recent reviews further emphasize the role of current and emerging sensors for observing TC fine structure, wind fields, and temporal evolution [Ricciardulli et al., 2023, Vinour et al., 2021]. Prior studies often treat these modalities as auxiliary features, whereas our work uses them to address a specific operational information gap in which recent reanalysis fields are unavailable inside the blind window. We therefore formulate realtime TC forecasting as state reconstruction from delayed, partial, noisy, and asynchronous observations. Rather than imputing the full missing reanalysis field, RECAST-TC reconstructs a forecast-sufficient latent storm state through availability-aware environmental encoding, quality-aware satellite fusion, adaptive lag selection, and latent bridge rollout.

5 Future Work and Limitations

RECAST-TC relies on realtime satellite observations to bridge missing recent environmental context, but these observations can be irregular, delayed, missing, or degraded by illumination, viewing geometry, sensor noise, and retrieval uncertainty. Although reliability-aware fusion and modality ablations partly address this issue, future operational systems should model sensor latency, quality flags, and missingness more explicitly. Our benchmark also differs from fully operational forecasting systems, where data delivery, preprocessing, and analysis cycles vary across centers. Nevertheless,

because many TC forecasting studies implicitly assume ERA5 or GFS-derived analysis fields are available at issuance time, explicitly modeling this availability gap remains necessary for realistic datasets, benchmarks, and multimodal forecasting methods.

RECAST-TC reconstructs a forecast-sufficient latent state rather than a physically complete atmospheric state, which improves blind-window prediction but limits direct interpretability. Future work should connect the recovered latent representation to physical diagnostics such as steering flow, vertical shear, moisture structure, inner-core organization, and ocean–atmosphere interaction. Longer lead times may also require explicit environmental rollout, ensemble prediction, or physics-guided constraints. Because RECAST-TC operates at the reduced storm-state level and does not enforce full conservation laws, it should be viewed as complementary to operational data assimilation and physics-based NWP systems. Finally, the current model produces deterministic forecasts, while operational TC forecasting requires calibrated uncertainty, especially for tail events and rapid intensification. Future versions should provide probabilistic track and intensity forecasts and be tested prospectively on realtime data streams, evolving satellite systems, and alternative operational protocols.

6 Conclusion

This work identifies a hidden evaluation mismatch in tropical cyclone forecasting: recent reanalysis fields are often used as inputs despite being unavailable at realtime issuance. We formalize this as forecasting through a reanalysis blind window and introduce a realtime-faithful benchmark with full-history, delayed-only, and realtime-bridge regimes.

We propose RECAST-TC, a timeliness-aware multimodal reconstruction framework that recovers a forecast-sufficient latent storm state from delayed environmental analyses and real-time satellite observations. By combining availability-aware environmental encoding, quality-aware satellite fusion, target-specific lag selection, and latent bridge rollout, RECAST-TC improves both track and intensity forecasting across global, basin-level, lead-time, tail-case, rapid-intensification, qualitative, and ablation evaluations.

Overall, our results show that real-time satellite observations can compensate for storm-relevant state components affected by reanalysis latency, framing operational TC forecasting as an information-constrained state-reconstruction problem under realistic real-time constraints. More broadly, RECAST-TC offers a transferable framework for scientific forecasting with asynchronous fast and slow data streams, integrating delayed high-fidelity context with timely partial observations to infer forecast-sufficient latent states.

Acknowledgments and Disclosure of Funding

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM908), the National Natural Science Foundation of China (Grant Nos. T2600685 and 92565111), and the Xiaomi Foundation. We also gratefully acknowledge support from the Shanghai Innovation Institute.

References

Ghassan J Alaka Jr, Jason A Sippel, Zhan Zhang, Hyun-Sook Kim, Frank D Marks, Vijay Tallapragada, Avichal Mehra, Xuejin Zhang, Aaron Poyer, and Sundararaman G Gopalakrishnan. Lifetime performance of the operational hurricane weather research and forecasting model (HWRF) for north Atlantic tropical cyclones. *Bulletin of the American Meteorological Society*, 105(6): E932–E961, 2024.

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, 2025.

Ross N Bannister. A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):607–633, 2017.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

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(7970):533–538, 2023.

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(8065):1180–1187, 2025.

John P Cangialosi, Eric Blake, Mark DeMaria, Andrew Penny, Andrew Latto, Edward Rappaport, and Vijay Tallapragada. Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Weather and Forecasting*, 35(5):1913–1922, 2020.

Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.

Daniel J Cecil and Edward J Zipser. Relationships between tropical cyclone intensity and satellite-based indicators of inner core convection: 85-GHz ice-scattering signature and lightning. *Monthly weather review*, 127(1):103–123, 1999.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023a.

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

Rui Chen, Weimin Zhang, and Xiang Wang. Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere*, 11(7):676, 2020.

Mark DeMaria, Michelle Mainelli, Lynn K Shay, John A Knaff, and John Kaplan. Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Weather and Forecasting*, 20(4):531–543, 2005.

