

Geo- χ -Former: A Geodynamically Informed Transformer for Short-Term Polar Motion Prediction in the Excitation Domain

Junhai Huang

Sun Yat-sen University

Boyang Zhu

Sun Yat-sen University

Xiaodong Liu

`liuxd36@mail.sysu.edu.cn`

Sun Yat-sen University <https://orcid.org/0000-0001-9329-0315>

Zhigang Wu

Sun Yat-sen University

Wei Wang

China Aerospace Science and Technology Corporation

Research Article

Keywords: Polar motion prediction, Polar motion excitation, Effective angular momentum, Transformer

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11063005/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Geo- χ -Former: A Geodynamically Informed Transformer for Short-Term Polar Motion Prediction in the Excitation Domain

Junhai Huang¹, Boyang Zhu¹, Xiaodong Liu^{1,2*}, Zhigang Wu¹, Wei Wang³

^{1*} School of Aeronautics and Astronautics, Sun Yat-sen University, Shenzhen Campus, Shenzhen, 518107, Guangdong, China .

² Shenzhen Key Laboratory of Intelligent Microsatellite Constellation, Sun Yat-sen University, Shenzhen Campus, Shenzhen, 518107, Guangdong, China .

³ China Aerospace Science and Technology Corporation, Beijing, 100048, Beijing, China .

*Corresponding author(s). E-mail(s): liuxd36@mail.sysu.edu.cn;

Abstract

Short-term polar motion (PM) prediction remains challenging because persistent oscillations dominate the observed PM series, making shorter-timescale excitation variability associated with atmospheric, oceanic, hydrological, and related geophysical processes difficult to learn directly in the coordinate domain. This paper proposes Geo- χ -Former, a geodynamically informed framework for 30-day PM prediction in the excitation domain. The method first derives the geodetic angular momentum function (GAM) from observed PM using Wilson’s discrete polar motion equation, forecasts its short and seasonal component with effective angular momentum functions (EAM) and calendar encodings, and then reconstructs PM through the corresponding discrete Wilson recurrence. This design places the learning problem in an excitation space linked to geophysical angular momentum while retaining evaluation in the observed PM domain. Geo- χ -Former combines historical excitation, EAM covariates, and calendar encodings through temporal attention and attention across variables. Using International Earth Rotation and Reference Systems Service (IERS) PM observations and EAM data from 1993 to 2025, we evaluate the model against archived IERS Bulletin A predictions and representative deep learning baselines. Over the cumulative 1–30-day forecast window, Geo- χ -Former reduces the root-mean-square error

(RMSE) relative to Bulletin A by 9.21% for $\mathbf{x}_p$ and 11.93% for $\mathbf{y}_p$. Ablation and interpretability analyses show that EAM covariates, attention across variables, temporal encoding, and the PM reconstruction loss are each associated with improved forecast skill, while attention and masking diagnostics indicate the strongest weighting and forecast sensitivity for atmospheric and oceanic angular momentum.

Keywords: Polar motion prediction, Polar motion excitation, Effective angular momentum, Transformer

1 Introduction

Polar motion (PM) is the displacement of Earth’s instantaneous rotation axis relative to the terrestrial reference frame. Accurate PM prediction is required for maintaining Earth orientation products and for applications that depend on precise transformations between terrestrial and celestial reference frames, including satellite orbit determination, positioning, navigation, and timing. PM is also a geodetic observable that integrates angular momentum exchange and mass redistribution within the Earth system (Chin et al. 2004; Gross 2015; Dill et al. 2020). PM prediction remains difficult because the pole coordinates combine persistent Chandler and annual oscillations with excitation variability from multiple geophysical processes. For short-range forecasts, the available effective angular momentum (EAM) functions mainly describe contributions from surface fluid layers and barystatic sea-level changes, including atmospheric angular momentum (AAM), oceanic angular momentum (OAM), hydrological angular momentum (HAM), and sea-level angular momentum (SLAM) (Barnes et al. 1983; Brzeziński 1992; Dill et al. 2019; Kiani Shahvandi et al. 2024). As a result, models that predict the pole coordinates directly can be dominated by persistent PM oscillations and may not fully use the physical information provided by EAM covariates.

Early PM prediction studies mainly treated the observed PM as empirical time series. Least-squares extrapolation, autoregressive modeling, and residual prediction methods were used to capture the predictable temporal structure of the pole coordinate series. Chao (1985a) showed that PM prediction can be improved by allowing the Chandler and annual wobble periods to vary with time rather than prescribing fixed periods in a harmonic model. Kosek et al. (1998) further extended this approach by combining least-squares extrapolation of the dominant deterministic terms with autocovariance prediction of the residuals. The same perspective is also reflected in operational Earth Orientation Parameters (EOP) prediction. Published descriptions of IERS Bulletin A indicate that its PM prediction procedure extrapolates the dominant deterministic variations in the observed pole coordinates, including Chandler and seasonal terms, and predicts the remaining residual variability with stochastic procedures (McCarthy and Luzum 1991; Luzum et al. 2001). As a routinely updated operational EOP prediction product, Bulletin A therefore provides an appropriate operational baseline for evaluating new PM forecasts. Nevertheless, approaches based

on pole coordinates primarily exploit the temporal structure of PM itself. Formulating the prediction problem in the excitation domain provides a more direct setting for combining geodetically inferred excitation with EAM covariates.

Beyond extrapolation of the pole coordinates, another line of research has incorporated excitation information into EOP prediction. Forecast EAM products have been introduced into short- and medium-range prediction schemes, often together with least-squares extrapolation and autoregressive modeling of the remaining residuals (Dill et al. 2019; Kong et al. 2026). These studies show that EAM covariates carry physical information relevant to forecasting and complement temporal extrapolation of pole coordinates. However, many schemes that use EAM still combine physical covariates with empirical residual models. They therefore represent the relationships among PM history, EAM sources, target axes, and forecast leads only partly and adaptively. Data-driven models may learn how EAM covariates relate to geodetically inferred excitation and how this information combines with the temporal memory of PM.

Machine learning methods have recently been introduced to PM prediction and the forecasting of other EOP components to better represent nonlinear temporal dependence and multiscale variability in the observed series. Recurrent, convolutional, attention-based, Transformer-based, and hybrid neural architectures have been applied to these tasks, with reported improvements over classical statistical baselines or empirical hybrid models (Yu et al. 2023; Bakhtiari et al. 2024; Wang et al. 2025; Zhu et al. 2026). These models are capable of capturing nonlinear sequential dependencies, extracting localized temporal patterns, and adaptively assigning weights to historically informative states or features. In parallel, geodynamically informed learning strategies have incorporated physical covariates and process knowledge, including EAM functions, tidal terms, neural differential formulations, and physics-informed analyses of surface, mantle, and core contributions (Kiani Shahvandi et al. 2022, 2023, 2024). Recent studies have also explored formulations in the excitation domain by using the Liouville relationship to connect PM, geodetic excitation, and EAM information (Kong et al. 2023). However, most existing machine-learning approaches still forecast directly in the PM coordinate domain, while excitation-related information such as EAM is typically introduced only as auxiliary covariates or residual terms. Whether excitation-domain modeling can provide consistent performance gains across different forecasting architectures, and how multi-source geophysical information can be effectively fused to further improve PM prediction, remain insufficiently studied.

In this study, we develop Geo- χ -Former, a geodynamically informed Transformer framework for PM forecasting in the excitation domain. Observed PM is the coordinate-domain manifestation of the Earth’s rotational response to external and internal excitation, and its time series therefore contains persistent Chandler wobble, annual variability, and shorter-timescale excitation responses. Direct forecasting in the PM coordinate domain requires a model to represent these persistent oscillations while simultaneously identifying weaker short-term excitation variability, coupling information across multiple timescales. To make the forecast target more directly reflect the excitation that drives PM variability, we transform observed PM into the geodetic angular momentum function (GAM), defined here as the dynamically equivalent excitation inferred from the pole coordinates, and formulate the forecasting task in the

excitation domain. This gives the forecast target a physical representation more consistent with EAM and provides a structured modeling space for jointly using GAM history and EAM information. A Transformer architecture is adopted because the forecast problem involves temporal dependencies within each variable and cross-source relationships among excitation, EAM, and calendar information, with contributions that may vary by source, axis, and forecast lead.

Geo- χ -Former jointly models GAM history, EAM covariates, and calendar encodings through temporal attention and attention across variables, and predicts the short and seasonal components of GAM. The predicted excitation is subsequently reconstructed into PM through the corresponding dynamical relationship, so that model training is conducted in the excitation domain while the final forecast is returned to the observable PM coordinate domain for evaluation. Forecasting GAM allows EAM information to enter the model in the same excitation-domain representation as the prediction target. Reconstructing PM from the predicted GAM permits direct comparison with existing PM prediction methods and operational products. We evaluate the framework using multidecadal PM and EAM data, compare it with IERS Bulletin A and representative deep learning baselines, and further use ablation experiments and diagnostic analyses to examine the contributions of excitation-domain modeling, physical covariates, and key model components.

This paper is organized as follows. Section 2 describes the Wilson-based excitation framework, forecast sample construction, Geo- χ -Former architecture, and training objective. Section 3 introduces the PM and EAM products, including EAM forecasts used as future inputs. Section 4 presents the forecast evaluation and ablation results. Section 5 discusses the role of EAM covariates, model diagnostics, and limitations. Section 6 summarizes the main findings.

2 Methodology

2.1 Excitation Derivation and PM Reconstruction

The observed PM series contains prominent quasi-periodic variability, especially the Chandler wobble with a period of approximately 433 days and the annual oscillation with a period close to one year (Gross 2015). Because PM is the integrated rotational response of the Earth system, direct prediction of the pole coordinates can be dominated by these persistent oscillations and may obscure excitation signals at shorter time scales associated with atmospheric, oceanic, hydrological, and barystatic sea-level processes. We therefore formulate the forecasting problem in the excitation domain and reconstruct PM after the future excitation has been predicted. The complete workflow of the proposed framework is summarized in Figure 1.

The theoretical connection between PM and equatorial excitation is given by the linearized Liouville equation. Following the conventional complex representation used in PM excitation studies, the two equatorial PM coordinates and excitation components are written as

$$p(t) = x_p(t) - iy_p(t), \quad \chi(t) = \chi_x(t) + i\chi_y(t), \quad (1)$$

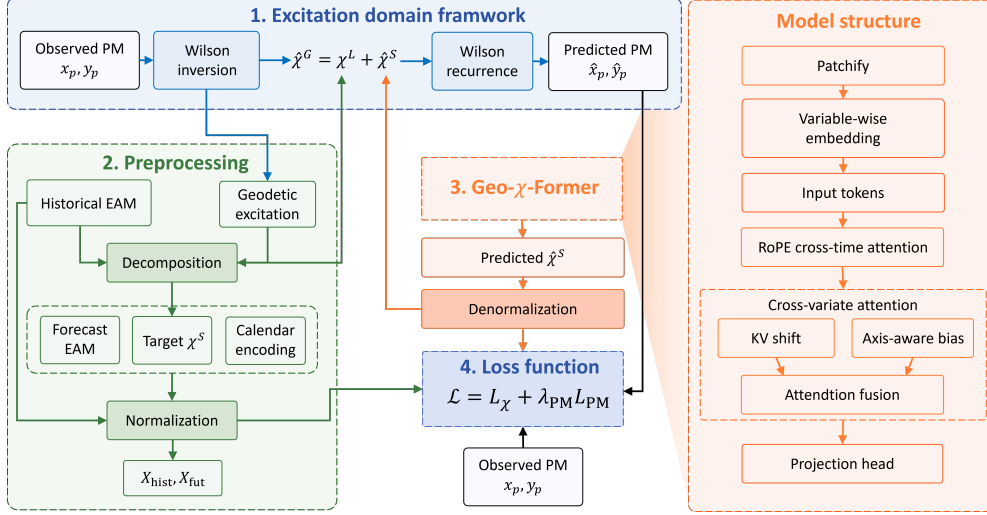


Fig. 1 Schematic of the Geo- χ -Former workflow for PM forecasting. Observed PM coordinates are first converted to GAM by Wilson inversion. The inferred GAM is separated into a slowly varying background χ^L and a short and seasonal component χ^S , while EAM covariates and calendar encodings are constructed and normalized as model inputs. Geo- χ -Former predicts $\hat{\chi}^S$ from the historical inputs and the covariates available over the forecast horizon. The prediction is then denormalized, combined with the background component, and reconstructed to predicted PM coordinates using the Wilson recurrence. The training objective combines an excitation loss with a PM reconstruction loss.

where $i = \sqrt{-1}$ is the imaginary unit; $x_p(t)$ and $y_p(t)$ denote the observed PM coordinates, and $\chi_x(t)$ and $\chi_y(t)$ denote the corresponding equatorial excitation components. Under this convention, the excitation inferred from geodetic PM observations is referred to here as the GAM. GAM represents the dynamically equivalent excitation implied by the observed PM, rather than an independently observed external forcing. GAM is the geodetically inferred prediction target, whereas EAM products are used as physically interpretable covariates (Brzeziński 1992; Gross 1992; Chao 1985b).

