PHYWARP: PARAMETER-EFFICIENT ADAPTATION AND KNOWLEDGE TRANSFER FOR TROPICAL CYCLONE FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

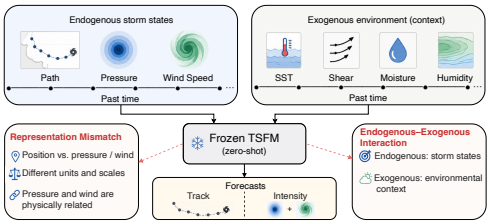
Time-series foundation models (TSFMs) offer reusable forecasting priors that can be adapted to downstream tasks without training dedicated models from scratch. However, tropical cyclone (TC) forecasting presents a challenging transfer setting: available storm records are limited, intensity and motion follow different dynamics, and storm evolution depends on multivariate environmental factors that vary across ocean basins. These characteristics limit the effectiveness of direct TSFM transfer. Moreover, the computational cost of TSFMs can hinder their deployment on resource-constrained meteorological devices. In this work, we introduce PHYWARP, a parameter-efficient framework that adapts frozen TSFMs to TC forecasting by representing endogenous storm states through multiple physically meaningful views and using exogenous environmental context to guide their combination. A shared frozen backbone produces forecasts from these complementary representations, which are combined through lightweight adaptation across lead times and prediction targets. To reduce computational cost, we further distill the adapted knowledge into a lightweight predictor. Experiments on Atlantic and eastern North Pacific data show consistent improvements across three representative TSFM backbones and lead times from 6 to 48 hours, while outperforming supervised time-series models. Ablation studies further demonstrate the complementary benefits of physically informed state representations and adaptive forecast fusion. The lightweight predictor retains strong forecasting performance while substantially reducing inference cost. The code and pretrained weights will be made publicly available.

1 INTRODUCTION

Tropical cyclones (TCs) pose serious threats to human safety and can cause substantial economic losses (Camargo & Wing, 2021; Young & Hsiang, 2024). Sea-level pressure, maximum wind speed, and storm-center position form a compact multivariate sequence that describes key aspects of TC evolution. TC evolution is influenced by multiple environmental factors, including atmospheric and oceanic conditions (Emanuel, 1988; DeMaria & Kaplan, 1994; 1999). From a time-series perspective, these storm-state variables can be viewed as endogenous variables to be forecast, while the environmental factors serve as exogenous variables (Wang et al., 2024b). Forecasting TC track and intensity can therefore be formulated as a multivariate time-series forecasting problem with exogenous variables. This formulation is challenging because storm records are limited, track and intensity follow different dynamics, and environmental influences vary across basins.

Recent time-series foundation models (TSFMs), such as TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), and Moirai (Woo et al., 2024), have demonstrated broad forecasting capabilities through pretraining on large and diverse time-series collections. Their transferability, however, can be sensitive to differences between downstream data and the distributions and representations encountered during pretraining (Kim et al., 2021; Gong et al., 2026). This issue is particularly relevant to TC forecasting, where track and intensity are represented by variables with different numerical structures and physical relationships. Geographic coordinates describe storm position and displacement, whereas pressure and wind provide different but physically related measures of intensity. These relationships are not explicitly reflected when the variables are presented to a generic

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



(a)

Lead	Frozen Chronos-2	PHYWARP Chronos-2	PHYWARP Student
6 h	3.19/4.01/52.53	2.71/3.71/20.45	2.39/3.50/20.93
12 h	5.57/7.44/68.69	4.51/6.50/49.46	4.37/6.22/50.42
24 h	10.23/13.77/175.64	7.61/10.90/131.80	7.81/10.90/133.52
48 h	16.18/23.28/551.26	11.56/16.23/378.93	11.79/16.48/371.08

(b)

Figure 1: Challenges and representative results for adapting frozen TSFMs to TC forecasting. (a) Endogenous storm states and exogenous environmental context present distinct representation challenges. (b) Representative results illustrate the progression from direct frozen-TSFM transfer to PHYWARP adaptation and subsequent knowledge distillation. Entries report MAE for pressure, wind speed, and track in hPa, kt, and km, respectively. The lower is better.

frozen TSFM in their native form. In addition, environmental predictors introduce a related challenge. Such variables have long been used to characterize conditions associated with TC evolution (DeMaria et al., 2005; Kaplan et al., 2015), but they serve as contextual information rather than forecast targets. Treating endogenous and exogenous variables through the same input representation may therefore limit the modeling of their interactions. Figure 1 summarizes these endogenous–exogenous representation challenges and provides representative results illustrating the benefit of adapting frozen TSFMs to TC forecasting.

Motivated by these challenges, we introduce PHYWARP, a parameter-efficient framework for adapting frozen TSFMs to joint TC track and intensity forecasting. Rather than relying on a single representation, PHYWARP constructs complementary physically informed views of endogenous storm states for a shared frozen TSFM. Exogenous environmental context guides lightweight adaptive fusion of the resulting candidate forecasts across prediction targets and lead times. An intensity residual head further corrects pressure and wind errors that cannot be resolved through fusion alone. This design enables TC-specific adaptation without updating the pretrained backbone. To reduce inference cost, we distill the adapted knowledge into a compact predictor that operates without TSFM inference. We evaluate PHYWARP across three TSFM backbones, two ocean basins, and lead times from 6 to 48 hours, together with ablation and distillation analyses. The contributions are summarized as follows:

- We introduce a parameter-efficient approach for adapting TSFMs to TC forecasting by representing endogenous storm states through complementary physically informed views and using exogenous environmental context for target- and lead-specific forecast fusion.
- We further distill the adapted TSFM knowledge into a compact predictor, enabling forecasting without TSFM inference while retaining strong forecasting performance.
- Across three frozen TSFMs, six supervised forecasters, two ocean basins, and lead times from 6 to 48 hours, PHYWARP consistently improves direct TSFM transfer and outperforms supervised time-series models. Extensive analyses further characterize the proposed adaptation and knowledge-transfer mechanisms.

2 RELATED WORK

2.1 STORM-STATE TROPICAL CYCLONE FORECASTING

Storm-state TC forecasting focuses on predicting compact storm-level variables, such as track and intensity, rather than full atmospheric fields. Statistical models use historical storm states and environmental context to predict future intensity and rapid intensification (DeMaria & Kaplan, 1994; 1999; DeMaria et al., 2005; DeMaria, 2009; Kaplan et al., 2010). Numerical weather prediction (NWP) models instead simulate the surrounding atmosphere to forecast track and intensity (Goldenberg et al., 2015; Hazelton et al., 2023), while consensus approaches combine multiple forecast systems (Simon et al., 2018). Although effective, the approaches can require complex configurations and substantial computational resources. Data-driven methods offer an alternative by learning storm

evolution directly from historical observations. In this work, we forecast TC track and intensity from historical storm-state sequences and environmental context, without relying on NWP fields.

2.2 Time-Series Foundation Models

In recent years, task-specific time-series architectures have provided strong baselines. TiDE uses a dense encoder-decoder structure (Das et al., 2023), PatchTST applies channel-independent patch attention (Nie et al., 2023), and iTransformer treats complete variates as tokens (Liu et al., 2024a). TimesNet models temporal variation through multi-period 2D representations (Wu et al., 2023), while TimeMixer performs decomposable multiscale mixing (Wang et al., 2024a). TimeXer explicitly separates endogenous targets from exogenous histories (Wang et al., 2024b). These models are typically trained or tuned for each target dataset. Time-series foundation models instead aim to provide reusable forecasting priors across datasets and tasks. Representative examples include Lag-Llama (Rasul et al., 2023), Chronos (Ansari et al., 2024), MOMENT (Goswami et al., 2024), Timer (Liu et al., 2024b), TimesFM (Das et al., 2024), Moirai (Woo et al., 2024), and Chronos-2 (Ansari et al., 2025). The models enable the reuse of broad time-series priors on relatively small downstream datasets. However, the native representation of TC variables may not fully reflect the physical relationships governing storm motion and intensity, limiting direct transfer of frozen TSFMs. PhyWarp addresses this through TC-specific physical representations and lightweight adaptation.

2.3 Time-Series Adaptation and Distillation

Time-series forecasting has explored various forms of input and model adaptation to improve robustness and transferability. RevIN (Kim et al., 2021) and Non-stationary Transformers (Liu et al., 2022) transform inputs to address distribution shift and non-stationarity, but primarily target generic statistical variation. More recently, TATO (Qiu et al., 2026), δ-Adapter (Liang et al., 2026), and TS-Memory (Lyu et al., 2026) adapt pretrained time-series forecasters while keeping their backbones frozen. Unlike these generic approaches, PhyWarp constructs complementary TC-specific physical representations and uses environmental context for target- and lead-specific adaptive fusion of their forecasts. Complementary to adaptation, knowledge distillation transfers predictive behavior from a larger model to a compact predictor (Hinton et al., 2015). Rather than distilling a generic pretrained model directly, we first adapt the frozen TSFM to TC forecasting and then distill the adapted knowledge, enabling inference without TSFM computation.

3 Problem Definition

Let $\mathbf{s}_t \in \mathbb{R}^{d_s}$ denote the cyclone state at time t, including intensity and position variables, and let $\mathbf{e}_t \in \mathbb{R}^{d_e}$ denote the environmental context observed at the same time. Given the historical state and environmental sequences $\mathbf{S}_{1:L} = [\mathbf{s}_1, \ldots, \mathbf{s}_L]$ and $\mathbf{E}_{1:L} = [\mathbf{e}_1, \ldots, \mathbf{e}_L]$, the task is to predict the next H cyclone states:

$$\widehat{\mathbf{S}}_{L+1:L+H} = f(\mathbf{S}_{1:L}, \mathbf{E}_{1:L}) \in \mathbb{R}^{H \times d_s}. \tag{1}$$

PhyWarp implements f using deterministic multi-view transformations and lightweight environmental, fusion, and correction modules, while keeping the pretrained TSFM F_θ frozen.

4 Methodology

Figure 2 summarizes the three-stage adaptation framework and its subsequent lightweight knowledge-transfer stage. First, complementary representations expose different aspects of endogenous storm-state evolution and are forecast by a shared frozen TSFM to form a candidate bank. Second, exogenous environmental context is used to estimate the relative reliability of these candidates and adaptively combine their forecasts. Finally, a lightweight intensity residual head adjusts pressure and wind errors that cannot be resolved through candidate fusion alone. During adaptation, only the lightweight environmental, fusion, and residual modules are trained; the TSFM backbone and state transformations remain fixed. To reduce computational cost, we further distill the adapted predictor into a lightweight student that operates without TSFM inference.