James D Doyle, Richard M Hodur, SUE Chen, YI Jin, Jonathan R Moskaitis, Shouping Wang, Eric A Hendricks, HAO Jin, and Travis A Smith. Tropical cyclone prediction using COAMPS-TC. *Oceanography*, 27(3):104–115, 2014.

Kerry A Emanuel. The dependence of hurricane intensity on climate. *Nature*, 326(6112):483–485, 1987.

Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

Robert Gall, James Franklin, Frank Marks, Edward N Rappaport, and Frederick Toepfer. The hurricane forecast improvement project. *Bulletin of the American Meteorological Society*, 94(3): 329–343, 2013.

Sophie Giffard-Roisin, Mo Yang, Guillaume Charpiat, Christina Kumler Bonfanti, Balázs Kégl, and Claire Monteleoni. Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Frontiers in big Data*, 3:1, 2020.

Franz Hamilton, Tyrus Berry, and Timothy Sauer. Kalman-Takens filtering in the presence of dynamical noise. *The European Physical Journal Special Topics*, 226(15):3239–3250, 2017.

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.

Cheng Huang, Pan Mu, Jinglin Zhang, Sixian Chan, Shiqi Zhang, Hanting Yan, Shengyong Chen, and Cong Bai. Benchmark dataset and deep learning method for global tropical cyclone forecasting. *Nature Communications*, 16(1):5923, 2025.

John Kaplan, Mark DeMaria, and John A Knaff. A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Weather and forecasting*, 25(1):220–241, 2010.

Margaret E Kieper and Haiyan Jiang. Predicting tropical cyclone rapid intensification using the 37 ghz ring pattern identified from passive microwave measurements. *Geophysical Research Letters*, 39(13), 2012.

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.

Christian Kummerow, William Barnes, Toshiaki Kozu, James Shiue, and Joanne Simpson. The tropical rainfall measuring mission (TRMM) sensor package. *Journal of atmospheric and oceanic technology*, 15(3):809–817, 1998.

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.

Timothy L Olander and Christopher S Velden. The advanced Dvorak technique (ADT) for estimating tropical cyclone intensity: Update and new capabilities. *Weather and Forecasting*, 34(4):905–922, 2019.

Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Muhammad Naufal Razin, Christopher J Slocum, John A Knaff, Paula J Brown, and Michael M Bell. Tropical cyclone precipitation, infrared, microwave, and environmental dataset (TC PRIMED). *Bulletin of the American Meteorological Society*, 104(11):E1980–E1998, 2023.

Lucrezia Ricciardulli, Brian Howell, Christopher R Jackson, Jeff Hawkins, Joe Courtney, Ad Stoffelen, Sebastian Langlade, Chris Fogarty, Alexis Mouche, William Blackwell, et al. Remote sensing and analysis of tropical cyclones: Current and emerging satellite sensors. *Tropical Cyclone Research and Review*, 12(4):267–293, 2023.

Robert Rogers, Sim Aberson, Altug Aksoy, Bachir Annane, Michael Black, Joseph Cione, Neal Dorst, Jason Dunion, John Gamache, Stan Goldenberg, et al. NOAA's hurricane intensity forecasting experiment: A progress report. *Bulletin of the American Meteorological Society*, 94(6):859–882, 2013.

Mario Rüttgers, Sangseung Lee, Soohwan Jeon, and Donghyun You. Prediction of a typhoon track using a generative adversarial network and satellite images. *Scientific reports*, 9(1):6057, 2019.

Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3): 579–616, 1991.

Timothy J Schmit, Jun Li, Jinlong Li, Wayne F Feltz, James J Gurka, Mitchell D Goldberg, and Kevin J Schrab. The GOES-R Advanced Baseline Imager and the continuation of current sounder products. *Journal of Applied Meteorology and Climatology*, 47(10):2696–2711, 2008.

William C Skamarock, Joseph B Klemp, Jimy Dudhia, David O Gill, Zhiquan Liu, Judith Berner, Wei Wang, Jordan G Powers, Michael G Duda, Dale M Barker, et al. A description of the advanced research WRF version 4. *NCAR tech. note ncar/tn-556+ str*, 145(10.5065), 2019.

Brian J Soden and Francis P Bretherton. Upper tropospheric relative humidity from the GOES 6.7 μm channel: Method and climatology for July 1987. *Journal of Geophysical Research: Atmospheres*, 98(D9):16669–16688, 1993.

Jaroslav Stark, David S Broomhead, Michael Evan Davies, and J Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods & Applications*, 30(9): 5303–5314, 1997.

Jaroslav Stark, David S Broomhead, Mike E Davies, and John Huke. Delay embeddings for forced systems. II. Stochastic forcing. *Journal of Nonlinear Science*, 13(6):519–577, 2003.

Hui Su, Longtao Wu, Jonathan H Jiang, Raksha Pai, Alex Liu, Albert J Zhai, Peyman Tavallali, and Mark DeMaria. Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophysical Research Letters*, 47(17): e2020GL089102, 2020.

Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381, Berlin, Heidelberg, 1981. Springer Berlin Heidelberg. ISBN 978-3-540-38945-3.