In continuous form, the linearized Liouville equation relates PM to equatorial excitation as (Gross 2015)

$$\chi(t) = p(t) + \frac{i}{\sigma_{cw}} \frac{dp(t)}{dt}, \quad (2)$$

where σ_{cw} is the complex Chandler angular frequency,

$$\sigma_{cw} = \frac{2\pi}{T_{cw}} \left(1 + \frac{i}{2Q_{cw}} \right). \quad (3)$$

Here, T_{cw} and Q_{cw} denote the Chandler wobble period and quality factor, respectively. In this study, the daily discrete implementation uses $\Delta t = 1$ day, $T_{cw} = 433$ days, and $Q_{cw} = 179$ (Gross 2015). With time expressed in days, σ_{cw} is expressed in radians per day.

Because both PM observations and EAM products are provided as discrete time series, we adopt the discrete PM equations of Wilson to estimate GAM from observed PM and to integrate predicted excitation back to PM (Wilson 1985). With the sign convention in Eq. (1), the GAM at epoch t is estimated from three consecutive PM samples as

$$\chi^G(t) = \frac{i}{2\sigma_{cw}\Delta t} \exp\left(\frac{i\pi\Delta t}{T_{cw}}\right) \left[p(t + \Delta t) + (1 - \exp(i\sigma_{cw}\Delta t))p(t) - \exp(i\sigma_{cw}\Delta t)p(t - \Delta t) \right]. \quad (4)$$

This operation transforms the observed PM series into the GAM series used to construct the learning target. Because the three-point inversion estimates excitation at the central epoch, no GAM value is obtained for the first or last PM sample. The PM and GAM series are therefore restricted to their common valid epochs before sample construction.

After the model predicts the future excitation sequence, the corresponding Wilson forward recurrence is used to reconstruct PM:

$$\hat{p}(t) = -\frac{i\sigma_{cw}\Delta t}{2} \exp\left(\frac{i\pi\Delta t}{T_{cw}}\right) [\hat{\chi}(t) + \hat{\chi}(t - \Delta t)] + \exp(i\sigma_{cw}\Delta t)\hat{p}(t - \Delta t). \quad (5)$$

The recurrence begins at $t = t_0 + \Delta t$ and is initialized using the observed PM and inferred GAM at the forecast initialization epoch, such that $\hat{p}(t_0) = p(t_0)$ and $\hat{\chi}^G(t_0) = \chi^G(t_0)$. Subsequent PM values are reconstructed recursively from the predicted excitation sequence. Wilson inversion maps observed PM to GAM, and the forward recurrence maps predicted GAM back to PM. The neural network learns the evolution of geodetic excitation, and the final forecast is evaluated in the directly observable PM domain.

2.2 Preprocessing

Let t_0 denote the forecast initialization epoch, L the length of the historical window, and H the forecast horizon. The main 30-day forecasting experiment uses $L = 360$ days and $H = 30$ days. After the Wilson inversion, the raw GAM is denoted by $\chi^G = [\chi_x^G, \chi_y^G]^T$. The forecasting task is formulated for the GAM component containing short-period and seasonal variability rather than for the PM coordinates.

$$\hat{\chi}_{t_0+1:t_0+H}^S = f_\theta(\mathbf{X}_{t_0}^{\text{hist}}, \mathbf{X}_{t_0}^{\text{fut}}), \quad (6)$$

where f_θ denotes Geo- χ -Former parameterized by the learnable parameters θ ; $\chi^S = [\chi_x^S, \chi_y^S]^T$ is the GAM component used as the model target; and $t_0 + 1 : t_0 + H$ denotes the sequence of forecast epochs from lead day 1 through lead day H . The two equatorial components are predicted jointly because they are coupled through the complex PM excitation representation and may share common geophysical forcing.

For each forecast initialization, the model input is divided into historical inputs and covariates available during the forecast period:

$$\begin{aligned}\mathbf{X}_{t_0}^{\text{hist}} &= [\boldsymbol{\chi}_{t_0-L+1:t_0}^S, \mathbf{E}_{t_0-L+1:t_0}, \mathbf{C}_{t_0-L+1:t_0}], \\ \mathbf{X}_{t_0}^{\text{fut}} &= [\mathbf{E}_{t_0+1:t_0+H}^f, \mathbf{C}_{t_0+1:t_0+H}].\end{aligned}\quad (7)$$

Here, $\mathbf{E}$ denotes EAM covariates, including the equatorial components of atmospheric angular momentum (AAM), oceanic angular momentum (OAM), hydrological angular momentum (HAM), and barystatic sea-level angular momentum (SLAM). These variables represent modeled atmospheric, oceanic, hydrological, and sea-level excitation sources. The calendar term $\mathbf{C}$ is constructed from day-of-year Fourier encodings with periods of one year, one-half year, one-third year, and one-quarter year, and can be extended into the forecast horizon deterministically. The notation $\mathbf{E}^f$ denotes EAM covariates over the forecast horizon. In operational settings, $\mathbf{E}^f$ is taken from EAM forecast products indexed by issue date and lead time. In sensitivity experiments, analyzed future EAM fields may also be used to examine the potential contribution of future excitation information. Neither future PM nor future GAM values are supplied to the model, and GAM remains the prediction target.

The raw GAM contains a slowly varying background together with excitation variability at shorter time scales. In addition, the low-frequency behavior of GAM can differ persistently from the summed EAM functions produced by geophysical models. Such differences may include contributions from longer-term processes, including core-mantle coupling, ice-sheet mass loss, glacial isostatic adjustment, and secular or decadal mass redistribution (Kiani Shahvandi et al. 2024). Because the forecast target in this study is short and seasonal PM variability, Geo- χ -Former estimates and removes a slowly varying background associated with the difference between GAM and the summed EAM functions before training. The extrapolated background is restored after prediction and before Wilson reconstruction.

In the implemented decomposition, the total EAM excitation is first formed by summing the selected atmospheric, oceanic, hydrological, and sea-level angular momentum components along each equatorial axis,

$$\mathbf{E}^{\text{sum}}(t) = \sum_{m \in \mathcal{M}} \mathbf{E}^m(t), \quad \mathcal{M} \subseteq \{\text{AAM}, \text{OAM}, \text{HAM}, \text{SLAM}\}. \quad (8)$$

where m indexes the EAM source, $\mathbf{E}_m(t) = [E_{m,x}(t), E_{m,y}(t)]^\top$, and $\mathcal{M}$ denotes the set of sources included in a given experiment. The gap between GAM and EAM is then defined as

$$\mathbf{g}(t) = \boldsymbol{\chi}^G(t) - \mathbf{E}^{\text{sum}}(t). \quad (9)$$

For a forecast initialized at epoch t_0 , the background is estimated separately for each component $a \in \{x, y\}$ using only the historical window $t = t_0 - L_b, \dots, t_0 - 1$, where L_b is the background estimation window. The default implementation uses $L_b = 1095$

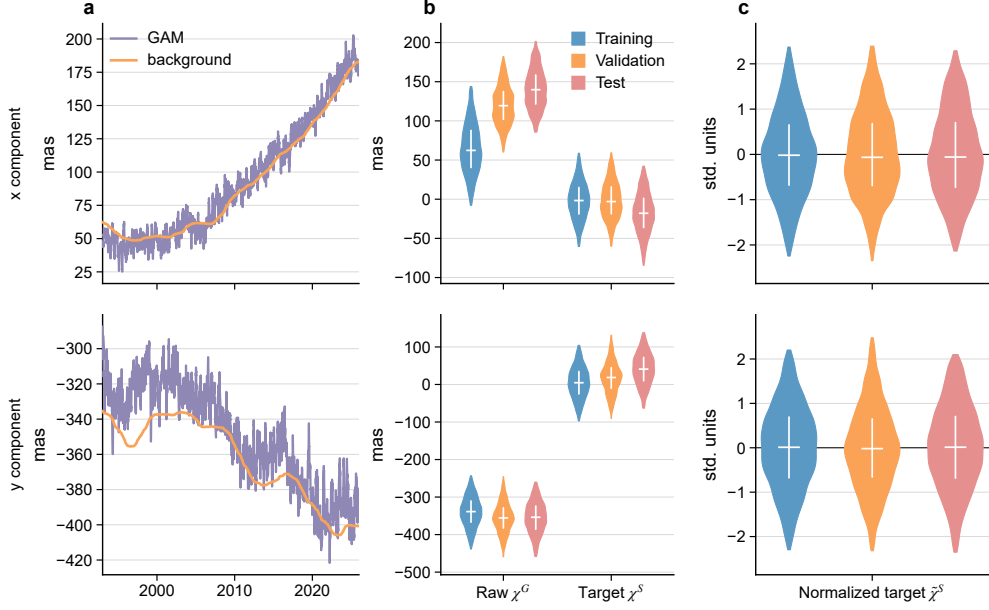


Fig. 2 Background removal and target normalization for the excitation domain forecast target. (a) Inferred GAM χ^G and the retained slowly varying background for the x and y components. (b) Removing the background reduces the offset between the training, validation, and test distributions, yielding the residual short and seasonal target χ^S . (c) Sample-wise normalization further places the target $\tilde{\chi}^S$ on a comparable scale across forecast samples. Upper and lower rows show the x and y components, respectively. Violin plots indicate sample distributions, with white markers showing the median and interquartile range.

days. For each component $a \in \{x, y\}$, a local regression is fitted to the gap,

$$g_a(t) = \beta_{a,0} + \beta_{a,1}\tau(t) + \sum_{P \in \mathcal{P}_b} \left[\alpha_{a,P} \sin\left(\frac{2\pi\tau(t)}{P}\right) + \gamma_{a,P} \cos\left(\frac{2\pi\tau(t)}{P}\right) \right] + \varepsilon_a(t), \quad (10)$$

where $\tau(t) = t - t_0$ is the relative time index, $\beta_{a,0}$ and $\beta_{a,1}$ are the local intercept and linear-trend coefficients; $\alpha_{a,P}$ and $\gamma_{a,P}$ are the sinusoidal coefficients associated with period P ; $\varepsilon_a(t)$ is the regression residual; and $\mathcal{P}_b = \{365.25, 182.625\}$ days denotes the annual and semiannual periods included in the local fit. A robust Huber regression is used to reduce the influence of transient outliers (Sun et al. 2020). Only the intercept and linear term are retained as the slowly varying background,

$$\hat{\chi}_a^L(t | t_0) = \hat{\beta}_{a,0} + \hat{\beta}_{a,1}\tau(t), \quad (11)$$

whereas the annual and semiannual terms are treated as nuisance terms to prevent seasonal variability from being absorbed into the estimated background. The same linear background is extrapolated into the forecast horizon using coefficients estimated only from samples before t_0 .