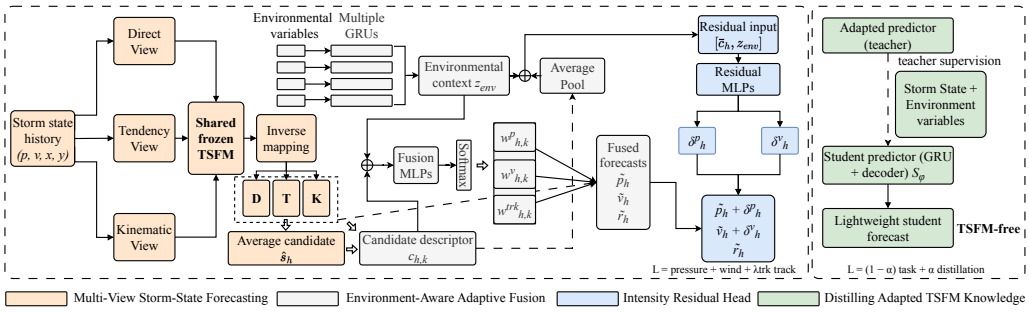


Figure 2: Architecture of PHYWARP. A shared frozen TSFM forecasts three complementary storm-state views, whose predictions are adaptively fused using environmental context and refined by an intensity residual head. Only the lightweight adaptation modules are trained. The adapted predictor then serves as an offline teacher for a compact student, enabling TSFM-free inference.

4.1 MULTI-VIEW STORM-STATE FORECASTING

We represent the storm state as $\mathbf{s}_t = (p_t, v_t, x_t, y_t)$, where p_t and v_t denote minimum sea-level pressure and maximum sustained wind, respectively, and (x_t, y_t) denotes the local east–north displacement relative to forecast initialization. Rather than presenting the storm state only in its native form, we construct three deterministic time-series representations:

$$D : (p, v, x, y), \quad T : \Delta(p, v, x, y), \quad K : (p, v, \Delta u_{\parallel}, \Delta u_{\perp}). \tag{2}$$

Let $\mathcal{V} = \{D, T, K\}$ denote the candidate views. Each transformed sequence is independently passed through the same frozen TSFM F_θ and deterministically mapped back to the common storm-state space. For view $k \in \mathcal{V}$ and prediction step $h = 1, \ldots, H$, we denote the candidate forecast for time $L + h$ by $\widehat{\mathbf{s}}_h^{(k)} = (\widehat{p}_h^{(k)}, \widehat{v}_h^{(k)}, \widehat{\mathbf{r}}_h^{(k)})$, where $\widehat{\mathbf{r}}_h^{(k)} = (\widehat{x}_h^{(k)}, \widehat{y}_h^{(k)})$.

The Direct view preserves the native storm state. The Tendency view represents temporal changes by forecasting state increments, which are accumulated from the latest observation to recover absolute states. The Kinematic view retains pressure and wind but represents storm displacement relative to its recent heading. For $\mathbf{r}_t = (x_t, y_t)$, $(\Delta u_{\parallel,t}, \Delta u_{\perp,t}) = R(\gamma)(\mathbf{r}_t - \mathbf{r}_{t-1})$, where $R(\gamma)$ is a two-dimensional rotation defined by the heading γ, estimated from the most recent 12 hours of observed storm motion. Thus, $\Delta u_{\parallel}$ and $\Delta u_{\perp}$ describe displacement along and across the recent heading direction, respectively. Predicted heading-relative displacements are transformed back and accumulated to recover storm positions. The explicit form of $R(\gamma)$ and the heading-estimation procedure are provided in the Appendix.

4.2 ENVIRONMENT-AWARE ADAPTIVE FUSION

Environmental variables serve as exogenous context rather than forecast targets. We therefore use their observed history to guide candidate fusion rather than treating them as additional storm-state variables forecast by the frozen TSFM. The selected SHIPS environmental variables are organized into four groups according to their physical characteristics: oceanic, thermodynamic, dynamical, and geographic. Each group is encoded by an independent lightweight GRU, and the resulting representations are concatenated as

$$\mathbf{z}_{\text{env}} = [\text{GRU}_{\text{ocean}}(\mathbf{E}_{1:L}^{(\text{ocean})}), \text{GRU}_{\text{thermo}}(\mathbf{E}_{1:L}^{(\text{thermo})}), \text{GRU}_{\text{dyn}}(\mathbf{E}_{1:L}^{(\text{dyn})}), \text{GRU}_{\text{geo}}(\mathbf{E}_{1:L}^{(\text{geo})})]. \tag{3}$$

Candidate reliability can vary with the current storm state, environmental conditions, prediction target, and forecast step. For candidate k at prediction step h, we define

$$\bar{\mathbf{s}}_h = \frac{1}{|\mathcal{V}|} \sum_{m \in \mathcal{V}} \widehat{\mathbf{s}}_h^{(m)}, \qquad \mathbf{c}_{h,k} = \left[\mathbf{s}_L, \widehat{\mathbf{s}}_h^{(k)}, \widehat{\mathbf{s}}_h^{(k)} - \bar{\mathbf{s}}_h, h/H, \text{onehot}(k) \right]. \tag{4}$$

Here, $\bar{\mathbf{s}}_h$ is the average candidate forecast, while $\mathbf{c}_{h,k}$ describes the current storm state, candidate forecast, candidate disagreement, forecast progress, and view identity. For prediction target $q \in \{p, v, \text{trk}\}$, the fusion weights are computed as

$$w_{h,k}^{q} = \text{softmax}_{k \in \mathcal{V}} \, G_q \left([\mathbf{c}_{h,k}, \mathbf{z}_{\text{env}}] \right), \tag{5}$$

where G_q is a lightweight fusion MLP for target q. The fused predictions are

$$\widetilde{p}_h = \sum_{k\in\mathcal{V}} w^p_{h,k}\widehat{p}^{(k)}_h, \qquad \widetilde{v}_h = \sum_{k\in\mathcal{V}} w^v_{h,k}\widehat{v}^{(k)}_h, \qquad \widetilde{\mathbf{r}}_h = \sum_{k\in\mathcal{V}} w^{\mathrm{trk}}_{h,k}\widehat{\mathbf{r}}^{(k)}_h. \tag{6}$$

Pressure, wind, and track therefore use separate candidate reliability weights, while the two track coordinates share the same weights.

4.3 Intensity Residual Head

Adaptive fusion forms a convex combination of candidate forecasts and is therefore limited to combinations of the available candidates. To correct intensity errors that cannot be resolved through fusion alone, we pool the candidate descriptors as $\overline{\mathbf{c}}_h = \frac{1}{|\mathcal{V}|}\sum_{k\in\mathcal{V}}\mathbf{c}_{h,k}$, and predict residual adjustments for pressure and wind: $\delta^q_h = C_q([\overline{\mathbf{c}}_h, \mathbf{z}_{\mathrm{env}}]), \qquad q \in \{p, v\}$, where C_q is a lightweight residual MLP for target q. Notably, the view-specific disagreement cancels under pooling but guides adaptive fusion. The final forecast is $\widehat{\mathbf{s}}_h = (\widetilde{p}_h + \delta^p_h,\ \widetilde{v}_h + \delta^v_h,\ \widetilde{\mathbf{r}}_h)$. The residual head modifies only pressure and wind and is disabled at the first prediction step. The adaptation modules are jointly optimized using

$$\mathcal{L} = [H(2+\lambda_{\mathrm{trk}})]^{-1}\sum\nolimits_h \mathbb{E}_{\mathrm{train}}\left[\ell^p_h + \ell^v_h + \lambda_{\mathrm{trk}}\ell^{\mathrm{trk}}_h\right], \qquad h = 1, \ldots, H, \tag{7}$$

where ℓ^p_h and ℓ^v_h are scaled absolute errors for pressure and wind, and ℓ^{trk}_h is the scaled Euclidean track error. This scaling places the three forecasting objectives on comparable scales during joint optimization. Each horizon-specific model is trained on all intermediate prediction steps, while evaluation is performed at the target lead time.

4.4 Distilling Adapted TSFM Knowledge

Although PhyWarp keeps the TSFM frozen, constructing the candidate bank still requires multiple backbone queries. We therefore use the adapted predictor as an offline teacher and transfer its forecasting capability to a compact student that operates without TSFM inference.

The student receives the observed storm-state and environmental histories, including environmental missingness indicators, but does not use the candidate bank. A single-layer GRU jointly encodes the storm-state and environmental histories together with the missingness indicators. Its final hidden state is combined with a learned prediction-step embedding and decoded into a state change relative to the latest observation:

$$\widehat{\mathbf{s}}^{\mathrm{stu}}_h = \mathbf{s}_L + \mathbf{d}_h \odot S_\phi\left(\mathbf{h}_{\mathrm{GRU}} + \mathbf{a}_h\right), \qquad h = 1, \ldots, H, \tag{8}$$

where $\mathbf{a}_h$ is the prediction-step embedding, S_ϕ is the student decoder, and $\mathbf{d}_h$ contains state-change scales estimated from the training partition. The student is trained using both observed targets and forecasts from the adapted teacher:

$$\mathcal{L}_{\mathrm{stu}} = (1-\alpha)\mathcal{L}_{\mathrm{task}}\left(\widehat{\mathbf{S}}^{\mathrm{stu}}_{L+1:L+H}, \mathbf{S}_{L+1:L+H}\right) + \alpha\mathcal{L}_{\mathrm{task}}\left(\widehat{\mathbf{S}}^{\mathrm{stu}}_{L+1:L+H}, \widehat{\mathbf{S}}^{\mathrm{tea}}_{L+1:L+H}\right), \tag{9}$$

where $\mathcal{L}_{\mathrm{task}}$ uses the same pressure, wind, and track errors as the adaptation objective, and α controls the contribution of teacher supervision. Teacher forecasts are generated offline and treated as fixed targets during student training. At inference, only the student is retained, eliminating candidate generation and TSFM inference.