Vijay Tallapragada, Ligia Bernardet, Mrinal K Biswas, Sundararaman Gopalakrishnan, Young Kwon, Qingfu Liu, Tim Marchok, Dmitry Sheinin, Mingjing Tong, Samuel Trahan, et al. Hurricane Weather Research and Forecasting (HWRF) model: 2013 scientific documentation. *HWRF Development Testbed Center Tech. Rep*, 99, 2014.

Christopher Velden, Bruce Harper, Frank Wells, John L Beven, Ray Zehr, Timothy Olander, Max Mayfield, Charles "CHIP" Guard, Mark Lander, Roger Edson, et al. The dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bulletin of the American Meteorological Society*, 87(9):1195–1210, 2006.

Christopher S Velden, Christopher M Hayden, Steven J W Nieman, W Paul Menzel, Steven Wanzong, and James S Goerss. Upper-tropospheric winds derived from geostationary satellite water vapor observations. *Bulletin of the American Meteorological Society*, 78(2):173–196, 1997.

Léo Vinour, Swen Jullien, Alexis Mouche, Clément Combot, and Morgan Mangeas. Observations of tropical cyclone inner-core fine-scale structure, and its link to intensity variations. *Journal of the Atmospheric Sciences*, 78(11):3651–3671, 2021.

Liang Wang, Bingcheng Wan, Shaohui Zhou, Haofei Sun, and Zhiqiu Gao. Forecasting tropical cyclone tracks in the northwestern pacific based on a deep-learning model. *Geoscientific Model Development*, 16(8):2167–2179, 2023.

Yuqing Wang and C-C Wu. Current understanding of tropical cyclone structure and intensity changes–a review. *Meteorology and Atmospheric Physics*, 87(4):257–278, 2004.

Anthony Wimmers, Christopher Velden, and Joshua H Cossuth. Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Monthly Weather Review*, 147 (6):2261–2282, 2019.

Anthony J Wimmers and Christopher S Velden. MIMIC: A new approach to visualizing satellite microwave imagery of tropical cyclones. *Bulletin of the American Meteorological Society*, 88(8): 1187–1196, 2007.

A An Illustrative Mixed-Observation Derivation

This appendix provides a minimal derivation illustrating why time-lagged observations are important for state recovery in numerical dynamical systems, and why a real-time auxiliary modality can compensate for missing recent analysis frames. The example is not intended as a complete model of atmospheric dynamics. Instead, it isolates a generic mechanism that broadly arises in discretized nonlinear systems: different latent components may require different temporal observation depths; removing a recent frame can destroy identifiability; and an additional real-time observation can restore identifiability when it provides an independent measurement direction.

We use a coarsened linear dynamical system as an analytically tractable local example. For nonlinear numerical systems, the same argument should be interpreted locally after linearization around a trajectory, where lagged observations and mixed observation channels determine whether the forecast-relevant state is sufficiently constrained.

A.1 Latent Dynamics and Observation Channels

As a simple analytically tractable instance of a time-lagged spatial dynamical system, we consider one local Fourier mode of a 2×2-coarsened two-dimensional heat-equation system. Let

$$\hat{z}^n = (\hat{X}^n, \hat{B}^n, \hat{C}^n, \hat{D}^n)^\top \in \mathbb{C}^4$$

denote the local latent state at discrete time n. Here, $\hat{X}^n$ is the observed coarse Fourier coefficient, while $\hat{B}^n, \hat{C}^n, \hat{D}^n$ represent unresolved sub-grid Fourier components inside the same coarse cell. The state evolves according to

$$\hat{z}^{n+1} = A\hat{z}^n + \omega^n,$$

where ω^n denotes process noise. The transition matrix is

$$A = \begin{bmatrix} a & rp & rq & 0 \\ r\bar{p} & a & 0 & rq \\ r\bar{q} & 0 & a & rp \\ 0 & r\bar{q} & r\bar{p} & a \end{bmatrix},$$

with

$$a = 1 - 4r, \qquad p = 1 + e^{-i\theta}, \qquad q = 1 + e^{-i\phi}.$$

Here, r is the diffusion parameter, while θ and ϕ are Fourier phases in the two spatial directions. The delayed ERA-like observation channel is

$$y_E^n = C_E\hat{z}^n + \varepsilon_E^n, \qquad C_E = (1, 0, 0, 0),$$

so that

$$y_E^n = \hat{X}^n + \varepsilon_E^n.$$

Thus, the ERA-like channel observes only the coarse component $\hat{X}^n$.

We also consider a real-time SAT-like observation channel. In the full latent space, this channel may be written abstractly as

$$y_S^n = C_S\hat{z}^n + \varepsilon_S^n.$$

In the main derivation, we only use its effective local direction on the unresolved edge subspace after linearization, propagation to the common reference time, and subtraction of components that are already identifiable from delayed observations. This effective direction is written as

$$u\hat{B}^{-3} + v\hat{C}^{-3} = s_S,$$

where u and v describe the satellite channel's local sensitivity to the two unresolved edge variables.

The local delayed window is

$$\hat{X}^{-3}, \quad \hat{X}^{-2}, \quad \hat{X}^{-1}, \quad \hat{X}^0.$$

The frame $\hat{X}^0$ is the most recent delayed coarse observation before forecast initialization. We compare three regimes:

full history, delayed-only blind window, real-time bridge.