The short and seasonal excitation target is

$$\chi_a^S(t) = \chi_a^G(t) - \hat{\chi}_a^L(t | t_0), \quad a \in \{x, y\}. \quad (12)$$

After prediction, the model output is restored to full GAM by adding back the extrapolated background,

$$\hat{\chi}_a^G(t) = \hat{\chi}_a^S(t) + \hat{\chi}_a^L(t | t_0). \quad (13)$$

The recovered excitation $\hat{\chi}^G$ is then converted back to PM using Eq. (5).

Figure 2 also shows that background removal reduces the offset between the training and test distributions in raw GAM. The remaining target, however, still has scale differences across components and samples. For each forecast sample and axis $a \in \{x, y\}$, the normalization statistics are computed only from the historical target sequence $\chi_a^S(t_0 - L + 1 : t_0)$. Let μ_a and σ_a denote its mean and standard deviation, respectively. A target or EAM variable $z_a(t)$ along the same axis is normalized as

$$\tilde{z}_a(t) = \frac{z_a(t) - \mu_a}{\sigma_a + \epsilon}, \quad a \in \{x, y\}. \quad (14)$$

where $\epsilon = 10^{-6}$ is a numerical-stability constant. All EAM variables along the x axis use the statistics of χ_x^S , and all EAM variables along the y axis use the statistics of χ_y^S . Calendar encodings are not normalized using excitation statistics because they are bounded Fourier features.

2.3 Geo- χ -Former Model Structure

Building on the sample construction described above, Geo- χ -Former forecasts the short and seasonal GAM component. The model separates temporal evolution within each variable from interactions between the target excitation and the covariates. It first applies variable-wise patch embedding and then stacks Geo- χ -Former layers composed of RoPE-based temporal attention and structure-aware attention across variables. Only the target excitation tokens are projected into the final forecast; EAM and calendar tokens condition the target representation.

2.3.1 Variable-Wise Patch Embedding

Following the patching strategy of PatchTST (Nie et al. 2023), each input variable is segmented into non-overlapping temporal patches. Let $\mathbf{X} \in \mathbb{R}^{B \times L \times C}$ denote a batch of input sequences, where B is the batch size, L is the sequence length, and C is the number of variables. Given patch length P , each variable is divided into $N = \lceil L/P \rceil$ patches. For the c -th variable and n -th patch,

$$\mathbf{Z}_{n,c} = \mathbf{W}_p \mathbf{X}_{n,c}^{(P)} + \mathbf{b}_p, \quad (15)$$

where $\mathbf{X}_{n,c}^{(P)} \in \mathbb{R}^P$ is the local temporal segment, $\mathbf{W}_p$ and $\mathbf{b}_p$ are the learnable projection matrix and bias, and the resulting latent dimension is D . The tensor

$\mathbf{Z} \in \mathbb{R}^{B \times N \times C \times D}$ preserves the variable dimension, avoiding premature mixing of GAM, EAM components, and calendar encodings.

A learnable source identity embedding is added to each projected patch token:

$$\mathbf{H}_{n,c} = \mathbf{Z}_{n,c} + \mathbf{e}_c, \quad (16)$$

where $\mathbf{e}_c \in \mathbb{R}^D$ is the embedding for variable c . This allows all variables to share a latent representation space while retaining source identity.

2.3.2 Temporal Attention With RoPE

The temporal attention sublayer models dependencies over time within each variable independently. Given $\mathbf{H} \in \mathbb{R}^{B \times N \times C \times D}$, the tensor is reshaped to $\mathbb{R}^{(BC) \times N \times D}$, so attention is computed along the patch sequence for each variable. Query, key, and value projections follow the standard Transformer form, with Rotary Position Embedding (RoPE) applied to the query and key vectors before scaled dot-product attention (Vaswani et al. 2017; Su et al. 2024):

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}} + \mathbf{M} \right) \mathbf{V}. \quad (17)$$

Here, d_h is the head dimension and $\mathbf{M}$ is a causal mask. The mask prevents a patch from attending to future target patches. RoPE supplies a relative temporal inductive bias for excitation sequences that contain persistence, seasonal variability that depends on phase, and lagged geophysical responses. RoPE is applied only in the temporal attention sublayer; the sublayer that attends across variables operates at a given patch step and does not use temporal position rotations.

2.3.3 Structure-Aware Attention Across Variables

After temporal attention, the target excitation, observed EAM, and deterministic covariate streams are represented as patch tokens with temporal context. The next sublayer models interactions among variables at each patch step while retaining source identity (Zhang and Yan 2023; Liu et al. 2024; Lim et al. 2021). It incorporates the availability of covariates during the forecast period and separate representations for the two axes.

For each patch step i , target excitation tokens are used as queries. The key-value set contains target tokens, observed EAM tokens, and deterministic known covariate tokens. Let $\mathbf{H}_i^X$ denote target excitation tokens, $\mathbf{H}_i^E$ observed EAM tokens, and $\mathbf{H}_i^K$ known covariate tokens. The attention inputs are

$$\mathbf{Q}_i = \mathbf{H}_i^X \mathbf{W}_Q, \quad \mathbf{K}_i = [\mathbf{H}_i^X, \mathbf{H}_i^E, \mathbf{H}_i^K] \mathbf{W}_K, \quad \mathbf{V}_i = [\mathbf{H}_i^X, \mathbf{H}_i^E, \mathbf{H}_{i+1}^K] \mathbf{W}_V. \quad (18)$$

The key-value (KV) shift is applied only to the value branch of deterministic known covariates, such as future calendar encodings, whose values are available over the forecast horizon. It conditions prediction of the next target patch on deterministic

phase information while preserving the input boundary at forecast time (Yamaguchi et al. 2026).

EAM inputs are treated as covariate tokens rather than deterministic known tokens. Historical analyzed EAM is used as an observed covariate in the historical window. When EAM products are available during the forecast period in a given experiment, they are appended to the same EAM covariate stream as auxiliary inputs. They are not processed by the value branch used for deterministic future covariates, and the model does not predict EAM itself.

The output of attention across variables is

$$\mathbf{A}_i = \text{softmax} \left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_h}} + \mathbf{B}_{\text{axis}} \right), \quad \mathbf{O}_i = \mathbf{A}_i \mathbf{V}_i \mathbf{W}_O, \quad (19)$$

where $\mathbf{B}_{\text{axis}}$ is an optional axis-aware bias. This bias introduces a weak prior by encouraging χ_x to attend more strongly to EAM components along the x axis and χ_y to components along the y axis.

In this sublayer, only the target excitation tokens form queries. EAM and calendar tokens provide contextual keys and values, so the layer updates the hidden states of the target excitation while using the covariates only as auxiliary inputs. The covariate tokens are not treated as forecast outputs. Standard Transformer operations, including residual connections, normalization, feed-forward layers, and the autoregressive rollout procedure, are summarized in Appendix B.

2.4 Loss Function and Training Strategy

Although Geo- χ -Former predicts the short and seasonal GAM component, the final objective is accurate PM forecasting. The training objective therefore combines a loss in the excitation domain with a PM reconstruction loss. The normalized excitation prediction is returned to the physical excitation scale, and the extrapolated background is restored. The recovered full GAM is then integrated using the Wilson forward recurrence in Eq. 5 to reconstruct PM.

Let $H = 30$ be the forecast horizon and $H_e \leq H$ the effective horizon used at training epoch e . The total loss is

$$\begin{aligned} \mathcal{L}^{(e)} &= \mathcal{L}_\chi^{(e)} + \lambda_{\text{PM}} \mathcal{L}_{\text{PM}}^{(e)} \\ &= \mathcal{H}_\delta \left(\hat{\chi}_{1:H_e}^G, \chi_{1:H_e}^G \right) + \lambda_{\text{PM}} \mathcal{H}_\delta \left(\hat{\mathbf{p}}_{1:H_e}, \mathbf{p}_{1:H_e} \right), \end{aligned} \quad (20)$$

where $\mathcal{L}_\chi^{(e)}$ constrains the GAM predictions on the physical scale after inverse normalization, $\mathcal{L}_{\text{PM}}^{(e)}$ constrains the PM trajectory reconstructed from the predicted GAM, and λ_{PM} controls the relative contribution of the reconstructed PM loss. Here, $\mathcal{H}_\delta$ denotes the Huber loss with threshold δ , averaged over the batch, effective forecast epochs, and two equatorial components. The Wilson integration is initialized using the observed PM and inferred GAM at t_0 . During training, a curriculum progressively extends the effective forecast horizon by patch. A robust Huber loss (Sun et al. 2020), AdamW optimization (Loshchilov and Hutter 2019), gradient clipping, and

early stopping according to validation performance are used in the implementation. Additional implementation and training details, including the Huber loss definition and curriculum formulation, are provided in Appendix B, and the experimental and training settings explicitly specified here are summarized in Table B.1.

3 Data

This section describes the data used for PM forecasting in the excitation domain. The workflow uses three types of data. First, an observed PM record provides the reference trajectory from which GAM is inferred and against which reconstructed PM forecasts are evaluated. Second, EAM products provide physically interpretable covariates associated with the atmosphere, ocean, land hydrology, and barystatic sea-level variability. Third, EAM arrays over the forecast horizon provide archived or retrospectively constructed covariates under the settings defined below.

3.1 Data Sources

The observational PM reference is the International Earth Rotation and Reference Systems Service (IERS) EOP 20u24 C04 series. The C04 series is the combined IERS Earth orientation solution and is consistent with the International Terrestrial Reference Frame 2020 (ITRF2020). It provides daily pole coordinates, UT1–UTC, celestial pole offsets, length of day, and formal errors. The pole coordinates combine space-geodetic solutions from very long baseline interferometry (VLBI), satellite laser ranging (SLR), global navigation satellite systems (GNSS), including the Global Positioning System (GPS), and Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS) (Bizouard et al. 2019). This study uses the daily x_p and y_p pole coordinates, converts them from arcseconds to milliarcseconds (mas), and applies Wilson inversion to derive the aligned GAM target.

The geophysical covariates are based on GFZ Earth System Modelling EAM products. We use AAM, OAM, HAM, and SLAM in the two equatorial directions because these components excite PM (Chin et al. 2004; Gross 2015). For AAM, OAM, and HAM, the mass and motion terms are summed for each equatorial component to obtain the total excitation associated with that source. Because SLAM contains only a mass term, it is used directly as provided. Before alignment with the GAM target derived from PM, the native EAM records are mapped to the daily UTC grid of the PM series. For subdaily EAM records, the daily value is formed from the eight samples centered on the target daily epoch using the symmetric weights [1, 1, 3, 4, 5, 4, 3, 1] normalized by their sum (Kong et al. 2023). All EAM components are then converted to milliarcseconds using the same excitation unit convention as the GAM target. Appendix A gives the detailed unit conversion and summation convention.

3.2 Forecast Data

Archived IERS Bulletin A forecasts, issued by the IERS Rapid Service/Prediction Centre, are used as the operational reference (Luzum et al. 2001). Bulletin A is not

supplied to Geo- χ -Former as an input. Bulletin A forecasts are compared with the PM reconstructed from predicted GAM.