5 Experimental Setup

5.1 Dataset

We use the Atlantic (AL) and eastern North Pacific (EP) archive of the Statistical Hurricane Intensity Prediction Scheme (SHIPS), a long-standing operational framework for tropical cyclone intensity forecasting that has been continuously developed and updated (DeMaria & Kaplan, 1994; 1999; DeMaria et al., 2005). The archive provides storm analyses and environmental diagnostics at six-hour intervals. The forecast state consists of minimum sea-level pressure, maximum sustained wind, and

storm-center location, which are treated as endogenous variables. We additionally use eight historical environmental predictors as exogenous context, grouped into four categories: Ocean (CSST and COHC), Thermodynamic (MPI deficit and RHMD), Dynamics (SHRD, D200, and V850), and Geography (DTL). We conduct four separate forecasting experiments at lead times of 6, 12, 24, and 48 hours. Each experiment uses the preceding 48 hours of observations to predict the storm state at the corresponding lead time. We retain tropical and subtropical cases with complete state histories and forecast targets, allowing multiple consecutive initialization times from the same storm. This yields 5,819 cases from 2001–2023, which are chronologically divided into a training period (2001–2011), a validation period (2012–2016), and a test period (2017–2023). Appendix provides detailed dataset statistics and preprocessing information.

5.2 Evaluation Metrics.

We consider four forecasting tasks corresponding to lead times of 6, 12, 24, and 48 hours and evaluate each task separately. We report MAE and RMSE for pressure and wind. Track error is the mean distance between predicted and observed storm centers, computed using the Vincenty inverse solution on WGS84 with a Haversine fallback upon non-convergence. Lower values indicate better performance for all metrics. Unless otherwise specified, results are averaged over three independent runs with different random seeds, and we report the mean and standard deviation. The same experimental settings are used across all runs.

5.3 Baselines

Supervised time-series models. We compare PhyWarp with six representative supervised time-series forecasting models: PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024a), TimeXer (Wang et al., 2024b), TiDE (Das et al., 2023), TimesNet (Wu et al., 2023), and TimeMixer (Wang et al., 2024a). PatchTST and TiDE use only the four storm-state variables because their channel-wise implementations do not support historical environmental conditioning through simple channel concatenation; TiDE's covariate interface also expects future inputs unavailable in our setting. The other four models, denoted by +Env., use the same eight historical SHIPS environmental variables as PhyWarp. All supervised baselines follow the same chronological data split and model-selection protocol. Detailed hyperparameter settings are provided in the Appendix.

Frozen TSFM backbones. Our framework is evaluated with three representative pretrained TSFM backbones: TimesFM (Das et al., 2024), Moirai (Woo et al., 2024), and Chronos-2 (Ansari et al., 2025), all of which remain frozen throughout adaptation. For each backbone, we construct a direct baseline that forecasts from the historical storm-state sequence without our adaptation framework. We also consider a +Env. baseline that directly incorporates the same historical environmental information. These baselines isolate the contribution of PhyWarp from that of the pretrained backbone and test whether simply adding environmental context is sufficient to improve TC forecasting. TimesFM (+Env.) is omitted because, under this direct-input construction, its channel-independent interface prevents the appended environmental channels from influencing the state forecasts, yielding predictions identical to those of direct TimesFM.

5.4 Implementation Details

The adaptation network consists of four environmental GRUs, three target-specific fusion heads for pressure, wind, and track, and two intensity residual heads for pressure and wind. Each fusion and residual head uses two 64-unit hidden layers with GELU activations and dropout of 0.1. A separate adaptation model is trained for each forecast horizon using AdamW for up to 300 epochs, with batch size 128, learning rate 3×10^{-3}, weight decay 10^{-3}, and early-stopping patience 30. We set $\lambda_{\mathrm{trk}} = 1.0$ for the joint forecasting objective. The residual head is disabled at the first prediction step. The pretrained TSFM backbone and all deterministic D/T/K transformations remain frozen. Candidate forecasts from the frozen backbone are precomputed and cached before adaptation training. The same optimization configuration is used across all TSFM backbones and forecast horizons. For lightweight knowledge transfer, we use PhyWarp-Chronos-2 as the teacher and set the distillation weight to $\alpha = 0.5$. The student is trained using the same training partition and forecast horizons as the adaptation models. Teacher forecasts are precomputed and cached before

Table 1: Forecast errors at 6- and 12-h lead times over the AL and EP basins. Pressure (P) and wind (V) are reported as MAE/RMSE in hPa and kt, respectively, and track error (D) as MAE in km. Frozen TSFMs involve no task-specific training and are reported once. The lightweight student is distilled from PHYWARP-Chronos-2 and requires no TSFM inference at deployment. Bold and underlined values indicate the best and second-best unrounded means, respectively.

Lead	Model	Atlantic (AL)			Eastern North Pacific (EP)		
		P (MAE/RMSE)	V (MAE/RMSE)	D (MAE)	P (MAE/RMSE)	V (MAE/RMSE)	D (MAE)
6 h	PatchTST	2.85±0.16/4.69±0.22	4.09±0.22/6.13±0.40	32.53±1.10	3.21±0.24/4.68±0.30	4.30±0.31/5.85±0.53	27.23±0.68
	iTransformer (+Env.)	2.78±0.04/4.65±0.04	4.02±0.07/6.03±0.11	27.89±0.49	3.04±0.06/4.58±0.04	4.07±0.13/5.68±0.16	22.42±1.80
	TimeXer (+Env.)	2.59±0.02/4.28±0.10	3.93±0.21/5.70±0.40	37.14±3.02	2.80±0.12/4.09±0.15	4.13±0.43/5.49±0.61	31.48±1.18
	TiDE	2.69±0.02/4.59±0.01	3.91±0.04/5.88±0.02	24.64±0.24	2.90±0.01/4.37±0.01	3.92±0.03/5.47±0.02	18.63±0.16
	TimesNet (+Env.)	3.33±0.35/5.05±0.52	4.71±0.34/6.69±0.50	48.50±0.92	3.84±0.49/5.51±0.69	4.92±0.52/6.66±0.66	37.77±1.99
	TimeMixer (+Env.)	3.02±0.21/4.68±0.34	4.46±0.31/6.40±0.49	34.85±3.26	3.42±0.22/4.86±0.33	4.57±0.33/6.11±0.44	26.48±1.99
	TimesFM	3.18/4.97	4.62/6.71	37.82	3.62/5.23	4.89/6.59	31.23
	Moirai	3.04/5.01	4.69/6.68	95.95	3.52/4.98	4.63/6.34	88.86
	Moirai (+Env.)	3.26/5.25	5.40/7.79	106.55	3.63/5.09	5.03/6.93	96.47
	Chronos-2	3.00/5.42	3.90/5.90	51.37	3.45/5.86	4.16/5.92	54.06
	Chronos-2 (+Env.)	3.16/5.83	4.01/6.22	57.44	3.82/6.73	4.49/6.52	55.78
	PHYWARP-TimesFM	2.52±0.01/4.14±0.02	3.69±0.01/5.49±0.04	27.05±0.05	2.74±0.01/4.09±0.02	3.74±0.01/5.18±0.02	19.92±0.09
	PHYWARP-Moirai	2.46±0.01/4.17±0.02	3.61±0.00/5.35±0.01	25.53±0.07	2.68±0.01/3.92±0.02	3.57±0.01/4.83±0.01	18.69±0.01
	PHYWARP-Chronos-2	2.65±0.00/4.60±0.02	3.74±0.04/5.62±0.09	**22.60±0.02**	2.80±0.02/4.49±0.01	3.68±0.02/5.22±0.02	**17.61±0.05**
	PHYWARP-Student	**2.32±0.01/3.84±0.02**	**3.50±0.02/5.00±0.03**	23.40±0.23	**2.49±0.01/3.72±0.01**	**3.49±0.01/4.76±0.03**	17.67±0.25
12 h	PatchTST	5.13±0.04/8.51±0.07	7.46±0.12/11.42±0.33	77.40±6.01	5.74±0.07/8.29±0.14	7.91±0.32/10.79±0.56	63.34±8.60
	iTransformer (+Env.)	5.32±0.07/8.78±0.08	7.66±0.07/11.81±0.20	61.55±2.81	5.85±0.17/8.64±0.20	8.06±0.22/11.34±0.36	45.06±2.02
	TimeXer (+Env.)	4.90±0.07/7.73±0.06	7.14±0.37/10.33±0.34	75.43±6.19	5.06±0.02/7.17±0.06	7.17±0.35/9.67±0.63	56.94±3.36
	TiDE	5.26±0.01/8.89±0.01	7.71±0.01/11.83±0.03	64.60±0.31	5.91±0.02/8.70±0.03	8.09±0.02/11.20±0.04	48.56±0.61
	TimesNet (+Env.)	5.09±0.02/7.72±0.08	7.21±0.07/10.39±0.10	88.98±1.92	5.75±0.09/8.12±0.16	7.68±0.08/10.37±0.15	66.18±0.18
	TimeMixer (+Env.)	5.62±0.32/8.61±0.52	7.97±0.50/11.52±0.70	82.47±8.84	6.32±0.62/8.79±0.80	8.59±0.97/11.37±1.13	60.80±6.58
	TimesFM	5.72/8.61	8.21/11.73	88.62	6.55/9.25	9.16/12.25	70.54
	Moirai	6.21/9.82	8.74/12.45	208.57	7.26/9.94	8.97/11.96	194.25
	Moirai (+Env.)	7.51/11.87	9.98/14.23	243.00	8.72/12.57	10.35/14.40	237.88
	Chronos-2	5.19/8.98	7.15/11.00	77.35	6.06/9.55	7.82/11.02	57.21
	Chronos-2 (+Env.)	5.45/9.73	7.40/11.69	88.09	6.62/10.99	8.45/12.42	69.15
	PHYWARP-TimesFM	4.51±0.02/6.79±0.11	6.65±0.02/9.28±0.17	66.49±0.16	4.71±0.07/6.63±0.15	6.61±0.10/8.79±0.16	47.19±0.07
	PHYWARP-Moirai	4.52±0.01/6.98±0.02	6.45±0.04/9.08±0.11	64.60±0.62	4.67±0.07/6.49±0.11	6.50±0.09/8.50±0.17	45.60±0.48
	PHYWARP-Chronos-2	4.43±0.02/6.76±0.05	6.50±0.01/9.14±0.04	**55.83±0.16**	4.61±0.05/6.40±0.12	6.50±0.04/8.55±0.02	**41.02±0.10**
	PHYWARP-Student	**4.34±0.02/6.62±0.09**	**6.23±0.01/8.85±0.12**	57.39±0.91	**4.42±0.05/6.23±0.08**	**6.21±0.10/8.29±0.12**	41.19±0.18

student training. Experiments are conducted on an Intel Xeon Platinum 8370C CPU and a single NVIDIA A800 GPU.