A.2 Full-History Recoverability

In the noiseless case, starting from

$$\hat{z}^{-3} = (\hat{X}^{-3}, \hat{B}^{-3}, \hat{C}^{-3}, \hat{D}^{-3})^{\top},$$

the subsequent coarse observations satisfy

$$\hat{X}^{-2} = C_E A \hat{z}^{-3}, \qquad \hat{X}^{-1} = C_E A^2 \hat{z}^{-3}, \qquad \hat{X}^{0} = C_E A^3 \hat{z}^{-3}.$$

The first delayed observation is

$$\hat{X}^{-2} = a\hat{X}^{-3} + rp\,\hat{B}^{-3} + rq\,\hat{C}^{-3}.$$

The second delayed observation is

$$\hat{X}^{-1} = (a^2 + r^2|p|^2 + r^2|q|^2)\hat{X}^{-3} + 2arp\,\hat{B}^{-3} + 2arq\,\hat{C}^{-3} + 2r^2pq\,\hat{D}^{-3}.$$

The coefficients of $\hat{B}^{-3}$ and $\hat{C}^{-3}$ in the second equation are exactly $2a$ times their coefficients in the first equation. Therefore,

$$\hat{X}^{-1} - 2a\hat{X}^{-2} = (r^2|p|^2 + r^2|q|^2 - a^2)\hat{X}^{-3} + 2r^2pq\,\hat{D}^{-3}.$$

Solving for $\hat{D}^{-3}$, we obtain

$$\hat{D}^{-3} = \frac{\hat{X}^{-1} - 2a\hat{X}^{-2} - (r^2|p|^2 + r^2|q|^2 - a^2)\hat{X}^{-3}}{2r^2pq}.$$

Hence, provided $r^2pq \neq 0$, the corner-like hidden variable $\hat{D}^{-3}$ is identifiable from the first three delayed coarse frames

$$(\hat{X}^{-3}, \hat{X}^{-2}, \hat{X}^{-1}).$$

However, the same three frames do not identify the edge variables $(\hat{B}^{-3}, \hat{C}^{-3})$. From the first delayed observation, they satisfy only one linear constraint:

$$rp\,\hat{B}^{-3} + rq\,\hat{C}^{-3} = s_1, \qquad s_1 := \hat{X}^{-2} - a\hat{X}^{-3}.$$

Thus, $(\hat{B}^{-3}, \hat{C}^{-3})$ remain underdetermined after three frames.

The fourth delayed frame supplies the missing independent constraint. Direct expansion of $C_E A^3 \hat{z}^{-3}$ gives

$$\hat{X}^{0} = \alpha_X \hat{X}^{-3} + \beta_B \hat{B}^{-3} + \beta_C \hat{C}^{-3} + \alpha_D \hat{D}^{-3},$$

where

$$\beta_B = rp\big(3a^2 + r^2|p|^2 + 3r^2|q|^2\big),$$

and

$$\beta_C = rq\big(3a^2 + 3r^2|p|^2 + r^2|q|^2\big).$$

The coefficients α_X and α_D multiply components that are already known or already identifiable at this point. Define

$$s_0 := \hat{X}^{0} - \alpha_X \hat{X}^{-3} - \alpha_D \hat{D}^{-3}.$$

Then the fourth frame provides

$$\beta_B \hat{B}^{-3} + \beta_C \hat{C}^{-3} = s_0.$$

Combining this equation with

$$rp\,\hat{B}^{-3} + rq\,\hat{C}^{-3} = s_1$$

yields the two-equation system

$$\begin{bmatrix} rp & rq \\ \beta_B & \beta_C \end{bmatrix} \begin{bmatrix} \hat{B}^{-3} \\ \hat{C}^{-3} \end{bmatrix} = \begin{bmatrix} s_1 \\ s_0 \end{bmatrix}.$$

Its determinant is

$$\Delta = rp\,\beta_C - rq\,\beta_B.$$

Substituting the expressions for β_B and β_C, we get

$$\Delta = rp \cdot rq\big(3a^2 + 3r^2|p|^2 + r^2|q|^2\big) - rq \cdot rp\big(3a^2 + r^2|p|^2 + 3r^2|q|^2\big),$$

and therefore

$$\Delta = 2r^4 pq\big(|p|^2 - |q|^2\big).$$

Hence, for generic non-degenerate modes satisfying

$$pq \neq 0, \qquad |p|^2 \neq |q|^2,$$

we have $\Delta \neq 0$, and the edge variables $(\hat{B}^{-3}, \hat{C}^{-3})$ are identifiable from the full delayed window.

This illustrates component-dependent recoverability in a time-lagged observation system. The corner-like component $\hat{D}^{-3}$ is identifiable from three coarse frames, whereas the edge variables $(\hat{B}^{-3}, \hat{C}^{-3})$ require the fourth and most recent frame $\hat{X}^0$.