EAM information available during the forecast period is handled separately from the analyzed historical EAM record. Historical EAM covariates are obtained from the aligned GFZ record. The forecast EAM archive follows the ETH Zurich Geodetic Prediction Center (GPC) product, which provides daily 14-day forecasts of AAM, OAM, HAM, and SLAM by using machine learning methods to correct and extend GFZ EAM forecasts (Shahvandi et al. 2022). In the local archive used here, forecast EAM is indexed by issue date, valid date, and lead day. It contains lead days 0 to 13, issue dates from 2021-05-20 to 2025-12-08, and valid dates through 2025-12-21.

The model uses the same input form during training and evaluation. If an archived GPC EAM forecast is available, the matching issue-date and lead-day value is used directly. Because the 30-day forecast horizon extends beyond the 14-day coverage of the GPC EAM forecasts, archived forecast EAM values are used only for the covered lead days. The remaining unavailable covariate entries are filled with zeros.

For earlier epochs without archived forecasts, the unavailable EAM sequence over the forecast period is approximated using perturbations calibrated separately by component and forecast lead from the 2021 to 2025 forecast-analysis overlap; the calibration equations and fitted parameters are provided in Appendix C and Table C.1. This construction gives earlier historical training samples the same input structure as later samples with real GPC forecasts. It is not an additional observational data set and is not treated as perfect future EAM information; rather, it approximates the forecast EAM information unavailable at earlier epochs to maintain consistent input conditions throughout training.

3.3 Data Processing and Sample Construction

The assembled PM and EAM table spans 1976-01-01 to 2025-12-31 before the end-point trimming required by Eq. (4). All variables are aligned on a daily UTC grid, sorted chronologically, deduplicated, and converted to floating-point values. Isolated missing values are linearly interpolated, followed by forward or backward filling at the boundaries when necessary.

Although EOP C04 provides PM estimates before 1993, GAM diagnostics show clear distortions in the early part of the record for the present learning task in the excitation domain. We retain samples initialized from 1993-01-01 onward. This cutoff is a preprocessing choice made for this task rather than a general rejection of earlier EOP data. It also aligns with the transition toward routine high-precision GNSS contributions to EOP determination (Bizouard et al. 2019).

The retained experiment period spans 1993-01-01 to 2025-12-31. The training split covers 1993-01-01 to 2015-12-31, the validation split covers 2016-01-01 to 2020-01-01, and the later issue dates starting on 2020-01-02 are assigned to the test split. Forecast-window start dates are generated with a fixed stride, and only samples with complete future labels are retained.

For each forecast issue epoch t_0 , a moving time window defines the model input and target. The main experiment uses a look-back length of $L = 360$ days and a forecast horizon of $H = 30$ days. The historical window contains short and seasonal

GAM, historical EAM covariates, and calendar encodings. The forecast window contains GAM labels and only those future covariates allowed by the experiment. Day of year Fourier features are deterministic calendar encodings and are available throughout the forecast window. Future EAM is included only in experiments using archived GPC values, retrospectively constructed forecasts, or future analyses used for sensitivity tests. This sampling design implements the input and target convention defined above and is consistent with the target excitation, EAM covariates, and deterministic calendar streams used by Geo- χ -Former.

4 Results

4.1 Experimental Design and Evaluation Metrics

All forecast errors are computed in the PM domain using observed PM from the IERS EOP 20u24 C04 series. The main experiment constructs samples with the sliding window in Section 3.3, using $L = 360$ days and $H = 30$ days.

The evaluated methods include IERS Bulletin A, DLinear, LSTM, PatchTST, and Geo- χ -Former (Hochreiter and Schmidhuber 1997; Zeng et al. 2023; Nie et al. 2023). Bulletin A is the archived operational reference. Each deep learning baseline is evaluated in two paired settings. In the PM setting, the model directly predicts future pole coordinates. In the excitation setting, the same baseline predicts the future short and seasonal excitation component χ^S , restores the low-frequency background, and reconstructs PM with Eq. (5). This paired design compares the combined pipeline in the excitation domain with direct prediction in the PM domain while keeping the model architecture fixed.

Let $\mathbf{p}_{n,h} = [x_{p,n,h}, y_{p,n,h}]^T$ denote the reference PM at lead day h for forecast issue n , and let $\hat{\mathbf{p}}_{n,h}$ denote the corresponding prediction. Forecast accuracy is evaluated with the PM component errors $e_{n,h}^c = \hat{p}_{n,h}^c - p_{n,h}^c$, where $c \in \{x, y\}$. Let $\mathcal{H}$ denote the set of lead days and $\mathcal{C} \subseteq \{x, y\}$ the set of components. For a single lead, $\mathcal{H} = \{h\}$; for statistics aggregated over a window, $\mathcal{H} = \{1, \dots, H_e\}$ with $H_e \in \{1, 7, 10, 14, 30\}$. Forecast accuracy is evaluated using the RMSE and mean absolute error (MAE). They are defined as

$$\text{RMSE}_{\mathcal{H}}^{\mathcal{C}} = \sqrt{\frac{1}{|\mathcal{C}|N|\mathcal{H}|} \sum_{c \in \mathcal{C}} \sum_{n=1}^N \sum_{h \in \mathcal{H}} \left(e_{n,h}^c\right)^2}, \quad (21)$$

$$\text{MAE}_{\mathcal{H}}^{\mathcal{C}} = \frac{1}{|\mathcal{C}|N|\mathcal{H}|} \sum_{c \in \mathcal{C}} \sum_{n=1}^N \sum_{h \in \mathcal{H}} |e_{n,h}^c|. \quad (22)$$

Here, N is the number of forecast issue epochs. All PM errors are reported in milliarcseconds.

To quantify relative performance with respect to Bulletin A, we further use a skill score with respect to Bulletin A, defined as

$$\text{SS}_{\mathcal{H}} = 1 - \frac{\text{RMSE}_{\mathcal{H}}^{\text{model}}}{\text{RMSE}_{\mathcal{H}}^{\text{BulletinA}}}. \quad (23)$$

$SS_{\mathcal{H}} > 0$ denotes lower RMSE than Bulletin A, whereas $SS_{\mathcal{H}} < 0$ denotes higher RMSE.

4.2 Polar Motion Prediction Performance

Table 1 Forecast accuracy for the x_p component of polar motion over selected cumulative forecast windows.

Model	Forecast window (days)				
	1	7	10	14	30
<i>RMSE (mas)</i>					
Bulletin A	0.055	1.462	2.142	2.967	5.766
DLinear-PM	0.443	2.055	2.754	3.607	6.479
DLinear- χ	0.090	1.521	2.198	2.992	5.615
LSTM-PM	7.335	7.064	7.189	7.379	8.765
LSTM- χ	0.311	1.951	2.544	3.246	5.417
PatchTST-PM	1.675	2.830	3.341	4.025	6.452
PatchTST- χ	0.160	1.596	2.246	3.030	5.663
Geo- χ -Former	0.100	1.474	2.116	2.833	5.235
<i>MAE (mas)</i>					
Bulletin A	0.043	0.965	1.449	2.036	4.073
DLinear-PM	0.360	1.450	1.956	2.582	4.732
DLinear- χ	0.070	1.008	1.491	2.066	4.016
LSTM-PM	5.455	5.351	5.493	5.712	6.902
LSTM- χ	0.253	1.390	1.827	2.350	4.021
PatchTST-PM	1.332	2.190	2.573	3.081	4.875
PatchTST- χ	0.128	1.070	1.529	2.093	4.071
Geo- χ -Former	0.080	0.990	1.438	1.952	3.770
<i>RMSE skill score</i>					
DLinear-PM	-6.984	-0.406	-0.286	-0.216	-0.124
DLinear- χ	-0.622	-0.041	-0.026	-0.008	0.026
LSTM-PM	-131.260	-3.832	-2.356	-1.487	-0.520
LSTM- χ	-4.612	-0.334	-0.188	-0.094	0.061
PatchTST-PM	-29.208	-0.936	-0.560	-0.356	-0.119
PatchTST- χ	-1.879	-0.091	-0.048	-0.021	0.018
Geo- χ -Former	-0.810	-0.008	0.012	0.045	0.092

Note. Boldface indicates the best value for each metric and cumulative forecast window.

We evaluate Geo- χ -Former against Bulletin A and three representative deep learning baselines. Each deep learning model was tuned by grid search on the validation period and then evaluated on the chronologically later test issue epochs. The comparison therefore uses the configuration with the best validation performance for each model family.

Tables 1 and 2 report RMSE, MAE, and RMSE skill score over selected cumulative forecast windows for the x_p and y_p components. Figure 3 summarizes the same comparison using combined PM RMSE. The suffix PM denotes direct prediction of pole coordinates, and the suffix χ denotes prediction in the excitation domain followed by PM reconstruction. For all three baselines, cumulative RMSE increases more rapidly

Table 2 Forecast accuracy for the y_p component of polar motion over selected cumulative forecast windows.

Model	Forecast window (days)				
	1	7	10	14	30
<i>RMSE (mas)</i>					
Bulletin A	0.064	0.939	1.353	1.828	3.478
DLinear-PM	0.352	1.344	1.726	2.217	4.087
DLinear- χ	0.105	0.995	1.358	1.743	3.011
LSTM-PM	6.109	6.111	6.306	6.590	8.083
LSTM- χ	0.223	1.192	1.511	1.875	3.176
PatchTST-PM	1.286	2.001	2.370	2.831	4.688
PatchTST- χ	0.131	1.024	1.378	1.778	3.075
Geo- χ -Former	0.102	0.938	1.292	1.696	3.063
<i>MAE (mas)</i>					
Bulletin A	0.054	0.616	0.905	1.249	2.452
DLinear-PM	0.275	0.990	1.264	1.594	2.887
DLinear- χ	0.084	0.676	0.937	1.209	2.090
LSTM-PM	4.585	4.790	4.961	5.217	6.470
LSTM- χ	0.176	0.859	1.098	1.362	2.281
PatchTST-PM	1.044	1.577	1.857	2.198	3.517
PatchTST- χ	0.101	0.705	0.965	1.261	2.180
Geo- χ -Former	0.082	0.635	0.896	1.185	2.156
<i>RMSE skill score</i>					
DLinear-PM	-4.521	-0.432	-0.276	-0.213	-0.175
DLinear- χ	-0.645	-0.060	-0.004	0.047	0.134
LSTM-PM	-94.694	-5.510	-3.661	-2.606	-1.325
LSTM- χ	-2.501	-0.269	-0.117	-0.026	0.087
PatchTST-PM	-19.147	-1.132	-0.752	-0.549	-0.348
PatchTST- χ	-1.049	-0.091	-0.019	0.027	0.116
Geo- χ -Former	-0.599	0.001	0.045	0.072	0.119

Note. Boldface indicates the best value for each metric and cumulative forecast window.

in the PM setting than in the excitation setting (Figure 3a). This contrast is especially clear for LSTM, for which LSTM-PM reaches a cumulative PM RMSE of 8.43 mas over days 1 to 30, compared with 4.44 mas for LSTM- χ .

Under the present experimental protocol, the paired PM-domain and excitation-domain experiments show that forecasting in the excitation domain improves PM prediction for all three tested baseline architectures. Over the cumulative 1–30-day forecast window, PM-domain errors exceed the corresponding excitation-domain errors by 3.99 mas for LSTM, 1.08 mas for PatchTST, and 0.91 mas for DLinear (Figure 3b). These positive differences show that the gain is not limited to one of the tested architectures. Within this comparison, forecasting excitation variability before reconstructing PM through the discrete Wilson equation is more effective than directly extrapolating the PM coordinates.