6 RESULTS

6.1 OVERALL PERFORMANCE COMPARISON

Table 1 compares 6- and 12-hour results over the AL and EP basins. Direct frozen TSFMs generally underperform the strongest supervised time-series models, particularly for track prediction, highlighting the difficulty of directly transferring generic forecasting priors to TC forecasting. Naively augmenting Moirai or Chronos-2 with historical environmental variables consistently degrades performance at 6 and 12 hours, suggesting that direct input augmentation does not effectively exploit environmental context at these lead times. In contrast, PHYWARP consistently improves all three TSFM backbones across both basins and forecasting targets, demonstrating that the proposed adaptation is not tied to a particular pretrained model. The best adapted models outperform the supervised baselines across all reported metrics. At 6 hours, PHYWARP-Chronos-2 reduces track MAE relative to TiDE by 8.3% in AL and 5.5% in EP. The distilled student further improves short-range intensity forecasting while remaining close to the adapted teacher in track prediction. At 12 hours, relative to TimeXer, it reduces pressure MAE by 11.4% and 12.6%, and wind MAE by 12.7% and 13.4%, in AL and EP, respectively. These results indicate that the knowledge acquired through TSFM adaptation can be effectively transferred to a compact student while retaining the performance gains over supervised models. Complete results at 24 and 48 hours are provided in the Appendix and show similar overall trends. A 6-hour forecast-observation visualization is also provided in Appendix.

6.2 ABLATION STUDIES AND FURTHER ANALYSIS

We conduct ablation studies with Chronos-2. Table 2 examines PHYWARP from two complementary perspectives. First, we progressively construct the candidate bank from the Direct view (A) to the

Table 2: Ablation analysis of PHYWARP with Chronos-2 on the pooled AL and EP test set, reporting MAE scores. A, B, and C denote the Direct, Tendency, and Kinematic views, respectively, and Full denotes $A + B + C$. The first three rows progressively construct the candidate bank, while the remaining rows remove environmental context (Env.), adaptive fusion (Fusion), or intensity correction (Corr.) from the full model.

Config	6 h			12 h			24 h			48 h		
	P	V	D	P	V	D	P	V	D	P	V	D
A	3.19±0.00	4.01±0.00	52.53±0.00	4.51±0.03	6.47±0.02	68.69±0.00	7.69±0.04	10.95±0.07	175.64±0.00	11.38±0.13	16.14±0.20	551.26±0.00
$A + B$	2.75±0.01	3.75±0.02	20.78±0.01	4.56±0.07	6.52±0.05	50.57±0.13	7.68±0.15	10.92±0.07	136.21±0.61	11.45±0.26	16.12±0.23	387.53±0.97
Full	2.71±0.01	3.71±0.02	20.45±0.03	4.51±0.01	6.50±0.01	49.46±0.08	7.61±0.02	10.90±0.12	131.80±0.79	11.56±0.23	16.23±0.28	378.93±1.04
w/o Env.	2.73±0.01	3.72±0.01	20.47±0.01	4.69±0.04	6.70±0.01	49.74±0.13	8.35±0.05	11.77±0.05	132.42±0.72	12.83±0.02	17.99±0.07	378.42±1.52
w/o Fusion	2.94±0.00	3.83±0.00	27.03±0.00	4.47±0.04	6.49±0.08	51.93±0.00	7.71±0.04	10.90±0.13	136.01±0.00	11.53±0.17	16.23±0.26	394.89±0.00
w/o Corr.	2.71±0.01	3.71±0.02	20.45±0.03	5.19±0.01	7.09±0.06	49.79±0.19	9.43±0.00	12.87±0.04	132.34±1.07	14.71±0.03	21.00±0.04	379.31±1.02

Direct and Tendency views ($A + B$), and finally to the full three-view configuration ($A + B + C$). Second, starting from the full model, we remove environmental context, adaptive fusion, and intensity correction individually to assess their respective contributions. For the w/o Fusion variant, D/T/K forecasts are uniformly averaged, while the remaining modules are retained and retrained.

Component contributions. The progressive construction of the candidate bank shows that additional views are particularly important for track forecasting. Adding the Tendency view to the Direct view reduces track MAE from 52.53 to 20.78 km at 6 hours and from 551.26 to 387.53 km at 48 hours. Incorporating the Kinematic view further reduces the errors to 20.45 and 378.93 km, respectively, with similar improvements at intermediate lead times. The additional views mainly improve track forecasting, with smaller intensity gains and slight degradation at 48 hours. To test whether Tendency alone explains these gains, we apply it to iTransformer (+Env.). Its track errors improve at 6 and 48 hours but worsen at 12 and 24 hours, while PHYWARP-Chronos-2 performs better at all horizons, which is presented in Appendix.

The component-wise ablations reveal distinct and complementary roles of the remaining mechanisms. Removing environmental context has little effect at short lead times but increasingly degrades intensity prediction as lead time grows. At 48 hours, pressure and wind MAE increase from 11.56 and 16.23 to 12.83 and 17.99, respectively, while track error remains nearly unchanged. Adaptive fusion contributes most clearly to track prediction: removing it increases track MAE from 20.45 to 27.03 km at 6 hours and from 378.93 to 394.89 km at 48 hours, while its effect on intensity is comparatively small. In contrast, the residual head primarily benefits medium- and long-range intensity forecasting. It is inactive at 6 hours, while its removal increases pressure and wind MAE by 3.15 hPa and 4.77 kt, respectively, at 48 hours. Overall, multi-view forecasting and adaptive fusion primarily improve track prediction, whereas environmental context and residual correction provide larger benefits for medium- and long-range intensity forecasting.

Adaptive fusion analysis. Figure 3 shows the learned fusion weights averaged over test cases. The weighting patterns vary across targets and lead times. At 6 hours, the Tendency view dominates pressure and wind forecasting, receiving 53.8% and 61.2% of the weights, respectively. For track, the Kinematic view receives the largest weight at every lead time, decreasing from 56.3% at 6 hours to 46.0% at 48 hours. Meanwhile, the Direct-view weight generally increases with lead time, indicating that the model progressively rebalances native and transformed representations. These patterns support the use of target- and lead-specific adaptive fusion. Fusion weights reflect reliance within the full candidate bank, rather than the marginal contribution of individual views.

Label efficiency. Table 3 (a) shows that PHYWARP retains most of its accuracy with only 50% of the training storms. Track MAE increases by at most 1.04%, compared with 4.85% for pressure and 5.54% for wind, indicating greater label efficiency for track adaptation. Results from 5% to 100% of the training data are provided in the Appendix.

Parameter efficiency. Table 3 (b) shows that PHYWARP uses only 23,427 trainable adapter parameters at 6 hours and 38,149 at 12–48 hours, independent of the frozen TSFM backbone. The 6-hour adapter omits the intensity residual heads. After candidate forecasts are cached, adapter training takes only 27.33–29.49 seconds on average.

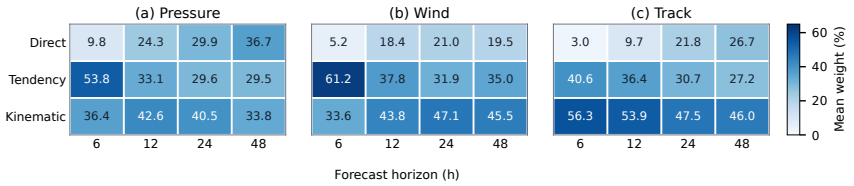


Figure 3: Learned fusion weights for the Direct, Tendency, and Kinematic views, averaged over test cases and three random seeds. The distinct patterns across targets and lead times illustrate the target- and lead-specific behavior of adaptive fusion.

(a) Label efficiency

Metric	6 h	12 h	24 h	48 h
Pressure	+0.18%	+3.10%	+4.85%	+4.68%
Wind	-0.04%	+3.12%	+4.75%	+5.54%
Track	+0.26%	+0.76%	+0.69%	+1.04%

(b) Parameter efficiency

Backbone	Adapter parameters 6 h / 12–48 h	Training time (s) Mean±std.
TimesFM	23,427 / 38,149	29.49±2.51
Moirai		27.33±3.54
Chronos-2		28.53±3.16

Table 3: Efficiency analysis of PHYWARP. (a) Relative MAE change using 50% rather than 100% of the training storms; positive values indicate increased error. (b) Trainable adapter parameters and training time after candidate forecasts are cached.

Table 4: Distillation analysis on the pooled AL and EP test set, reporting MAE scores. Direct-KD distills frozen Chronos-2, whereas PHYWARP-KD distills adapted PHYWARP-Chronos-2. CPU latency denotes batch-one inference after preprocessing, excluding model loading and disk I/O. All student variants share the same inference latency.

Configuration	6 h			12 h			24 h			48 h			Inference (ms/sample)
	P	*V*	*D*	*P*	*V*	*D*	*P*	*V*	*D*	*P*	*V*	*D*	
Teacher	2.71±0.01	3.71±0.03	**20.45±0.03**	4.51±0.01	6.50±0.01	**49.46±0.08**	**7.61±0.02**	10.90±0.12	**131.80±0.79**	**11.56±0.23**	**16.23±0.28**	378.93±1.04	340.92±6.22
Scratch	2.46±0.02	3.50±0.03	24.53±0.28	4.50±0.02	6.38±0.01	56.61±0.67	7.93±0.11	10.95±0.13	144.08±1.65	11.99±0.11	16.68±0.20	377.01±0.97	
Direct-KD	2.49±0.04	3.52±0.01	28.65±0.82	4.72±0.06	6.64±0.07	58.00±0.48	8.73±0.07	12.18±0.09	154.42±0.50	13.36±0.08	18.99±0.19	463.98±3.86	**3.66±0.02**
PHYWARP-KD	**2.39±0.00**	**3.50±0.01**	20.93±0.24	**4.37±0.03**	**6.22±0.05**	50.42±0.44	7.81±0.14	**10.90±0.11**	133.52±2.20	11.79±0.28	16.48±0.33	**371.08±4.84**	

6.3 DISTILLATION ANALYSIS

Table 4 shows that distilling the adapted teacher is more effective than distilling frozen Chronos-2 directly. The scratch student already achieves competitive performance, particularly in short-range intensity forecasting. Direct-KD consistently degrades its performance, whereas Student + PHYWARP-KD improves nearly all metrics, reducing track MAE by 14.7%, 10.9%, 7.3%, and 1.6% at 6, 12, 24, and 48 hours, respectively. The distilled student remains close to the adapted teacher across horizons while retaining strong intensity forecasting performance. Without TSFM inference, the student reduces CPU latency from 340.92 to 3.66 ms per sample, a 93.1× speedup. These results demonstrate the benefits of distilling task-adapted TSFM knowledge into a compact predictor. Additional basin-specific results are provided in the Appendix.