A.3 Blind-Window Failure

Now suppose the recent frame $\hat{X}^0$ lies inside the blind window and is therefore unavailable. Then the second edge constraint

$$\beta_B \hat{B}^{-3} + \beta_C \hat{C}^{-3} = s_0$$

is lost. The edge variables are constrained only by

$$rp\,\hat{B}^{-3} + rq\,\hat{C}^{-3} = s_1,$$

or equivalently,

$$\begin{bmatrix} rp & rq \end{bmatrix} \begin{bmatrix} \hat{B}^{-3} \\ \hat{C}^{-3} \end{bmatrix} = s_1.$$

This system has rank one. Therefore, there are infinitely many pairs $(\hat{B}^{-3}, \hat{C}^{-3})$ consistent with the same delayed coarse history. The blind window thus destroys identifiability by removing the particular recent observation that closed the observability gap.

A.4 Real-Time Rank Restoration

Suppose a real-time SAT-like observation is available inside the blind window. After local linearization and projection onto the unresolved edge subspace, it provides the effective constraint

$$u\hat{B}^{-3} + v\hat{C}^{-3} = s_S.$$

Combining this with the remaining delayed ERA-like constraint yields

$$\begin{bmatrix} rp & rq \\ u & v \end{bmatrix} \begin{bmatrix} \hat{B}^{-3} \\ \hat{C}^{-3} \end{bmatrix} = \begin{bmatrix} s_1 \\ s_S \end{bmatrix}.$$

Let

$$M_{\mathrm{mix}} = \begin{bmatrix} rp & rq \\ u & v \end{bmatrix}.$$

The determinant of this mixed observation matrix is

$$\Delta_{\mathrm{mix}} = \det M_{\mathrm{mix}} = rp\,v - rq\,u = r(pv - qu).$$

Therefore, if

$$\Delta_{\mathrm{mix}} \neq 0,$$

then M_{mix} has full rank and the unresolved edge variables are identifiable. The explicit solution is

$$\hat{B}^{-3} = \frac{vs_1 - rq\,s_S}{rp\,v - rq\,u},$$

and

$$\hat{C}^{-3} = \frac{rp\,s_S - us_1}{rp\,v - rq\,u}.$$

Thus, the real-time observation can help recover the unresolved variables not by reconstructing the missing coarse frame $\hat{X}^0$, but by supplying an independent observation direction. This is the algebraic mechanism behind the real-time bridge in Fig. 1(e).

A.5 Noisy Stable Recovery

Finally, we return to the noisy setting. The delayed and satellite observations contain both process noise and measurement noise:

$$\hat{z}^{n+1} = A\hat{z}^n + \omega^n, \qquad y_E^n = \hat{X}^n + \varepsilon_E^n, \qquad y_S^n = C_S\hat{z}^n + \varepsilon_S^n.$$

After collecting the two effective constraints on the edge variables, the mixed system can be written as

$$\tilde{b} = M_{\mathrm{mix}}\theta + \zeta,$$

where

$$\theta = \begin{bmatrix} \hat{B}^{-3} \\ \hat{C}^{-3} \end{bmatrix}, \qquad M_{\mathrm{mix}} = \begin{bmatrix} rp & rq \\ u & v \end{bmatrix}.$$

Here, $\tilde{b}$ denotes the noisy observed right-hand side, and ζ absorbs both measurement noise and propagated process noise. The inverse estimate is

$$\hat{\theta} = M_{\mathrm{mix}}^{-1}\tilde{b},$$

so

$$\hat{\theta} - \theta = M_{\mathrm{mix}}^{-1}\zeta.$$

Taking norms gives

$$\|\hat{\theta} - \theta\| \le \|M_{\mathrm{mix}}^{-1}\|\,\|\zeta\| = \frac{\|\zeta\|}{\sigma_{\min}(M_{\mathrm{mix}})}.$$

Thus, exact recovery in the noiseless case becomes stable recovery in the noisy case. The bridge is robust when the real-time satellite direction is well-conditioned relative to the delayed ERA-like direction. Conversely, if the satellite direction is nearly collinear with the delayed direction, then $\sigma_{\min}(M_{\mathrm{mix}})$ becomes small and the reconstruction becomes noise-sensitive.

B Implementation Details and Hyperparameters

This section gives the implementation details of RECAST-TC used in our experiments. The overall training procedure is summarized in Algorithm 1. Each sample contains an 8-step historical window and a 4-step forecast window. To simulate realtime issuance, we expose only the earliest T_E environmental steps to the model and mask the remaining $8 - T_E$ recent steps inside the blind window. In our default setting, $T_E = 4$. The masked environmental input is used by the student branch, while the complete environmental sequence is kept as a teacher signal for latent bridge supervision. The model therefore learns to recover forecast-sufficient latent tokens rather than directly reconstructing the missing ERA/GPH fields in pixel space.

The trajectory state contains four variables: longitude, latitude, pressure, and wind. The trajectory encoder and decoder are one-layer LSTMs. The default embedding dimension is 32, the generator encoder and decoder hidden dimensions are both 64, and the discriminator hidden dimension is 128. The generator predicts relative future states and converts them to absolute forecasts through autoregressive decoding. When multiple samples are enabled, the model generates $K = 6$ candidates and applies a best-of-K supervised loss. When soft-MoE decoding is used, the effective number of samples is set to one and the prediction is obtained by a soft mixture over expert outputs.