Among all tested methods, Geo- χ -Former has the lowest combined PM RMSE estimate over the cumulative forecast window from days 1 to 30, at 4.29 mas (Figure 3c). The corresponding estimates are 4.44 mas for LSTM- χ , 4.51 mas for DLinear- χ , and 4.56 mas for PatchTST- χ . Compared with DLinear-PM, the best baseline that predicts

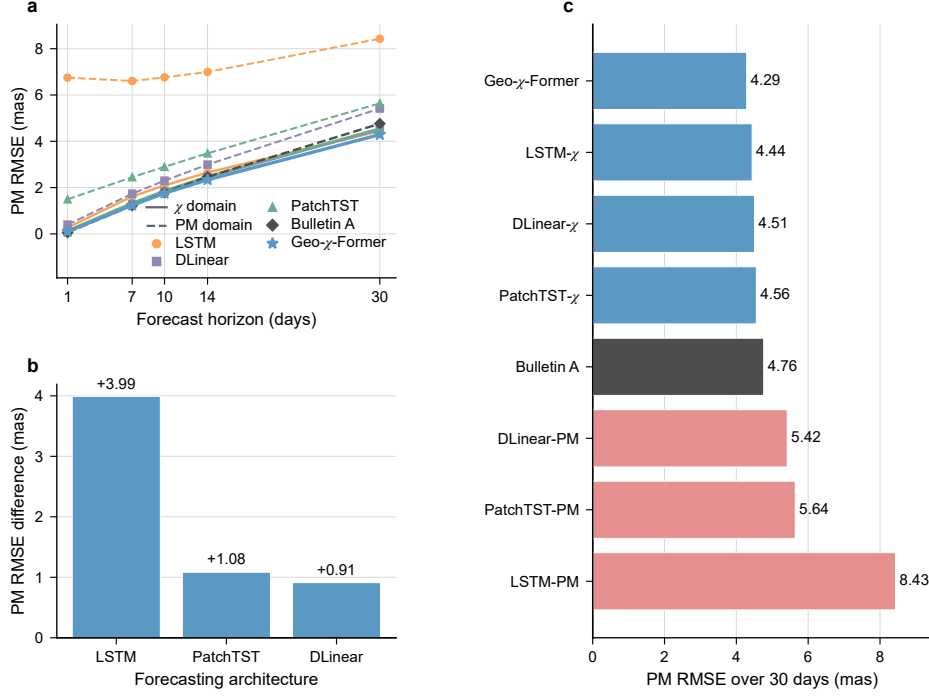


Fig. 3 Comparison of overall PM forecast accuracy among Bulletin A, general deep learning baselines, and Geo- χ -Former. (a) Cumulative PM RMSE for windows ending at forecast days 1, 7, 10, 14, and 30. Solid lines denote forecasts in the excitation domain, whereas dashed lines denote direct forecasts in the PM domain. (b) Difference in PM RMSE over the cumulative window from days 1 to 30 between the PM and excitation settings for each deep learning baseline architecture. (c) Ranking of the methods by PM RMSE over the cumulative forecast window from days 1 to 30.

PM directly (5.42 mas), the Geo- χ -Former estimate is lower by 1.13 mas, or about 21%. Using the same excitation-domain pipeline, Geo- χ -Former reduced 1–30-day combined PM RMSE by 0.15–0.27 mas relative to LSTM- χ , DLinear- χ , and PatchTST- χ .

The difference between Geo- χ -Former and Bulletin A is smaller at the shortest lead times and larger over the longer cumulative windows. Over the cumulative forecast window from days 1 to 30, the combined PM RMSE estimate for Geo- χ -Former is 4.29 mas, compared with 4.76 mas for archived Bulletin A, a difference of 0.47 mas, or 9.87%. For x_p and y_p , the corresponding estimates change from 5.77 to 5.24 mas and from 3.48 to 3.06 mas, respectively, or 9.21% and 11.93%.

The two components show different results. For x_p , Geo- χ -Former gives the lowest RMSE estimates over the cumulative windows from days 1 to 10, 1 to 14, and 1 to 30. For y_p , the methods that forecast excitation are closer. DLinear- χ gives a slightly lower RMSE estimate over days 1 to 30 than Geo- χ -Former. Because the x_p error is larger in absolute magnitude, its ordering contributes substantially to the combined PM ranking.

Geo- χ -Former has the lowest combined estimate among the tested methods in the excitation domain. Relative to archived Bulletin A, it also has lower cumulative RMSE over the windows from days 1 to 10, 1 to 14, and 1 to 30.

5 Discussion

This section interprets the retrospective estimates in terms of predictability in the excitation domain and the information provided by the EAM functions. We first examine the patterns shared across model architectures and then use input and structural ablations to characterize differences associated with individual EAM channels and model components. Attention weights and masking results are used only as diagnostics of model weighting and forecast sensitivity.

5.1 Ablation Studies

All ablation experiments use the same test issue epochs and evaluation metrics as Section 4.1. Relative changes are computed against the full Geo- χ -Former; positive values indicate larger PM RMSE.

5.1.1 Historical and Future EAM Information

The EAM input ablation tests whether EAM covariates improve the excitation domain forecasting stage used for PM prediction. The full network input includes historical GAM, historical EAM covariates, future EAM inputs, and deterministic calendar encodings. The variants alter overall EAM availability, EAM availability during the forecast period, individual EAM sources, or retention of a single EAM component; detailed experiment settings and numerical summaries are provided in Tables D.1 and D.2 in Appendix D.

The clearest difference among the input ablations is the case labeled No EAM in the figures. Removing all EAM channels from the network input produces the largest aggregate degradation over days 1 to 30 among the broad input tests (+5.03%). The No-EAM variant produces higher RMSE than the full model over every reported cumulative window ending between days 7 and 30 (Figure 4). The performance degradation after removing EAM suggests that these inputs provide complementary information to the historical GAM sequence and enhance the model’s predictive capability.

Removing all EAM channels increases the cumulative RMSE over the forecast window from days 1 to 30 by 5.03%, whereas removing EAM covariates throughout the forecast period increases the error by 2.22%. The latter setting is labeled Future off in Figure 4, with historical EAM inputs retained. The Future 7 d setting provides EAM covariates only for the first 7 lead days and fills the remaining forecast-period entries with zeros, resulting in an RMSE increase of 3.64%. The historical EAM record provides important information about the excitation state, while EAM covariates available during the forecast period provide further predictive gain. However, this additional gain depends on the availability and quality of future EAM sequences and is therefore constrained by the accuracy and reliability of EAM forecasts.

Removing individual EAM sources produces clear component-dependent differences. Removing HAM increases the cumulative PM RMSE over the forecast window

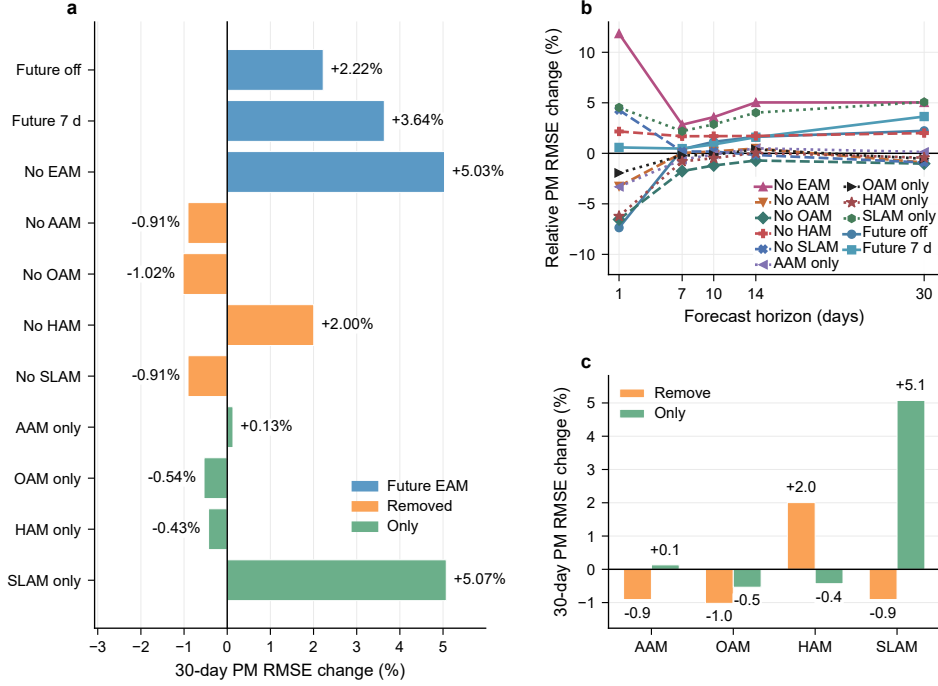


Fig. 4 EAM input ablation results for Geo- χ -Former. (a) Relative change in PM RMSE over the cumulative forecast window from days 1 to 30 for future EAM inputs, individual EAM removal, and single EAM component experiments. (b) Relative PM RMSE change over selected cumulative forecast windows. (c) Comparison between removing each EAM component from the full input set and retaining that component as the only EAM source. All values are computed relative to the full Geo- χ -Former; positive values indicate larger PM RMSE.

from days 1 to 30 by 2.00%, whereas removing AAM, OAM, or SLAM produces small decreases of approximately 0.9%, 1.0%, and 0.9%, respectively, relative to the full model. Under the current multi-source input configuration, HAM provides a clearer incremental contribution, whereas the marginal contributions of AAM, OAM, and SLAM are more limited. This may be related to the more distinct and stable periodic temporal structure of HAM, as further examined in the temporal attention analysis below. A slight decrease in RMSE after removing one component does not imply that the corresponding source is physically unimportant or unused by the model. It may instead indicate partial redundancy among EAM components. Overlapping high-frequency variations or source-specific errors may introduce additional noise during multi-source fusion, so removing one component can occasionally reduce this noise and slightly improve forecast accuracy.

Figure 4c further examines the predictive capability of each EAM source by retaining only one component at a time. With AAM, OAM, or HAM as the only EAM input, the cumulative PM RMSE over the forecast window from days 1 to 30 remains close to that of the full model, with relative changes of approximately +0.1%, -0.5%, and

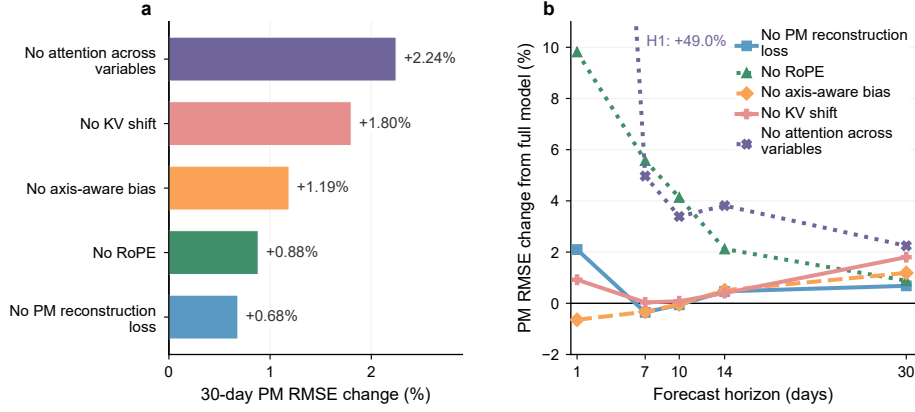


Fig. 5 Structural ablation results for Geo- χ -Former. (a) Relative change in PM RMSE over the cumulative 1-30-day forecast window after removing individual model components. (b) Relative PM RMSE change over selected cumulative forecast windows. All values are computed relative to the full Geo- χ -Former; positive values indicate larger PM RMSE. H1 denotes the day-1 degradation of the variant without attention across variables, which is truncated for display.