7 CONCLUSION AND DISCUSSION

We introduced PHYWARP, a parameter-efficient framework for adapting frozen time-series foundation models to tropical cyclone forecasting through physically informed storm-state representations and environment-aware adaptive fusion. Experiments across three TSFM backbones, two ocean basins, and lead times from 6 to 48 hours demonstrate consistent improvements over direct TSFM transfer, with the best adapted models outperforming the evaluated supervised time-series baselines. We further demonstrate that the adapted forecasting knowledge can be distilled into a compact student with near-teacher accuracy and substantially lower computational cost. Our evaluation is limited to the AL and EP basins and lead times up to 48 hours, while requiring complete targets through 48 hours may underrepresent cases of early storm dissipation.

AI USE STATEMENT

We used ChatGPT 5.5 to assist with language polishing, improving the clarity and readability of the manuscript, and supporting literature searches. All AI-generated suggestions and retrieved references were carefully reviewed, revised where necessary, and verified by the authors. The authors retain full responsibility for the scientific content, claims, references, and final presentation of the work.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Suzana J Camargo and Allison A Wing. Increased tropical cyclone risk to coasts. *Science*, 371 (6528):458–459, 2021.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan K Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=pCbC3aQB5W.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Mark DeMaria. A simplified dynamical system for tropical cyclone intensity prediction. *Monthly Weather Review*, 137(1):68–82, 2009.

Mark DeMaria and John Kaplan. A statistical hurricane intensity prediction scheme (ships) for the atlantic basin. *Weather and Forecasting*, 9(2):209–220, 1994.

Mark DeMaria and John Kaplan. An updated statistical hurricane intensity prediction scheme (ships) for the atlantic and eastern north pacific basins. *Weather and Forecasting*, 14(3):326–337, 1999.

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.

Kerry A Emanuel. The maximum intensity of hurricanes. *J. Atmos. Sci*, 45(7):1143–1155, 1988.

Stanley B Goldenberg, Sundararaman G Gopalakrishnan, Vijay Tallapragada, Thiago Quirino, Frank Marks Jr, Samuel Trahan, Xuejin Zhang, and Robert Atlas. The 2012 triply nested, high-resolution operational version of the hurricane weather research and forecasting model (hwrf): Track and intensity forecast verifications. *Weather and Forecasting*, 30(3):710–729, 2015.

Peiliang Gong, Emadeldeen Eldele, Min Wu, Zhenghua Chen, Xiaoli Li, and Daoqiang Zhang. Bridging distribution gaps in time series foundation model pretraining with prototype-guided normalization. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Andrew Hazelton, Ghassan J Alaka Jr, Lew Gramer, William Ramstrom, Sarah Ditchek, Xiaomin Chen, Bin Liu, Zhan Zhang, Lin Zhu, Weiguo Wang, et al. 2022 real-time hurricane forecasts from an experimental version of the hurricane analysis and forecast system (hafsv0. 3s). *Frontiers in Earth Science*, 11:1264969, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

John Kaplan, Christopher M Rozoff, Mark DeMaria, Charles R Sampson, James P Kossin, Christopher S Velden, Joseph J Cione, Jason P Dunion, John A Knaff, Jun A Zhang, et al. Evaluating environmental impacts on tropical cyclone rapid intensification predictability utilizing statistical models. *Weather and Forecasting*, 30(5):1374–1396, 2015.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *Proceedings of the International conference on learning representations*, 2021.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=syfWdclGE1.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Proceedings of the Advances in neural information processing systems*, 35:9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of the International conference on learning representations*, volume 2024, pp. 11116–11140, 2024a.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/liu24cb.html.

Sisuo Lyu, Siru Zhong, Tiegang Chen, Weilin Ruan, Qingxiang Liu, Taiqiang Lv, Qingsong Wen, Raymond Chi-Wing Wong, and Yuxuan Liang. Ts-memory: Plug-and-play memory for time series foundation models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 3562–3572, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817614. URL https://doi.org/10.1145/3770855.3817614.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Yunzhong Qiu, Zhiyao Cen, Zhongyi Pei, Chen Wang, and Jianmin Wang. Adapt data to model: Adaptive transformation optimization for domain-shared time series foundation models. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uTK1SNgi1N.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, Sahil Garg, Alexandre Drouin, Nicolas Chapados, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023. URL https://openreview.net/forum?id=jYluzCLFDM.

Anu Simon, Andrew B Penny, Mark DeMaria, James L Franklin, Richard J Pasch, Edward N Rappaport, and David A Zelinsky. A description of the real-time hfip corrected consensus approach (hcca) for tropical cyclone track and intensity guidance. *Weather and Forecasting*, 33(1):37–57, 2018.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *Proceedings of the International conference on learning representations*, volume 2024, pp. 38626–38652, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Proceedings of the Advances in neural information processing systems*, 37:469–498, 2024b.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=Yd8eHMY1wz.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Rachel Young and Solomon Hsiang. Mortality caused by tropical cyclones in the united states. *Nature*, 635(8037):121–128, 2024.

A Appendix

A.1 Dataset Construction

This section provides additional details on case construction, data filtering, and preprocessing for the SHIPS dataset used in the main experiments.

Table 5: Case counts for the chronological data partitions and sequential filtering steps.

Partition or filter	Cases
Training, 2001–2011	2,623
Validation, 2012–2016	1,273
Test, 2017–2023	1,923
Total	5,819

A.1.1 Case Construction

We retain the native six-hour sampling interval without temporal interpolation. For each initialization, the historical context consists of nine storm-state observations at $-48, -42, \ldots, 0$ hours. To maintain an identical set of cases across the lead-time experiments, we require the storm state to be available at every six-hour interval from $+6$ through $+48$ hours. Each case is indexed by basin, storm identifier, and initialization time. A single storm may therefore contribute multiple consecutive initialization cases. Storm-state variables are required to be available throughout the historical context and required target times. Missing environmental values are retained during case selection, imputed using medians estimated from the training partition, and accompanied by corresponding missingness masks for models that use environmental inputs. Table 5 summarizes the resulting case counts.

A.1.2 Environmental Variable Definitions

Table 6 provides the definitions of the environmental variables used in the experiments. MPI deficit is derived at each historical time step as the difference between SHIPS VMPI and the observed VMAX.

Table 6: Definitions and physical grouping of the historical SHIPS environmental variables used as exogenous context.

Code	Group	Interpretation	Typical physical role
CSST	Ocean	Sea-surface temperature	Surface energy supply
COHC	Ocean	Ocean heat content	Subsurface energy reservoir
MPI deficit	Thermodynamic	VMPI minus current VMAX	Remaining intensity potential
RHMD	Thermodynamic	Mid-level relative humidity	Atmospheric moisture
SHRD	Dynamics	Deep-layer vertical wind shear	Environmental shear
D200	Dynamics	200-hPa divergence	Upper-level outflow
V850	Dynamics	850-hPa tangential wind	Environmental circulation
DTL	Geography	Distance to land	Land interaction

A.2 Additional Experimental Results

A.2.1 Visualization Analysis

Figure 4 visualizes the 6-hour predictions across all test cases. Predictions for all four state variables are concentrated around the perfect-agreement line, indicating good overall forecasting accuracy. Latitude and longitude show tighter agreement than pressure and wind, suggesting that intensity prediction is more challenging than short-range track prediction. Within the intensity variables, weaker storms, characterized by higher pressure and lower wind speed, exhibit tighter agreement,

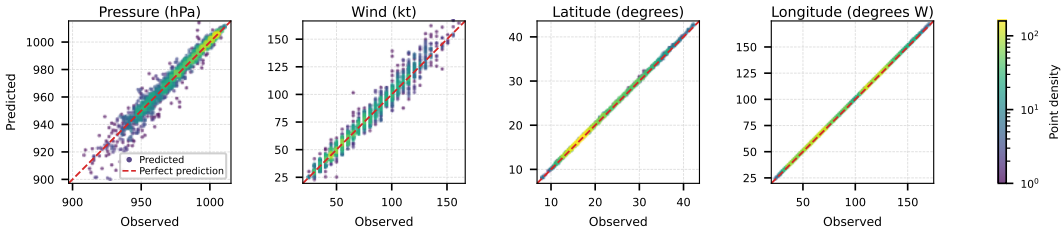


Figure 4: Visualization of the regression results between predictions and observations for the 6-hour PHYWARP-Chronos-2 experiment, including sea-level pressure, maximum wind speed, latitude prediction, and longitude prediction.

Table 7: Complete 24-hour full-state errors by basin.

Atlantic (AL) Model	Pressure		Wind		Latitude		Longitude		Distance
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Supervised time-series models									
PatchTST	9.04±0.06	14.35±0.14	13.06±0.05	20.29±0.13	1.083±0.045	1.379±0.046	1.283±0.027	1.652±0.024	193.15±4.56
iTransformer (+Env.)	9.34±0.10	15.04±0.22	13.37±0.11	21.19±0.18	0.942±0.008	1.237±0.008	1.148±0.012	1.525±0.018	169.17±0.81
TimeXer (+Env.)	8.79±0.09	13.26±0.03	12.28±0.04	17.54±0.03	0.928±0.036	1.211±0.037	1.155±0.019	1.506±0.022	169.33±4.82
TiDE	9.63±0.01	15.77±0.02	13.79±0.02	21.48±0.02	1.090±0.011	1.396±0.013	1.288±0.013	1.669±0.016	193.09±1.97
TimesNet (+Env.)	8.94±0.06	12.93±0.14	12.37±0.11	17.19±0.16	1.057±0.010	1.398±0.011	1.392±0.007	1.858±0.009	199.56±1.30
TimeMixer (+Env.)	9.81±0.27	14.12±0.37	13.56±0.21	18.69±0.27	1.150±0.044	1.503±0.053	1.540±0.059	2.043±0.059	217.01±8.18
Frozen TSFMs									
TimesFM	9.86	14.13	13.89	19.33	1.331	1.692	1.472	1.920	231.58
Moirai	10.64	15.25	14.68	20.42	2.074	2.598	3.259	3.941	452.22
Moirai (+Env.)	12.60	17.96	16.35	21.80	2.502	3.068	3.662	4.439	520.07
Chronos-2	9.29	15.81	12.82	19.84	1.111	1.418	1.270	1.685	195.11
Chronos-2 (+Env.)	9.75	17.14	13.37	21.19	1.028	1.362	1.265	1.692	186.60
PhyWarp models (ours)									
PHYWARP-TimesFM	7.70±0.03	10.85±0.03	10.96±0.13	15.05±0.15	0.987±0.006	1.297±0.005	1.199±0.014	1.614±0.021	176.88±0.60
PHYWARP-Moirai	7.90±0.14	11.21±0.19	10.98±0.15	15.31±0.24	0.973±0.006	1.300±0.007	1.167±0.015	1.572±0.027	173.24±0.83
PHYWARP-Chronos-2	**7.47±0.09**	**10.77±0.10**	**10.67±0.11**	**14.86±0.15**	**0.854±0.006**	**1.123±0.008**	**1.018±0.003**	**1.369±0.006**	**151.99±0.83**
PHYWARP-Student	7.70±0.25	11.13±0.20	10.71±0.18	14.89±0.19	0.856±0.013	1.125±0.011	1.072±0.030	1.445±0.034	155.82±3.30