The satellite branch processes four realtime modalities: VIS, IR, WV, and PMW. Each modality is encoded by a lightweight frame encoder with three convolutional blocks, channel widths 16, 32, and 64, followed by adaptive average pooling and a linear projection to a 64-dimensional feature. VIS is additionally equipped with an implicit completion module. This module predicts a completed VIS feature from IR, WV, PMW, storm-history context, environmental context, and VIS metadata, and then combines the observed and completed VIS features using a learned reliability score. For modality fusion, each channel receives a quality-aware score computed from its feature and side information, including quality flags, VIS availability, day/night flag, and solar zenith information. The final satellite feature is the softmax-weighted sum of the enabled modality features.

Temporal satellite context is encoded by a GRU. When satellite TALR is enabled, the GRU hidden size is 256, the attention hidden size is 128, and the final satellite context combines recurrent and attention-pooled features. A future-token decoder then autoregressively generates 4 future satellite

Algorithm 1 RECAST-TC training under the blind-window protocol

Require: Storm history $X_{1:T}$, future target $Y_{1:H}$, environmental fields $E_{1:T}$, satellite observations $S_{1:T}$, ERA mask $M_{1:T}$

Require: History length $T = 8$, forecast horizon $H = 4$, visible ERA length T_E

1: Mask the blind-window ERA input:

$$\tilde{E}_{1:T} = \mathrm{Mask}(E_{1:T}, T_E).$$

2: Encode delayed environmental information:

$$Z_E = \mathrm{EnvEnc}(\tilde{E}_{1:T}, M_{1:T}).$$

3: Encode realtime satellite modalities and fuse them with quality-aware weights:

$$Z_S = \sum_{m \in \{\mathrm{VIS,IR,WV,PMW}\}} \alpha_m \mathrm{SatEnc}_m(S^m_{1:T}).$$

4: Reconstruct blind-window latent tokens:

$$\hat{Z}_E = M \odot Z_E + (1 - M) \odot \mathrm{Bridge}(Z_E, Z_S, M).$$

5: Fuse storm history, reconstructed environmental tokens, and satellite tokens:

$$C = \mathrm{Fuse}(X_{1:T}, \hat{Z}_E, Z_S).$$

6: Decode future track and intensity:

$$\hat{Y}_{1:H} = \mathrm{Decoder}(C).$$

7: Optimize the forecasting loss with auxiliary bridge, VIS-completion, VQ, and adversarial losses:

$$\mathcal{L} = \mathcal{L}_{\mathrm{forecast}} + \lambda_{\mathrm{bridge}}\mathcal{L}_{\mathrm{bridge}} + \lambda_{\mathrm{vis}}\mathcal{L}_{\mathrm{vis}} + \lambda_{\mathrm{vq}}\mathcal{L}_{\mathrm{vq}} + \mathcal{L}_{\mathrm{adv}}.$$

8: Update model parameters with AdamW.

tokens using a GRUCell. These observed and predicted satellite tokens are projected to the same token dimension as the trajectory and environmental representations.

The latent bridge rollout module operates on 64-dimensional environmental and satellite tokens. It first projects satellite tokens to the environmental token dimension and concatenates environmental tokens, satellite tokens, and the binary availability mask. The concatenated representation is mapped to 64 dimensions and passed through a two-layer Transformer encoder with four attention heads. The bridge predicts latent tokens for all historical steps, but the completed token sequence keeps observed environmental tokens unchanged and replaces only unavailable blind-window tokens with predicted tokens. A GRU summary over the completed token sequence provides the bridge-level global feature used by the downstream decoder.

The environmental image branch uses a compact 3D U-Net. The input tensor is organized as (B, C, T, H, W). The encoder uses channel widths 16, 32, 64, 128, and 128, while the decoder upsamples with skip connections and channel widths 64, 32, 16, and 16. All multi-scale decoder features are resized to a common spatiotemporal size and fused along the channel dimension, which makes the network compatible with different effective analysis lengths such as 4, 6, and 8.

The optional vector-quantization module uses an EMA codebook with 256 entries and 64-dimensional codes. The commitment weight is 0.25 and the EMA decay is 0.99. This module regularizes the recovered latent representation and encourages a discrete, stable bridge state. During training, the auxiliary outputs include bridge reconstruction loss, VIS completion loss, VQ codebook loss, and VQ commitment loss.

The generator objective combines the best-of-K forecasting loss, adversarial loss, image reconstruction loss, and optional auxiliary regularization terms:

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{forecast}} + \mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\mathrm{img}} + \lambda_{\mathrm{bridge}}\mathcal{L}_{\mathrm{bridge}} + \lambda_{\mathrm{vis}}\mathcal{L}_{\mathrm{vis}} + \lambda_{\mathrm{code}}\mathcal{L}_{\mathrm{code}} + \lambda_{\mathrm{commit}}\mathcal{L}_{\mathrm{commit}}.$$

Table 9: Default implementation hyperparameters used for RECAST-TC.