−0.4%, respectively. Retaining only SLAM increases RMSE by approximately 5.1%. The removal and single-source experiments indicate that AAM, OAM, and HAM contain predictive information with different degrees of substitutability. AAM and OAM alone maintain performance close to the full model but provide limited additional gain when the other EAM components are available. HAM also performs well as the sole EAM input, and its removal degrades the full model, consistent with a more complementary contribution. SLAM has weaker standalone performance and a limited marginal contribution in the full model. The contribution of each EAM component depends on both its information content and its redundancy with the other inputs.

5.1.2 Model components and PM reconstruction constraint

The structural ablation experiments evaluate whether Geo- χ -Former’s main architectural components improve PM forecasting after the excitation domain target and EAM covariates have been fixed. Each variant is retrained under the same training, validation, and test splits as the full model; detailed ablation settings and PM-domain numerical results are provided in Tables D.3 and D.4 in Appendix D.

Removing attention across variables produced the largest aggregate degradation. Cumulative RMSE over days 1 to 30 increased from 4.289 to 4.385 mas, with the largest increases over the shorter windows also occurring in this variant. The degradation suggests that learning from the additional covariate information provided by EAM and calendar encodings modestly improves the model’s forecast performance.

RMSE increased at every evaluated horizon when RoPE was removed. The increase was more pronounced over the early and intermediate windows, while the cumulative effect over days 1 to 30 remained below that observed after removing the KV shift

or axis-aware bias. RoPE appears to help the model represent relative timing in the excitation sequence, which may support the modeling of persistence, seasonal phase, and lagged responses.

The KV shift has little effect at short intervals but a clearer effect near the end of the forecast window. Removing the shift changes RMSE by only +0.03% over days 1 to 7 and +0.08% over days 1 to 10, whereas the degradation increases to +1.80% over days 1 to 30. This pattern suggests that deterministic known-covariate conditioning contributes during later recursive prediction steps, when accumulated errors are larger.

Removing the axis-aware bias and PM reconstruction loss produces smaller RMSE increases than removing attention across variables or the KV shift. The effect of the axis-aware bias is larger over the longer cumulative windows, consistent with its role in distinguishing same-axis dependencies from cross-axis interactions. Removing the PM reconstruction loss produces a modest degradation over days 1 to 30, suggesting that this term provides a PM domain constraint aligned with the final forecast. The loss links excitation prediction to PM reconstruction through the Wilson forward recurrence, but does not by itself establish full dynamical consistency. Its role may be clearer at longer horizons, where accumulated PM reconstruction errors can provide a stronger supervisory signal than in the present 30-day setting.

5.2 Interpretability

The previous sections report the forecast results and quantify the effects of individual input variables and model components. This section examines whether the trained model organizes temporal information and EAM covariates in a structured and nonuniform manner. Specifically, attention weights characterize how the model distributes emphasis across historical periods and input variables, while masking experiments evaluate the sensitivity of forecast performance to individual EAM channels.

5.2.1 Temporal attention analysis

Temporal attention is examined to assess whether the model uses the lag and periodic structure relevant to excitation forecasting. For comparison with the learned temporal weighting, we compute the absolute autocorrelation of each variable and axis combination c as

$$R_c(\tau) = |\text{corr}[z_c(t + \tau), z_c(t)]|, \quad (24)$$

where $z_c(t)$ denotes the GAM or EAM component indexed by c , and τ is the physical lag. This diagnostic is computed directly from the source series and is independent of the trained model.

For temporal attention, the latest query patch is retained, and key patch k is assigned the physical lag $\tau_k = (k_{\max} - k)P$, where P is the patch length. After averaging the attention weights over test issue dates, recursive rollout blocks, layers, and heads, the lag profile is normalized by its mean:

$$E_c(\tau_k) = \frac{\overline{A}_c(\tau_k)}{\frac{1}{|\mathcal{K}_c|} \sum_{k' \in \mathcal{K}_c} \overline{A}_c(\tau_{k'})}, \quad (25)$$

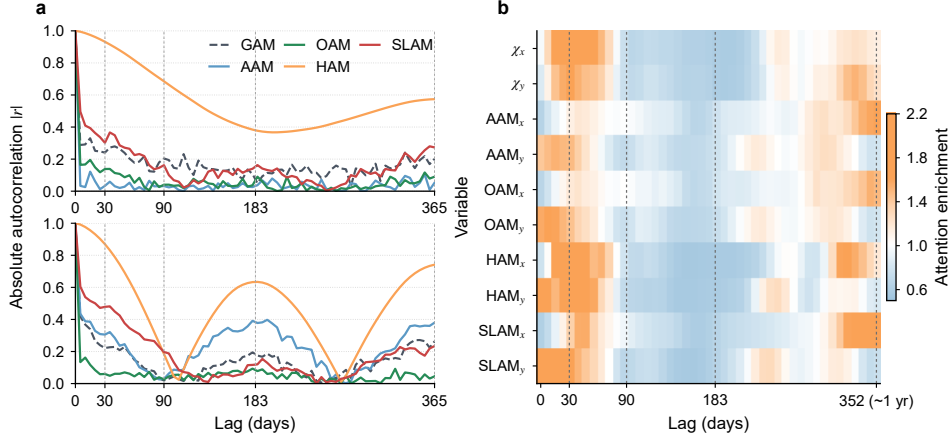


Fig. 6 Temporal attention analyses for Geo- χ -Former. (a) Absolute autocorrelation of GAM and EAM variables during the test period. Upper and lower axes show the x and y components. (b) Temporal attention enrichment by variable and lag after averaging over issue dates, rollout blocks, layers, and heads. Dashed lines mark 30, 90, 183 days, and the maximum visible lag of 352 days.

where $\overline{A}_c(\tau_k)$ denotes the temporal attention weight averaged over test issue dates, recursive rollout blocks, layers, and heads; $\mathcal{K}_c$ is the set of visible historical key patches. Accordingly, $E_c(\tau_k) = 1$ represents the mean attention for component c , whereas values above or below 1 indicate relatively enhanced or reduced attention, respectively. With $P = 8$ days and a 360-day input window, the attention map covers lags of up to 352 days. Similar patterns in R_c and E_c suggest that the learned temporal weighting reflects persistence and recurrence in the input series.

PM excitation contains rapid variability at short lags and recurring seasonal signals at longer lags, so both recent persistence and seasonal recurrence are relevant. Figure 6 compares the attention pattern learned by the trained temporal attention layer with autocorrelation computed from each source variable over the test period. The attention map covers historical lags from 0 to 352 days. This range allows the attention diagnostics to examine recurrence close to the annual scale.

Figure 6a summarizes the lag and period structure of the input variables during the test period. Most variables have their strongest autocorrelation within the first 30 to 90 days, indicating pronounced short-lag persistence. Seasonal recurrence is more evident in the y components, where several variables show renewed correlation near the semiannual and annual lags, whereas the x components exhibit weaker periodic structure. HAM has the strongest autocorrelation in both axes. GAM and SLAM show weaker but still visible recovery near the annual lag, especially in the y component. AAM is more axis-dependent, with rapid decay in AAM_x and clearer semiannual and annual recurrence in AAM_y . OAM remains weakly correlated over most lags. These

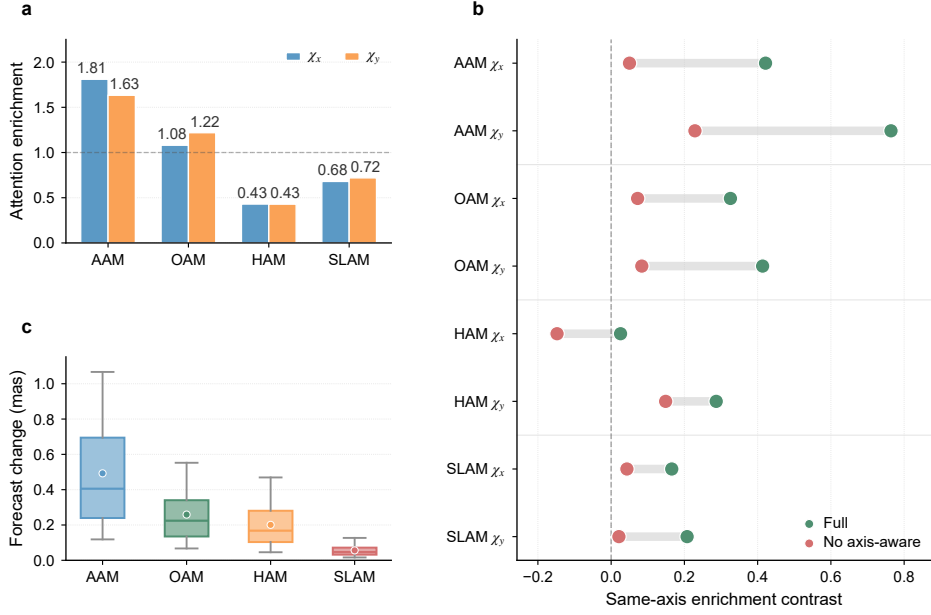


Fig. 7 Source use diagnostics for Geo- χ -Former. (a) Attention enrichment across AAM, OAM, HAM, and SLAM in χ_x and χ_y target queries. The dashed line marks the uniform attention reference. (b) Difference between enrichment along the same and opposite axes for the full model and the variant without the axis-aware bias. Positive values indicate stronger attention along the same axis. (c) Absolute change in PM forecasts after masking each EAM source. Boxplots summarize test issue dates, and white markers denote the mean.

results show that the source variables contain both recent persistence and seasonal recurrence, rather than a single monotonic decay pattern.

Figure 6b shows a shared lag structure across the target excitation and EAM channels. Most variables receive elevated attention over the first 60 to 90 days. Attention weakens over intermediate lags and increases again toward the end of the visible lag window, with the strongest late responses concentrated between 320 and 352 days. Figure 6a shows semiannual autocorrelation in HAM_y and AAM_y , together with a weaker feature in SLAM_y , but panel b shows no common attention increase near 183 days. The absence of a shared semiannual response may reflect the concentration of this periodicity in only a few variables, which provides a less consistent signal than recent persistence and recurrence close to the annual scale.

5.2.2 Attention Across Variables and Source Sensitivity

Attention across variables allows the model to weight different inputs unequally. This provides a mechanism for emphasizing EAM sources that contain more information for the target excitation. We examine this behavior in Geo- χ -Former using the test issue

dates from the forecast experiments. Attention enrichment measures how the trained network distributes weight across variables, while masking measures the sensitivity of PM reconstructed with the Wilson recurrence to controlled input perturbations.

Figure 7 characterizes how the trained model uses the different EAM sources from three perspectives: attention allocation, axis selectivity, and input perturbation. Figure 7a shows that attention is distributed nonuniformly across the EAM sources. AAM has the highest attention enrichment for the two target axes, reaching 1.81 and 1.63, respectively, followed by OAM with values of 1.08 and 1.22. HAM and SLAM remain below the uniform-attention reference. The model assigned more attention to AAM and OAM, consistent with earlier studies showing that atmospheric and oceanic angular momentum are major contributors to PM excitation (Gross et al. 2003; Dill et al. 2019). The masking results in Figure 7c further support this pattern: perturbing AAM produces the largest forecast change, followed by OAM, whereas HAM and SLAM show weaker responses overall.