Eastern North Pacific (EP) Model	Pressure		Wind		Latitude		Longitude		Distance
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Supervised time-series models									
PatchTST	10.47±0.17	14.69±0.30	15.25±0.15	20.93±0.24	0.668±0.018	0.861±0.021	1.180±0.076	1.447±0.077	160.29±6.98
iTransformer (+Env.)	10.67±0.21	15.17±0.37	15.19±0.22	21.21±0.43	0.627±0.004	0.817±0.004	0.781±0.024	1.026±0.022	119.97±2.07
TimeXer (+Env.)	8.85±0.08	12.37±0.07	12.46±0.04	16.72±0.03	0.585±0.018	0.759±0.022	0.923±0.037	1.140±0.038	129.64±4.83
TiDE	11.81±0.07	16.82±0.06	16.48±0.09	22.67±0.10	0.668±0.006	0.859±0.007	1.119±0.018	1.359±0.020	154.50±2.08
TimesNet (+Env.)	9.53±0.07	13.23±0.07	13.05±0.09	17.40±0.12	0.640±0.010	0.843±0.016	1.056±0.031	1.342±0.034	145.95±3.29
TimeMixer (+Env.)	10.56±0.29	14.49±0.35	14.44±0.53	19.17±0.72	0.745±0.043	0.956±0.049	1.115±0.065	1.382±0.095	157.87±9.05
Frozen TSFMs									
TimesFM	11.30	15.66	16.21	21.83	0.903	1.183	1.197	1.538	181.09
Moirai	11.73	15.85	17.69	23.73	1.083	1.396	4.007	4.561	465.70
Moirai (+Env.)	13.35	18.09	19.07	24.94	1.370	1.745	4.533	5.171	534.64
Chronos-2	11.47	17.32	15.02	20.87	0.696	0.914	1.052	1.328	149.84
Chronos-2 (+Env.)	12.42	19.82	16.14	23.46	0.707	0.907	0.850	1.106	132.56
PhyWarp models (ours)									
PHYWARP-TimesFM	7.97±0.13	10.95±0.19	11.20±0.52	14.94±0.75	0.657±0.004	0.847±0.003	0.775±0.006	1.017±0.008	121.70±0.36
PHYWARP-Moirai	7.91±0.24	10.70±0.35	11.33±0.36	14.88±0.71	0.608±0.002	0.777±0.002	0.775±0.009	0.982±0.010	117.16±0.69
PHYWARP-Chronos-2	**7.78±0.07**	**10.52±0.03**	11.20±0.13	**14.67±0.28**	**0.552±0.002**	0.709±0.002	0.679±0.010	0.888±0.012	105.04±0.76
PHYWARP-Student	7.96±0.07	10.90±0.09	**11.15±0.04**	14.76±0.12	0.552±0.008	**0.703±0.008**	**0.671±0.005**	**0.863±0.004**	**103.96±1.09**

whereas stronger storms show greater prediction dispersion. This pattern further highlights the increased difficulty of intensity prediction for stronger storms.

A.2.2 COMPLETE LONG-HORIZON PREDICTION RESULTS

Tables 7 and 8 provide complete basin-wise results for the 24- and 48-hour forecasting tasks. The advantages of PHYWARP become particularly evident at longer lead times: the adapted TSFMs consistently improve over their corresponding frozen backbones, while achieving the strongest overall performance for intensity and track forecasting across both basins. At 24 hours, PHYWARP-Chronos-2 achieves strong overall results in both the Atlantic and eastern North Pacific. At 48 hours, different PHYWARP variants achieve the lowest intensity errors, while the distilled student obtains the lowest track-distance error in both basins. Pressure and wind RMSE show trends broadly consistent with MAE, indicating that the gains are not limited to reductions in average error but also extend to larger forecast deviations.

Table 8: Complete 48-hour full-state errors by basin.

Atlantic (AL) Model	Pressure		Wind		Latitude		Longitude		Distance
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Supervised time-series models									
PatchTST	13.64±0.05	19.28±0.15	21.00±0.05	30.00±0.13	3.172±0.141	3.889±0.139	3.393±0.099	4.721±0.060	529.73±19.64
iTransformer (+Env.)	14.45±0.22	20.95±0.27	21.18±0.40	30.45±0.75	2.881±0.145	3.637±0.141	3.456±0.042	4.782±0.030	509.67±15.81
TimeXer (+Env.)	13.61±0.09	18.90±0.06	18.88±0.04	25.22±0.04	2.448±0.029	3.142±0.035	3.264±0.021	4.516±0.076	454.72±3.68
TiDE	14.67±0.01	21.80±0.02	21.17±0.02	30.45±0.03	3.568±0.002	4.285±0.002	4.127±0.004	5.280±0.003	624.41±0.47
TimesNet (+Env.)	13.68±0.08	19.03±0.09	18.96±0.25	25.44±0.20	2.647±0.024	3.391±0.021	3.703±0.045	5.120±0.038	507.38±5.09
TimeMixer (+Env.)	14.06±0.30	19.23±0.38	19.04±0.26	25.39±0.36	2.688±0.066	3.435±0.101	3.726±0.177	5.153±0.214	508.19±18.95
Frozen TSFMs									
TimesFM	14.42	20.13	20.91	28.43	3.802	4.639	4.198	5.468	653.19
Moirai	15.32	21.06	21.90	28.52	4.984	5.940	6.613	8.089	962.41
Moirai (+Env.)	15.75	21.54	23.82	31.82	5.248	6.224	6.621	8.073	984.23
Chronos-2	14.72	22.92	21.38	31.28	3.196	3.939	3.854	5.128	573.50
Chronos-2 (+Env.)	15.36	24.88	22.25	33.40	2.979	3.740	3.680	4.971	538.21
PhyWarp models (ours)									
PHYWARP-TimesFM	12.18±0.25	16.66±0.40	16.93±0.26	22.81±0.50	2.717±0.014	3.441±0.015	3.481±0.017	4.792±0.022	490.81±1.39
PHYWARP-Moirai	12.25±0.33	16.50±0.42	16.73±0.25	22.78±0.61	2.798±0.013	3.570±0.007	3.511±0.040	4.806±0.047	499.20±2.06
PHYWARP-Chronos-2	**11.90±0.24**	**16.31±0.46**	16.37±0.43	22.14±0.55	2.484±0.031	3.178±0.030	**3.104±0.016**	**4.362±0.011**	442.85±1.55
PHYWARP-Student	12.06±0.24	16.55±0.34	**16.28±0.34**	**22.07±0.47**	**2.428±0.031**	**3.103±0.047**	3.172±0.062	4.445±0.045	**439.97±6.55**

Eastern North Pacific (EP) Model	Pressure		Wind		Latitude		Longitude		Distance
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Supervised time-series models									
PatchTST	14.81±0.05	19.77±0.09	24.92±0.23	33.75±0.43	1.795±0.038	2.325±0.039	2.861±0.323	3.551±0.328	402.95±32.84
iTransformer (+Env.)	15.53±0.13	21.08±0.27	23.45±0.25	31.49±0.36	1.829±0.022	2.364±0.024	2.667±0.210	3.352±0.229	383.85±21.96
TimeXer (+Env.)	12.45±0.12	16.99±0.12	18.36±0.08	23.98±0.06	1.420±0.013	1.847±0.016	2.433±0.114	2.988±0.122	330.15±12.24
TiDE	17.90±0.02	25.14±0.03	26.69±0.04	36.08±0.05	1.915±0.001	2.423±0.001	4.533±0.009	5.251±0.009	567.83±0.90
TimesNet (+Env.)	12.37±0.02	17.06±0.10	18.11±0.12	24.07±0.19	1.546±0.010	2.029±0.016	2.865±0.011	3.561±0.031	377.93±2.05
TimeMixer (+Env.)	12.60±0.34	17.13±0.41	18.35±0.49	24.07±0.47	1.585±0.082	2.030±0.088	2.674±0.116	3.295±0.159	363.46±16.47
Frozen TSFMs									
TimesFM	17.04	23.33	26.18	35.16	2.227	2.875	3.869	4.639	526.17
Moirai	16.89	21.85	25.11	31.61	2.811	3.480	8.900	10.114	1052.25
Moirai (+Env.)	16.64	21.54	31.83	40.53	2.841	3.511	8.616	9.782	1025.28
Chronos-2	18.10	26.49	25.80	35.75	1.788	2.345	4.149	4.832	521.80
Chronos-2 (+Env.)	19.67	30.20	27.51	39.69	1.758	2.297	3.495	4.147	457.46
PhyWarp models (ours)									
PHYWARP-TimesFM	10.90±0.14	14.81±0.34	**15.89±0.21**	21.11±0.38	1.654±0.003	2.124±0.003	2.138±0.031	2.791±0.033	319.84±2.47
PHYWARP-Moirai	**10.56±0.32**	**14.61±0.62**	16.08±0.51	21.29±0.83	1.656±0.009	2.124±0.006	2.308±0.049	2.921±0.048	335.21±4.86
PHYWARP-Chronos-2	11.11±0.27	14.85±0.35	16.05±0.23	**21.10±0.26**	1.469±0.016	1.883±0.018	2.018±0.005	2.550±0.011	294.23±1.06
PHYWARP-Student	11.43±0.33	15.58±0.39	16.74±0.33	22.09±0.45	**1.362±0.015**	**1.752±0.032**	**1.942±0.024**	**2.442±0.030**	**279.78±3.70**

Table 9: Direct versus Tendency representations for iTransformer (+Env.) on the pooled, AL, and EP test sets.