Component	Default setting
Historical length	8 steps
Forecast length	4 steps
Visible ERA/GPH length	$T_E = 4$ steps
Blind-window mask	Latest $8 - T_E$ environmental steps masked
Satellite modalities	VIS, IR, WV, PMW
Image resolution	64×64
Trajectory variables	Longitude, latitude, pressure, wind
Trajectory embedding dim.	32
Generator LSTM hidden dim.	64
Discriminator hidden dim.	128
MLP hidden dim.	128
Noise dimension	16
Best-of-K samples	$K = 6$
Satellite feature dim.	64
Satellite TALR GRU hidden dim.	256
Satellite TALR attention hidden dim.	128
Bridge token dim.	64
Bridge Transformer	2 layers, 4 heads
VQ codebook size	256
VQ code dim.	64
VQ commitment weight	0.25
VQ EMA decay	0.99
Batch size	64
Optimizer	AdamW
Generator learning rate	10^{-4}
Discriminator learning rate	10^{-4}
Generator steps	1
Discriminator steps	2
Number of epochs	100

We set

$$\lambda_{\text{bridge}} = 1.0, \quad \lambda_{\text{vis}} = 1.0, \quad \lambda_{\text{code}} = 1.0, \quad \lambda_{\text{commit}} = 0.25.$$

The discriminator is trained on complete observed-plus-future sequences and uses both adversarial classification loss and image consistency loss. All linear layers are initialized with Kaiming initialization, and gradients are clipped only when the corresponding clipping threshold is positive.

For reproducibility, all experiments were conducted on a server equipped with 8 NVIDIA H100 GPUs, 15 CPU cores, 100 GB system memory, and 100 GB shared memory. The experiments were run in parallel on the 8 GPUs, and the longest full training run of the default RECAST-TC configuration takes approximately 30 hours. Runs with different learning-rate settings may vary slightly in execution time, but all experiments can be completed within about 30 hours.

C Blind-Window Sensitivity Across Basins

We further evaluate how the amount of visible environmental history affects forecasting performance under the blind-window protocol. The full-history setting uses all eight historical ERA/GPH steps and serves as an oracle reference. For delayed settings, only the earliest T_E environmental steps are visible, while the remaining $8 - T_E$ recent steps are masked before forecast initialization.

Tables 10 and 11 summarize basin-level track and intensity errors under different visible-history settings. Track errors generally increase as the blind window becomes larger, especially over NA, SP, and WP, indicating that recent environmental context is important for preserving storm motion. Intensity errors show a similar degradation when fewer recent ERA/GPH steps are available, confirming that pressure and wind prediction also benefit from recent environmental constraints.

Broader Impacts. This work aims to improve the realism of tropical cyclone forecasting by accounting for data-availability constraints, which may contribute to more reliable evaluation and

Table 10: Track errors across basins under different blind-window settings. Each entry reports AVG TDE / Step-4 TDE in km, corresponding to ADE/FDE.

Setting	EP	NA	NI	SI	SP	WP	Global
Full history (8/8)	**70.2/122.1**	**101.4/174.4**	**97.6/160.9**	**98.8/163.7**	**116.6/191.9**	**94.5/156.1**	**93.3/157.3**
ERA obs. 7/8	81.5/143.0	118.5/209.8	101.2/171.1	106.4/181.7	129.9/221.1	107.6/181.0	105.8/182.5
ERA obs. 6/8	80.6/143.8	131.9/235.3	117.8/203.5	114.6/198.5	139.1/236.6	120.7/208.0	115.1/201.5
ERA obs. 5/8	88.4/159.7	140.7/251.9	119.7/203.6	120.6/208.4	152.1/262.0	136.6/238.6	125.2/220.6
ERA obs. 4/8	98.7/174.4	155.4/279.5	134.8/233.1	130.2/225.5	152.5/262.8	150.6/264.5	137.1/241.7
ERA obs. 3/8	84.2/152.0	149.7/271.2	125.4/218.2	125.1/218.9	153.1/266.1	140.6/247.8	128.8/229.0
ERA obs. 2/8	83.7/151.0	150.8/271.7	116.2/198.2	127.9/222.2	156.0/270.6	138.7/243.7	128.8/228.1

Table 11: Intensity errors across basins under different blind-window settings. Each entry reports Pressure MAE (hPa) / Wind MAE (m/s).