Figure 7b shows an axis-selective pattern. Compared with the model without the axis-aware bias, the full model exhibits a larger difference between same-axis and opposite-axis attention for almost all EAM sources, with the clearest increases for AAM and OAM. The axis-aware bias increases attention to EAM components aligned with the target excitation axis, so the fusion depends on both physical source and axis correspondence. These diagnostics characterize information use and forecast sensitivity within the fixed trained model. The greater attention and masking responses of AAM and OAM are compatible with the stronger marginal contribution of HAM in the retrained removal experiments because HAM may supply complementary information that the other inputs do not readily replace.

The masking experiment in Figure 7c further examines sensitivity to individual EAM sources from the PM output side. Masking AAM produces the largest change in PM forecasts and the widest distribution across test cases, indicating the greatest sensitivity to perturbations in AAM among the EAM sources considered. Masking OAM also produces a clear response, whereas the response to HAM is weaker but remains non-negligible, and SLAM produces the smallest change.

The masking responses of AAM and OAM broadly follow their attention ranking, whereas HAM shows a different pattern. HAM has relatively low average attention enrichment across variables but still produces a visible PM response when masked. Its contribution may therefore depend more on learned temporal structure than on high attention across variables. As shown in Section 5.2.1, HAM exhibits persistence at recent lags and recurrence close to the annual scale. Masking HAM perturbs this temporal structure and changes the reconstructed PM forecast.

AAM and OAM receive greater attention across variables and produce larger output changes under masking. HAM has lower average attention but contains temporal structure that is not fully replaced by the other inputs. The axis-aware mechanism further separates same-axis from opposite-axis contributions.

6 Conclusions

Geo- χ -Former was developed as a geodynamically informed framework for 30-day PM prediction. The method reformulates direct PM coordinate forecasting as prediction in the excitation domain. After separating the slowly varying background, the model focuses on the short and seasonal GAM components associated with short-term PM variability and combines them with EAM covariates and calendar encodings to incorporate geophysical angular momentum information and deterministic seasonal phase information. The predicted excitation is subsequently reconstructed into PM through the Wilson forward recurrence, forming a complete forecasting pathway from observed PM to excitation prediction and PM reconstruction. This design addresses a central difficulty in data-driven PM forecasting by avoiding complete reliance on pole coordinate series dominated by persistent oscillations to identify and extrapolate shorter-timescale excitation variability, while returning the final forecast to the observable PM domain for direct comparison with observations and existing prediction products.

The paired architecture comparisons show that forecasting in the excitation domain lowers PM error relative to direct prediction in the pole coordinate domain for all tested neural architectures. Relative to IERS Bulletin A, Geo- χ -Former reduces the combined PM RMSE over the cumulative forecast window from days 1 to 30 from 4.76 to 4.29 mas, with corresponding reductions of 9.21% for x_p and 11.93% for y_p . Among the structural ablations, removing attention across variables produces the largest increase in cumulative RMSE over days 1 to 30. Removing the KV shift, axis-aware bias, RoPE-based temporal encoding, or PM reconstruction loss also increases the forecast error, although by smaller amounts. The diagnostic analyses identify temporal and geophysical structure in the trained model’s input preferences. Temporal attention emphasizes recent historical information and recurrence close to the annual scale, while attention across variables and masking indicate greater weighting of AAM and OAM and stronger forecast sensitivity to perturbations in these inputs.

The excitation-domain formulation combines GAM evolution, EAM covariates, and temporal dependencies within one forecasting model. The present study focuses on short and seasonal excitation variability relevant to short-term PM prediction. Operational performance remains limited by the availability and forecast uncertainty of EAM products. Future evaluation should use improved EAM forecasts, rapid EOP series, subdaily EOP estimates, and uncertainty models that vary with forecast lead. Extending the formulation to longer time scales would also allow tests of signals related to surface mass redistribution and interactions between the core and mantle.

Acknowledgements. The authors acknowledge the IERS, GFZ Earth System Modelling group, and ETH Zurich Geodetic Prediction Center for providing the Earth orientation and angular momentum data products used in this study. This work was supported by the National Natural Science Foundation of China (Nos. 62388101 and 12472048) and the Shenzhen Science and Technology Program (Grant No. ZDSYS20210623091808026).

Data availability. The polar motion reference series used in this study is the IERS EOP 20u24 C04 series, available at <https://www.iers.org/iers/en/dataproducts/earthorientationdata/eop>. The IERS Bulletin A products used for comparison are available from the IERS Rapid Service/Prediction Center at <https://datacenter.iers.org/availableVersions.php?id=6>. The geophysical covariates are based on GFZ Earth System Modelling effective angular momentum products, available at <https://rz-vm480.gfz.de/>. The EAM forecast products used over the forecast period are provided by the ETH Zurich Geodetic Prediction Center at <https://gpc.ethz.ch/products/EAM/>.

Author Contributions. HJH developed the model, designed and conducted the experiments, performed the data analysis, and drafted the manuscript; ZBY assisted with the experiments and model development; XDL supervised the study, guided the research direction, and reviewed and revised the manuscript; ZGW and WW reviewed the manuscript and provided comments and suggestions. All authors reviewed and approved the final manuscript.

Conflict of interest. The authors declare no conflicts of interest relevant to this study.

Appendix A EAM preprocessing details

The geophysical covariates use the atmospheric angular momentum (AAM), oceanic angular momentum (OAM), hydrological angular momentum (HAM), and sea-level angular momentum (SLAM) products described in the main text. For AAM, OAM, and HAM, the supplied mass and motion terms are summed for each equatorial component,

$$\begin{aligned} \text{AAM}_a &= \text{AAM}_a^{\text{mass}} + \text{AAM}_a^{\text{motion}}, \\ \text{OAM}_a &= \text{OAM}_a^{\text{mass}} + \text{OAM}_a^{\text{motion}}, \quad a \in \{x, y\}. \\ \text{HAM}_a &= \text{HAM}_a^{\text{mass}} + \text{HAM}_a^{\text{motion}}, \end{aligned} \tag{A.1}$$

Because SLAM contains only a mass term, it is used directly as provided.

Before alignment with the GAM target derived from PM, the native EAM records are mapped to the daily UTC grid of the PM series. For subdaily EAM records, the daily value is formed from the eight samples centered on the target daily epoch using the symmetric weights [1, 1, 3, 4, 5, 4, 3, 1] normalized by their sum (Kong et al. 2023). All EAM components are then converted to milliarcseconds using the same excitation unit convention as the GAM target.

Appendix B Geo- χ -Former implementation and training details

The main text retains the components that define the proposed architecture, including variable-wise patch embedding, temporal attention with RoPE, attention across variables, the KV shift, and the axis-aware bias. Standard Transformer layer operations and the autoregressive patch rollout are summarized here. Each Geo- χ -Former layer uses residual addition, layer normalization, and a feed-forward network after attention,

$$\mathbf{U} = \text{LN}(\mathbf{H} + \text{Attn}(\mathbf{H})), \quad \mathbf{H}' = \text{LN}(\mathbf{U} + \text{FFN}(\mathbf{U})). \quad (\text{B.1})$$

The feed-forward network consists of two linear transformations with GELU activation and dropout. During autoregressive patch rollout, the final target-patch representation is projected to the next target patch,

$$\hat{\mathbf{Y}}_{i+1} = \mathbf{W}_o \mathbf{h}_i^x + \mathbf{b}_o. \quad (\text{B.2})$$

The predicted patch is appended to the target history and used to generate subsequent patches until the full forecast horizon is reached. The same recursive strategy is used during training and inference.

The robust distance used for both the excitation loss and PM reconstruction loss is the mean Huber loss,

$$\mathcal{H}_\delta(\hat{\mathbf{y}}_{1:H_e}, \mathbf{y}_{1:H_e}) = \frac{1}{H_e d} \sum_{t=1}^{H_e} \sum_{j=1}^d \ell_\delta(\hat{y}_{t,j} - y_{t,j}), \quad (\text{B.3})$$

with

$$\ell_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \leq \delta, \\ \delta(|e| - \frac{1}{2}\delta), & |e| > \delta. \end{cases} \quad (\text{B.4})$$

A patch-wise horizon curriculum is used during autoregressive training,

$$H_e = \min(m_e P, H), \quad m_e \in \left\{1, \dots, \left\lceil \frac{H}{P} \right\rceil\right\}, \quad (\text{B.5})$$

where P is the patch length and m_e increases according to training progress. The model is optimized with AdamW (Loshchilov and Hutter 2019), gradient clipping is applied during autoregressive training, and the final model is selected according to validation performance with early stopping. All deep-learning baselines are likewise selected using the validation period before evaluation on the chronologically later test issue epochs. Table B.1 summarizes the experimental and training settings explicitly specified in the main text.

Table B.1 Experimental and training settings explicitly specified in the main text.

Item	Setting
Reference PM series	IERS EOP 20u24 C04 x_p and y_p
Training period	1993-01-01–2015-12-31
Validation period	2016-01-01–2020-01-01
Test issue period	2020-01-02–2025-11-27
Forecast issue stride	7 days
Look-back length	$L = 360$ days
Forecast horizon	$H = 30$ days
Patch length	$P = 8$ days
Reported cumulative windows	1, 7, 10, 14, and 30 days
Deep-learning baselines	DLinear, LSTM, and PatchTST
Prediction domains	PM domain and excitation domain
Covariates over the forecast period	EAM forecasts when available and deterministic calendar encodings
Evaluation domain	PM domain for all methods after PM reconstruction where required
Optimizer	AdamW
Training safeguards	Gradient clipping and validation-based early stopping
Model selection	Grid search on the validation period before test evaluation

Appendix C EAM information during the forecast period and retrospective construction

Archived EAM forecasts are indexed by issue date and lead day and are used directly when available. For earlier epochs without archived forecasts, the unavailable EAM sequence over the forecast period is approximated using lead-dependent perturbations calibrated from the 2021–2025 GPC forecast–analysis overlap. The forecast error for EAM component j and lead day ℓ is

$$e_j(t_0, \ell) = E_j^f(t_0, \ell) - E_j^{\text{ana}}(t_0 + \ell), \quad (\text{C.1})$$

and its normalized lead-dependent amplitude is

$$s_j(\ell) = \frac{\text{std}_{t_0} [e_j(t_0, \ell)]}{\sigma_j}. \quad (\text{C.2})$$

The lead dependence is approximated by

$$s_j(\ell) \approx r_j(1 + g_j \ell), \quad (\text{C.3})$$

where the fitted parameters are listed in Table C.1. The retrospectively constructed EAM value is then

$$\tilde{E}_j^f(t_0, \ell) = E_j^{\text{ana}}(t_0 + \ell) + \eta_j(t_0, \ell), \quad \eta_j(t_0, \ell) \sim \mathcal{N}\left(0, [r_j(1 + g_j\ell)\sigma_j]^2\right). \quad (\text{C.4})$$

This construction approximates the forecast EAM information unavailable at earlier epochs so that earlier historical training samples have the same input structure as later samples with real GPC forecasts. It is not an additional observational data set and is not treated as perfect future EAM information.

Table C.1 Lead-dependent perturbation parameters for retrospectively constructed EAM forecasts, estimated from the 2021–2025 GPC forecast–analysis overlap.