Basin	Represent	6 h			12 h			24 h			48 h		
		P	*V*	*D*	*P*	*V*	*D*	*P*	*V*	*D*	*P*	*V*	*D*
Pooled	Direct	2.90 ± 0.05	4.04 ± 0.10	25.54 ± 1.05	5.55 ± 0.11	7.83 ± 0.13	54.46 ± 2.46	9.91 ± 0.15	14.15 ± 0.16	148.01 ± 1.24	14.91 ± 0.18	22.15 ± 0.28	455.56 ± 18.34
	Tendency	2.69 ± 0.00	3.84 ± 0.02	23.33 ± 0.37	5.48 ± 0.02	7.87 ± 0.05	57.80 ± 0.20	11.25 ± 0.05	15.73 ± 0.12	152.62 ± 0.34	21.38 ± 0.25	30.28 ± 0.05	418.85 ± 3.95
AL	Direct	2.78 ± 0.04	4.02 ± 0.07	27.89 ± 0.49	5.32 ± 0.07	7.66 ± 0.07	61.55 ± 2.81	9.34 ± 0.10	13.37 ± 0.11	169.17 ± 0.81	14.45 ± 0.22	21.18 ± 0.40	509.67 ± 15.81
	Tendency	2.59 ± 0.01	3.83 ± 0.01	26.42 ± 0.48	5.21 ± 0.04	7.69 ± 0.05	66.61 ± 0.21	10.55 ± 0.07	14.73 ± 0.15	178.50 ± 0.34	19.95 ± 0.19	27.40 ± 0.28	501.35 ± 6.45
EP	Direct	3.04 ± 0.06	4.07 ± 0.13	22.42 ± 1.80	5.85 ± 0.17	8.06 ± 0.22	45.06 ± 2.02	10.67 ± 0.21	15.19 ± 0.22	119.97 ± 2.07	15.53 ± 0.13	23.45 ± 0.25	383.85 ± 21.96
	Tendency	2.83 ± 0.01	3.86 ± 0.03	19.24 ± 0.23	5.83 ± 0.00	8.12 ± 0.07	46.13 ± 0.20	12.18 ± 0.20	17.07 ± 0.13	118.33 ± 1.18	23.27 ± 0.44	34.10 ± 0.43	309.52 ± 0.75

A.2.3 TENDENCY REPRESENTATION CONTROL

Here, we train iTransformer (+Env.) using either the Direct or Tendency representation. For Tendency, both historical storm states and forecast targets are expressed as six-hour increments, and predictions are reconstructed by cumulative summation. Environmental variables and missingness indicators remain unchanged.

Table 9 shows that Tendency improves pooled track MAE at 6 and 48 hours but degrades it at 12 and 24 hours. It also increases intensity errors at longer leads, with varying effects across basins. In contrast, PHYWARP-Chronos-2 achieves lower pooled track MAE at all four horizons (Table 2), indicating that its improvements cannot be explained by the Tendency representation alone.

A.2.4 BASIN-SPECIFIC DISTILLATION ANALYSIS

Table 10 shows that the scratch student outperforms the adapted teacher in 6-hour intensity forecasting in both basins, although its track errors are higher through 24 hours. Distillation from PHYWARP improves the scratch student's track accuracy in both basins through 24 hours while largely preserving its intensity accuracy. In contrast, Direct-KD generally degrades performance. At 48 hours, the effect becomes basin-dependent: PHYWARP-KD improves track accuracy in EP but slightly degrades it in AL relative to scratch.

Table 10: Basin-specific distillation results. Direct-KD uses frozen Chronos-2 as the teacher, whereas PHYWARP-KD uses the adapted PHYWARP-Chronos-2 teacher.

Configuration	Basin	6 h $P/V/D$	12 h $P/V/D$	24 h $P/V/D$	48 h $P/V/D$
Teacher	AL	2.65±0.00/3.74±0.04/22.60±0.02	4.43±0.02/6.50±0.01/55.83±0.16	7.47±0.09/10.67±0.11/151.99±0.83	11.90±0.24/16.37±0.43/442.85±1.55
	EP	2.80±0.02/3.68±0.02/17.61±0.05	4.61±0.05/6.50±0.04/41.02±0.10	7.78±0.07/11.20±0.13/105.04±0.76	11.11±0.27/16.05±0.23/294.23±1.06
Student (scratch)	AL	2.41±0.02/3.51±0.02/27.81±0.31	4.51±0.03/6.43±0.01/64.63±1.22	8.02±0.13/10.91±0.11/165.73±1.98	12.66±0.15/16.85±0.18/433.43±3.16
	EP	2.51±0.03/3.50±0.05/20.18±0.31	4.48±0.01/6.31±0.02/45.97±0.13	7.82±0.09/10.99±0.15/115.38±2.52	11.10±0.26/16.45±0.41/302.24±3.23
Student + Direct-KD	AL	2.38±0.02/3.50±0.01/31.07±1.61	4.48±0.05/6.51±0.06/66.03±1.15	8.16±0.08/11.57±0.12/176.32±1.61	12.72±0.08/17.90±0.24/511.38±0.98
	EP	2.63±0.06/3.56±0.02/25.45±0.54	5.03±0.08/6.82±0.07/47.35±0.48	9.49±0.08/12.98±0.10/125.40±1.77	14.20±0.08/20.43±0.13/401.16±10.23
Student + PHYWARP-KD	AL	2.32±0.01/3.50±0.02/23.40±0.23	4.34±0.02/6.23±0.01/57.39±0.91	7.70±0.25/10.71±0.18/155.82±3.30	12.06±0.24/16.28±0.34/439.97±6.55
	EP	2.49±0.01/3.49±0.01/17.67±0.25	4.42±0.05/6.21±0.10/41.19±0.18	7.96±0.07/11.15±0.04/103.96±1.09	11.43±0.33/16.74±0.33/279.78±3.70

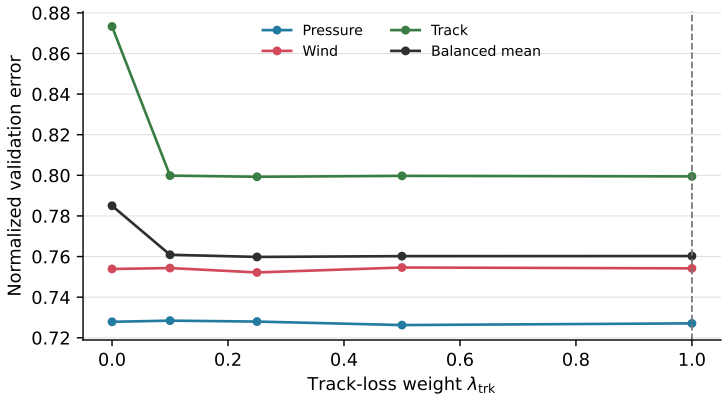


Figure 5: Sensitivity to the track-loss weight. Curves report validation errors averaged over the 6-, 12-, 24-, and 48-hour experiments; the dashed line marks the value used in all main experiments.

A.2.5 TRACK-LOSS SENSITIVITY

Figure 5 varies the track-loss weight λ_{trk} over $\{0, 0.1, 0.25, 0.5, 1.0\}$ and averages validation errors across the four horizon-specific experiments. Removing the track term substantially increases track and balanced error, whereas performance remains stable over 0.1–1.0 and pressure and wind errors change little. We use $\lambda_{\text{trk}} = 1.0$, corresponding to equal weighting of the pressure, wind, and track errors.

A.2.6 DISTILLATION-WEIGHT SENSITIVITY

Figure 6 examines the sensitivity to the distillation weight α. Teacher supervision improves mean track accuracy across all evaluated positive weights, with larger gains at shorter lead times. At $\alpha = 0.5$, track MAE decreases by 14.7%, 10.9%, 7.3%, and 1.6% at 6, 12, 24, and 48 hours, respectively. Larger weights provide additional gains for several targets and horizons but can degrade short-range intensity forecasting. In particular, relying solely on teacher supervision ($\alpha = 1$) increases both pressure and wind errors at 6 hours. We retain $\alpha = 0.5$ as a fixed, equal-weight combination of observed targets and teacher supervision across all main experiments.

A.2.7 ENVIRONMENTAL ENCODER DESIGN

Table 11 compares the grouped environmental GRUs with a single GRU over all environmental variables. Grouped encoding provides modest pressure and wind improvements from 12 to 48 hours, reaching 1.71% and 2.79% at 48 hours, respectively. The single encoder performs slightly better for track at 24 and 48 hours, with differences below 0.8%. We retain the grouped design because it preserves the physical grouping of the environmental variables while providing modest medium-to-long-range intensity improvements.

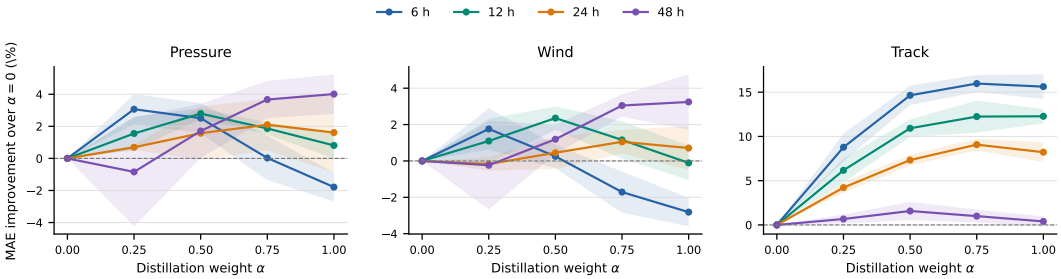


Figure 6: Sensitivity to the distillation weight α. Curves report the relative MAE improvement over the scratch student ($\alpha = 0$), with shaded regions indicating one standard deviation across three seeds. Positive values indicate improvement.

Table 11: Relative MAE reduction (%) of grouped environmental encoders over a single environmental encoder with Chronos-2. Positive values favor grouped encoding; results are computed from three-seed means.

Lead	Pressure	Wind	Distance
6 h	−0.31	−0.33	+0.12
12 h	+0.57	+0.49	+0.12
24 h	+1.02	+0.88	−0.31
48 h	+1.71	+2.79	−0.78

A.2.8 Label Efficiency

Table 12 reports the storm-level label-efficiency experiment across all four forecast horizons. Smaller label fractions are constructed as nested subsets of the full training split, with all forecast cases from each selected storm retained together. The validation and test sets remain unchanged. With 50% of the training storms, track MAE increases by at most 1.1%, while pressure and wind MAE increase by at most 5.6%. With only 5% of the training storms, the 6-hour errors remain close to the full-data results, whereas the 48-hour pressure and wind errors increase by approximately 24%. These results indicate that track adaptation is relatively label-efficient, while long-range intensity forecasting benefits more from additional labeled storms.