Setting	EP	NA	NI	SI	SP	WP	Global
Full history (8/8)	**4.58/3.70**	**4.89/3.93**	**6.55/5.13**	**6.27/4.24**	**7.12/4.55**	**6.01/3.70**	**5.56/3.95**
ERA obs. 7/8	4.94/4.07	5.40/4.07	6.34/5.35	6.09/4.28	6.76/4.43	6.84/4.23	5.93/4.21
ERA obs. 6/8	4.60/3.72	4.88/3.77	6.19/5.20	5.91/4.07	6.96/4.56	6.27/3.68	5.55/3.88
ERA obs. 5/8	5.74/4.52	6.98/5.44	7.64/6.42	7.17/4.80	7.99/5.09	7.96/4.98	7.10/5.01
ERA obs. 4/8	8.57/6.71	7.75/5.98	9.35/7.56	7.26/5.06	7.91/5.43	10.34/7.16	8.60/6.30
ERA obs. 3/8	4.81/3.83	5.25/4.03	7.33/5.98	6.10/4.16	7.05/4.57	6.39/3.87	5.79/4.06
ERA obs. 2/8	4.89/3.87	5.20/3.95	7.24/5.62	6.47/4.37	7.12/4.72	6.44/3.85	5.88/4.07

potentially benefit disaster preparedness and risk management. However, the proposed approach may also introduce risks if misused in operational settings. In particular, over-reliance on deterministic forecasts without proper uncertainty quantification could lead to suboptimal or overconfident decisions. In addition, differences in data availability or computational resources may limit equitable access to such models across regions. Care should therefore be taken when deploying these methods, including incorporating uncertainty estimation and adhering to established forecasting protocols.

Table 12: Sources and usage terms of external assets used in this work.

Asset	Source / Owner	License / Terms of Use
ERA5 reanalysis	ECMWF / Copernicus Climate Change Service	Copernicus data license; freely available for research and non-commercial use subject to Copernicus terms.
IBTrACS best-track data	NOAA NCEI	Publicly available tropical cyclone archive; distributed under NOAA open-data policy / WDC full-and-open data access terms.
GraphCast baseline	Google DeepMind	Used as a published research baseline following the original paper and released model/code terms.
Pangu-Weather baseline	Huawei Cloud	Released model weights are provided under CC BY-NC-SA 4.0 terms; commercial use is not permitted.
FuXi baseline	Fudan University	Released model weights are provided under CC BY-NC-SA 4.0 terms; commercial use is not permitted.
HWRF / HWFI reference	NOAA / NCEP operational forecast products	Used as operational forecast references following NOAA/NCEP data access and citation policies.

Licenses for Existing Assets. We summarize the licenses and usage terms of the datasets and models used in this work in Table 12.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction describe the reanalysis-latency mismatch, the blind-window benchmark, and RECAST-TC. Section 2 defines the evaluation protocol, and Section 3.2 reports global and basin-level performance, capacity-matched comparisons, ablations, and case studies supporting these contributions.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Section 5 discusses irregular or degraded satellite observations, differences from fully operational forecasting, limited physical interpretability, the absence of full conservation-law enforcement, deterministic forecasts, and the need for prospective evaluation and calibrated uncertainty.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]

Justification: Appendix A states the assumptions of the illustrative linear mixed-observation model and derives full-history identifiability, rank loss under missing observations, rank restoration with a complementary modality, and noise-dependent recovery error. Section 1 summarizes the connection to the model design.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 2 specifies the storm-disjoint temporal split, input modalities, forecast horizons, and input-timing protocol. Section 3 describes the model and comparison settings. Appendix B provides the training algorithm, architecture details, losses, and default hyperparameters needed to implement the method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data are publicly available. The data protocol is described in Section 2, and Appendix B provides the implementation details and training hyperparameters to support reproduction of the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 2 specifies the input-availability regimes and storm-disjoint train/validation/test split. Section 3.2 describes the shared settings for the capacity-matched comparisons. Appendix B reports the history and forecast lengths, model dimensions, losses, optimizer, learning rates, batch size, and training epochs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 4 reports RECAST-TC performance as mean$\pm$standard deviation over 20 random seeds, and Table 5 reports the same statistics for all capacity-matched methods. The standard deviations describe variation across seeds. The single-seed boundary references, ablations, and basin-level results are reported separately.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B reports the server configuration: eight NVIDIA H100 GPUs, 15 CPU cores, 100 GB of system memory, and 100 GB of shared memory. It also describes parallel execution across GPUs and an approximately 30-hour training runtime for the default RECAST-TC configuration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses meteorological and satellite data for tropical cyclone forecasting research and does not involve human subjects, private personal data, or deceptive deployment. The study is intended to improve realistic evaluation and responsible deployment-oriented forecasting.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader Impacts paragraph in Appendix C discusses potential benefits for forecast evaluation, disaster preparedness, and risk management, together with risks from over-reliance on deterministic forecasts and unequal access to data or computing resources. Section 5 explains the method's relationship to operational forecasting and the need for uncertainty quantification.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The study concerns tropical cyclone forecasting with meteorological and satellite observations. It does not release a general-purpose generative model, scraped personal data, or other assets requiring the safeguards described in this question.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the source datasets and baseline methods. Table 12 in Appendix C summarizes the sources and usage terms of the listed reanalysis, best-track, and forecasting-model assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Section 2 documents the proposed benchmark's data sources, temporal coverage, modalities, input-availability regimes, split protocol, and intended use. Appendix B describes the RECAST-TC implementation, and Appendix C provides additional benchmark results and external-asset usage terms.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing, user studies, or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve human subjects or study participants, so IRB approval or equivalent human-subject review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not a component of the RECAST-TC architecture, training objective, or forecasting inference. Assistance with language editing or LaTeX formatting is limited to manuscript preparation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.