Table 12: Label efficiency of PHYWARP-Chronos-2. Label fractions are nested storm-level subsets of the training split, while validation and test sets remain fixed.

Labels	6 h			12 h			24 h			48 h		
	P	V	D	P	V	D	P	V	D	P	V	D
5%	2.77±0.03	3.80±0.02	20.72±0.18	5.13±0.17	7.23±0.20	51.38±1.13	9.39±0.46	13.21±0.17	137.43±3.25	14.36±0.54	20.10±0.42	397.80±11.53
10%	2.76±0.02	3.78±0.02	20.67±0.05	5.02±0.23	7.14±0.20	50.41±0.09	9.13±0.48	12.74±0.29	135.21±1.18	13.76±0.94	20.56±1.15	386.36±11.33
25%	2.76±0.03	3.75±0.03	20.57±0.11	4.93±0.30	6.93±0.21	50.59±0.41	8.42±0.79	12.11±1.20	133.93±1.14	12.71±1.38	17.85±1.47	384.48±5.55
50%	2.72±0.01	3.71±0.00	20.51±0.05	4.65±0.09	6.70±0.09	49.84±0.11	7.97±0.30	11.42±0.28	132.70±0.83	12.10±0.46	17.12±0.67	382.87±4.47
100%	2.71±0.01	3.71±0.02	20.45±0.03	4.51±0.01	6.50±0.01	49.46±0.08	7.61±0.02	10.90±0.12	131.80±0.79	11.56±0.23	16.23±0.28	378.93±1.04

A.3 IMPLEMENTATION AND REPRODUCIBILITY

This section provides implementation details omitted from the main text, including the exact view transformations, loss normalization, backbone inference procedures, and baseline configurations.

A.3.1 HEADING ESTIMATION AND VIEW RECONSTRUCTION

The Kinematic view expresses storm displacement relative to the recent motion direction. Because the data are sampled every six hours, we estimate the heading from the most recent 12-hour displacement:

$$\gamma = \operatorname{atan2}\left(y_L - y_{L-2},\, x_L - x_{L-2}\right). \tag{10}$$

The corresponding rotation matrix is

$$R(\gamma) = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix}. \tag{11}$$

For a six-hour displacement $\Delta\mathbf{r}_t = (\Delta x_t, \Delta y_t)$, the heading-relative components are

$$\begin{bmatrix} \Delta u_{\parallel,t} \\ \Delta u_{\perp,t} \end{bmatrix} = R(\gamma) \begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix}. \tag{12}$$

Here, $\Delta u_{\parallel,t}$ and $\Delta u_{\perp,t}$ describe displacement along and across the recent storm heading, respectively.

We use $\epsilon = 10^{-6}$ km to identify an undefined heading. If

$$\|\mathbf{r}_L - \mathbf{r}_{L-2}\|_2 < \epsilon, \tag{13}$$

the heading is instead estimated from the latest six-hour displacement, $\mathbf{r}_L - \mathbf{r}_{L-1}$. If this displacement also has norm below ϵ, we set $R(\gamma) = I$, corresponding to the identity rotation. The threshold is sufficiently small that it only handles stationary or numerically near-stationary cases.

For the Tendency view, predicted increments are accumulated from the latest observed state to recover absolute forecasts. For the Kinematic view, predicted heading-relative displacements are first mapped back to the local east–north coordinate system using $R(\gamma)^\top$ and then accumulated:

$$\widehat{\mathbf{r}}_h^{(K)} = \mathbf{r}_L + \sum_{\ell=1}^{h} R(\gamma)^\top \begin{bmatrix} \Delta\widehat{u}_{\parallel,\ell} \\ \Delta\widehat{u}_{\perp,\ell} \end{bmatrix}. \tag{14}$$

The heading estimated from the observed history is kept fixed throughout the forecast horizon.

A.3.2 ADAPTER INPUTS AND TRAINING OBJECTIVE

Continuous candidate descriptors and environmental inputs are normalized using statistics computed from the training partition and clipped to $[-10, 10]$. Each environmental group is encoded by a single-layer GRU with hidden size 8, yielding a 32-dimensional environmental representation after concatenating the four group outputs. Environmental inputs contain the normalized historical variables together with their missingness indicators.

For completeness, the error terms in Eq. 7 are implemented as

$$\ell_h^p = \frac{|\widehat{p}_{n,h} - p_{n,h}|}{\sigma_h^p}, \qquad \ell_h^v = \frac{|\widehat{v}_{n,h} - v_{n,h}|}{\sigma_h^v}, \qquad \ell_h^{\mathrm{trk}} = \frac{\|\widehat{\mathbf{r}}_{n,h} - \mathbf{r}_{n,h}\|_2}{\sigma_h^{\mathrm{trk}}}. \tag{15}$$

Table 13: Frozen-backbone inference configuration.

Backbone	Point forecast	Batch size	Queries/case
TimesFM	Point head	128	3
Moirai	Median of 100 samples	32	3
Chronos-2	0.5 quantile	64	3

Here, n indexes forecast cases and $\mathbf{r}_{n,h} = (x_{n,h}, y_{n,h})$ denotes the local east–north displacement in kilometers. This local-coordinate track loss is used for optimization, whereas the reported track MAE is the mean geodesic distance after converting predicted displacements back to latitude and longitude. The scales σ_h^p, σ_h^v, and σ_h^{trk} are estimated from the training errors of the uniformly averaged D/T/K candidate forecasts at prediction step h.

Gradients are clipped to a maximum norm of 5. Unless otherwise specified, trainable models are evaluated using seeds 7, 42, and 2026, and results are reported as mean and standard deviation. For each backbone and terminal horizon, the adapter is trained on the training partition, with validation error used for early stopping and checkpoint selection. The checkpoint with the best validation performance is restored before evaluation on the held-out test set.

A.3.3 Backbone Inference

Table 13 summarizes point-forecast extraction and inference settings. We use the frozen checkpoints `google/timesfm-2.5-200m-pytorch`, `Salesforce/moirai-1.1-R-small`, and `amazon/chronos-2`. Each horizon-specific experiment receives nine six-hour context steps and predicts $H \in \{1, 2, 4, 8\}$ future steps for terminal lead times of 6, 12, 24, and 48 hours, respectively. The final Direct, Tendency, and Kinematic candidate bank requires three backbone queries per case, whereas the Direct baseline requires one.

TimesFM uses input normalization, its continuous quantile head, flip invariance, and quantile-crossing repair. Chronos-2 performs joint four-variate forecasting, while Moirai uses automatic patch selection, four-dimensional targets, and 100 forecast samples. For the Naive+Env diagnostic, Chronos-2 and Moirai additionally receive the eight historical environmental variables and their missingness indicators as past-only covariates. TimesFM does not include a corresponding Naive+Env variant because its forecasting interface processes channels independently.

A.3.4 Supervised Baselines

We train a separate supervised model for each terminal forecast horizon. Each model predicts all six-hourly storm states through that horizon and is evaluated at the terminal lead. Storm-state and environmental variables are z-score normalized using statistics from the training partition. Missing environmental values are replaced by training-set medians and accompanied by binary missingness indicators. Predictions are transformed back to physical units before evaluation.

The supervised models are trained using seeds 7, 42, and 2026. Unless otherwise specified, common settings are 100 maximum epochs, early-stopping patience 15, batch size 128, learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 1, model dimension 128, eight attention heads, two encoder layers, two decoder layers when applicable, feed-forward dimension 256, and dropout 0.1. The models optimize a lead- and target-normalized mean-squared error. For TiDE, we select the learning rate from $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$ using validation data and train for up to 300 epochs with patience 30. The selected learning rate is 10^{-3} for all horizons.

PatchTST uses patch length three and stride one. TimesNet uses three dominant periods and six convolution kernels. TimeMixer uses one average-pooling downsampling layer with window size two. TimeXer applies its target–exogenous interface separately to each of the four storm-state variables. PatchTST and TiDE receive only the four storm-state channels, whereas iTransformer, TimeXer, TimesNet, and TimeMixer additionally receive the eight historical environmental variables and their missingness indicators.

For each model and seed, we first train on the training partition and use validation error to select the number of training epochs. The model is then reinitialized with the same seed and retrained on

the training partition for the selected number of epochs before evaluation on the held-out test set. Validation cases are not included in the final model fitting.

A.3.5 Student and Coordinate Details

The student receives 20 features per historical step: four storm-state variables, eight environmental variables, and eight corresponding missingness indicators. A single-layer GRU with hidden size 64 encodes the history. Its final hidden state is combined with a learned 64-dimensional prediction-step embedding and passed through a $64 \to 64 \to 4$ decoder with GELU activation and dropout of 0.1. The output layer is zero-initialized. For forecast case n and prediction step h, the student predicts

$$\widehat{\mathbf{s}}^{\mathrm{stu}}_{n,h} = \mathbf{s}_{n,L} + \mathbf{d}_h \odot S_\phi(\mathbf{h}_n + \mathbf{a}_h), \tag{16}$$

where $\mathbf{h}_n$ is the final GRU hidden state and $\mathbf{a}_h$ is the prediction-step embedding. The state-change scales are computed from the training partition as

$$d_{h,q} = \max\left\{\mathrm{Std}_{n\in\mathcal{D}_{\mathrm{train}}}\left(s_{n,L+h,q} - s_{n,L,q}\right), m_q\right\}, \qquad \mathbf{m} = (1\ \mathrm{hPa}, 1\ \mathrm{kt}, 5\ \mathrm{km}, 5\ \mathrm{km}), \tag{17}$$

where q indexes pressure, wind, and the two position coordinates.

For position variables, west-positive SHIPS longitude is converted to east-positive longitude. Relative to the forecast initialization (ϕ_0, λ_0), local east–north displacements are computed as

$$x_t = R_{\mathrm{E}} \cos(\phi_0)\, \mathrm{wrap}(\lambda_t - \lambda_0), \qquad y_t = R_{\mathrm{E}}(\phi_t - \phi_0), \tag{18}$$

where $R_{\mathrm{E}} = 6371.0088$ km is the mean Earth radius, angles are in radians, and wrap maps angles to $[-\pi, \pi)$. The inverse transformation is

$$\phi_t = \phi_0 + \frac{y_t}{R_{\mathrm{E}}}, \qquad \lambda_t = \mathrm{wrap}\left(\lambda_0 + \frac{x_t}{R_{\mathrm{E}} \cos(\phi_0)}\right). \tag{19}$$

Track MAE is the mean distance between reconstructed and observed storm-center coordinates, computed using the Vincenty inverse solution on WGS84 with a Haversine fallback upon non-convergence.