EAM component	r_j	g_j
AAM _{<i>x</i>}	0.517	0.156
AAM _{<i>y</i>}	0.354	0.161
OAM _{<i>x</i>}	0.560	0.083
OAM _{<i>y</i>}	0.538	0.087
HAM _{<i>x</i>}	0.030	0.284
HAM _{<i>y</i>}	0.050	0.257
SLAM _{<i>x</i>}	0.196	0.349
SLAM _{<i>y</i>}	0.217	0.247

Appendix D Ablation definitions and experiment settings

All ablation experiments use the same test issue epochs and PM-domain evaluation metrics as the full Geo- χ -Former. For ablation experiment m , the relative RMSE change with respect to the full model is

$$\Delta_{\mathcal{H}}^{(m)} = \frac{\text{RMSE}_{\mathcal{H}}^{(m)} - \text{RMSE}_{\mathcal{H}}^{\text{Full}}}{\text{RMSE}_{\mathcal{H}}^{\text{Full}}} \times 100\%, \quad (\text{D.1})$$

where $\mathcal{H}$ denotes the evaluated lead-day set. Positive values indicate larger forecast error than the full model. Tables D.1 and D.2 summarize the EAM input experiments, including future EAM availability, individual EAM removal, and single EAM component settings. Tables D.3 and D.4 summarize the structural ablations.

Table D.1 Input settings for the EAM ablation experiments.

Experiment	Input setting
Full Geo- χ -Former	Historical GAM, historical EAM covariates, future EAM inputs under the stated availability convention, and deterministic calendar encodings
No EAM	Historical GAM and deterministic calendar encodings only; all historical and future EAM inputs removed
Future off	Historical EAM inputs retained; EAM covariates throughout the forecast period set to zero
Future 7 d	Historical EAM inputs retained; EAM covariates supplied for the first 7 lead days and set to zero thereafter
No AAM	AAM _x and AAM _y removed from both historical and future EAM inputs
No OAM	OAM _x and OAM _y removed from both historical and future EAM inputs
No HAM	HAM _x and HAM _y removed from both historical and future EAM inputs
No SLAM	SLAM _x and SLAM _y removed from both historical and future EAM inputs
AAM only	AAM is the only EAM source retained; historical GAM and deterministic calendar encodings remain available
OAM only	OAM is the only EAM source retained; historical GAM and deterministic calendar encodings remain available
HAM only	HAM is the only EAM source retained; historical GAM and deterministic calendar encodings remain available
SLAM only	SLAM is the only EAM source retained; historical GAM and deterministic calendar encodings remain available

Note. Deterministic calendar encodings are retained in all experiments.

Table D.2 PM RMSE and relative changes for the EAM input ablation experiments.

Model variant	Cumulative forecast window (days)				
	1	7	10	14	30
<i>RMSE (mas)</i>					
Full Geo- χ -Former	0.101	1.235	1.753	2.335	4.289
No EAM	0.113	1.270	1.816	2.453	4.505
Future off	0.094	1.241	1.773	2.373	4.384
Future 7 d	0.102	1.241	1.768	2.372	4.445
No AAM	0.098	1.236	1.756	2.346	4.250
No OAM	0.095	1.214	1.732	2.318	4.245
No HAM	0.103	1.256	1.783	2.375	4.375
No SLAM	0.106	1.238	1.756	2.331	4.250
AAM only	0.098	1.228	1.751	2.347	4.295
OAM only	0.099	1.233	1.750	2.344	4.266
HAM only	0.095	1.226	1.745	2.336	4.270
SLAM only	0.106	1.263	1.803	2.429	4.507
<i>Relative change, Δ (%)</i>					
No EAM	+11.85	+2.83	+3.58	+5.04	+5.03
Future off	-7.36	+0.41	+1.14	+1.63	+2.22
Future 7 d	+0.58	+0.47	+0.88	+1.59	+3.64
No AAM	-3.27	+0.04	+0.18	+0.48	-0.91
No OAM	-6.55	-1.76	-1.21	-0.71	-1.02
No HAM	+2.17	+1.69	+1.70	+1.72	+2.00
No SLAM	+4.28	+0.18	+0.17	-0.16	-0.91
AAM only	-3.31	-0.62	-0.12	+0.49	+0.13
OAM only	-1.94	-0.21	-0.15	+0.38	-0.54
HAM only	-6.22	-0.80	-0.48	+0.04	-0.43
SLAM only	+4.54	+2.19	+2.87	+4.02	+5.07

Note. Relative changes are calculated according to Eq. D.1; positive values denote higher forecast error relative to the full model.

Table D.3 Settings for the structural ablation experiments of Geo- χ -Former.

Ablation variant	Modification
Full Geo- χ -Former	All proposed modules and the joint excitation and PM reconstruction losses are used
No attention across variables	Removes the attention across variables sublayer and replaces it with the corresponding feature-mixing variant
No RoPE	Removes Rotary Position Embedding (RoPE) from temporal attention
No axis-aware bias	Removes the axis-aware attention-logit bias
No PM reconstruction loss	Sets $\lambda_{\text{PM}} = 0$ and uses only the excitation loss $\mathcal{L}_\chi$
No KV shift	Removes the one-patch value shift applied to deterministic known covariates

Note. All variants use the same input variables and data splits unless otherwise specified.

Table D.4 PM RMSE and relative changes for the structural ablation experiments.

Model variant	Cumulative forecast window (days)				
	1	7	10	14	30
<i>RMSE (mas)</i>					
Full Geo- χ -Former	0.101	1.235	1.753	2.335	4.289
No attention across variables	0.151	1.297	1.812	2.424	4.385
No RoPE	0.111	1.304	1.826	2.385	4.327
No axis-aware bias	0.101	1.231	1.752	2.347	4.340
No PM reconstruction loss	0.103	1.231	1.752	2.346	4.318
No KV shift	0.102	1.236	1.754	2.344	4.366
<i>Relative change, Δ (%)</i>					
No attention across variables	+49.01	+4.97	+3.39	+3.81	+2.24
No RoPE	+9.82	+5.58	+4.14	+2.12	+0.88
No axis-aware bias	-0.64	-0.33	-0.04	+0.51	+1.19
No PM reconstruction loss	+2.10	-0.35	-0.06	+0.47	+0.68
No KV shift	+0.92	+0.03	+0.08	+0.40	+1.80

Note. Relative changes are calculated according to Eq. D.1; positive values denote higher forecast error relative to the full model.

References

- Bakhtiari M, Panahyazdan A, Sabzy S. Prediction of Earth Orientation Parameters using a hybrid Attention-based CNN-GRU Model with a Coordinate Transformation Approach. *Journal of Aerospace Science and Technology*. 2024;17(2):125–136.
- Barnes R, Hide R, White A, Wilson C. Atmospheric angular momentum fluctuations, length-of-day changes and polar motion. *Proceedings of the Royal Society of London A Mathematical and Physical Sciences*. 1983;387(1792):31–73.
- Bizouard C, Lambert S, Gattano C, Becker O, Richard JY. The IERS EOP 14C04 Solution for Earth Orientation Parameters Consistent with ITRF 2014. *Journal of Geodesy*. 2019;93(5):621–633.
- Brzeziński A. Polar motion excitation by variations of the effective angular momentum function: considerations concerning deconvolution problem. *manuscripta geodaetica*. 1992;17(1):3–20.
- Chao BF. Predictability of the Earth’s polar motion. *Bulletin géodésique (1976-1995)*. 1985;59(1):81–93.
- Chao BF. On the Excitation of the Earth’s Polar Motion. *Geophysical Research Letters*. 1985;12(8):526–529.
- Chin TM, Gross RS, Dickey JO. Modeling and forecast of the polar motion excitation functions for short-term polar motion prediction. *Journal of Geodesy*. 2004;78(6):343–353.
- Dill R, Dobslaw H, Thomas M. Improved 90-day Earth orientation predictions from angular momentum forecasts of atmosphere, ocean, and terrestrial hydrosphere. *Journal of Geodesy*. 2019;93(3):287–295.
- Dill R, Dobslaw H, Hellmers H, Kehm A, Blossfeld M, Thomas M, et al. Evaluating processing choices for the geodetic estimation of Earth orientation parameters with numerical models of global geophysical fluids. *Journal of Geophysical Research: Solid Earth*. 2020;125(9):e2020JB020025.
- Gross RS. Correspondence between theory and observations of polar motion. *Geophysical Journal International*. 1992;109(1):162–170.
- Gross RS. Earth Rotation Variations–Long Period. In: Schubert G, editor. *Treatise on Geophysics* vol. 3, 2 ed. Oxford: Elsevier; 2015. p. 215–261.
- Gross RS, Fukumori I, Menemenlis D. Atmospheric and oceanic excitation of the Earth’s wobbles during 1980–2000. *Journal of Geophysical Research: Solid Earth*. 2003;108(B8).

- Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*. 1997;9(8):1735–1780.
- Kiani Shahvandi M, Adhikari S, Dumbery M, Modiri S, Heinkelmann R, Schuh H, et al. Contributions of core, mantle and climatological processes to Earth’s polar motion. *Nature Geoscience*. 2024;17(7):705–710.
- Kiani Shahvandi M, Dill R, Dobslaw H, Kehm A, Bloßfeld M, Schartner M, et al. Geophysically informed machine learning for improving rapid estimation and short-term prediction of Earth orientation parameters. *Journal of Geophysical Research: Solid Earth*. 2023;128(10):e2023JB026720.
- Kiani Shahvandi M, Schartner M, Soja B. Neural ODE differential learning and its application in polar motion prediction. *Journal of Geophysical Research: Solid Earth*. 2022;127(11):e2022JB024775.
- Kong Q, Bo G, Li X, Han J, Zhang L, Wu Y, et al. Improved polar motion prediction with denoised geophysical fluid excitation information and long short term memory neural network. *Earth, Planets and Space*. 2026;78(1):47.
- Kong Q, Han J, Wu Y, Wang T, Chen Y. High-precision polar motion prediction using EOP_20_C04 and EAM based on CSLS+ AR and CSLS+ LSTM methods. *Geophysical Journal International*. 2023;235(2):1658–1670.
- Kosek W, McCarthy D, Luzum B. Possible improvement of Earth orientation forecast using autocovariance prediction procedures. *Journal of Geodesy*. 1998;72(4):189–199.
- Lim B, Arik SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*. 2021;37(4):1748–1764.
- Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, et al. itransformer: Inverted transformers are effective for time series forecasting. In: *International conference on learning representations*, vol. 2024; 2024. p. 11116–11140.
- Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: *International Conference on Learning Representations*; 2019. .
- Luzum BJ, Ray JR, Carter MS, Josties FJ. Recent improvements to IERS Bulletin A combination and prediction. *GPS Solutions*. 2001;4(3):34–40.
- McCarthy DD, Luzum BJ. Prediction of Earth orientation. *Bulletin géodésique (1976-1995)*. 1991;65(1):18–21.
- Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In: *International Conference on Learning*

Representations; 2023. .

- Shahvandi MK, Gou J, Schartner M, Soja B. Data driven approaches for the prediction of Earth's effective angular momentum functions. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium IEEE; 2022. p. 6550–6553.
- Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*. 2024;568:127063.
- Sun Q, Zhou WX, Fan J. Adaptive Huber Regression. *Journal of the American Statistical Association*. 2020;115(529):254–265.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in neural information processing systems*. 2017;30.
- Wang L, Que H, Wu F. The CNN-LSTM-attention model for short term prediction of the polar motion. *Measurement Science and Technology*. 2025;36(1):016323.
- Wilson CR. Discrete polar motion equations. *Geophysical Journal International*. 1985;80(2):551–554.
- Yamaguchi Y, Suemitsu I, Wei W. CITRAS: Covariate-Informed Transformer for Time Series Forecasting. *IEEE Access*. 2026;14:77983–77998.
- Yu K, Yang K, Shen T, Li L, Shi H, Song X. Estimation of earth rotation parameters and prediction of polar motion using hybrid CNN–LSTM model. *Remote Sensing*. 2023;15(2):427.
- Zeng A, Chen M, Zhang L, Xu Q. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023;37(9):11121–11128.
- Zhang Y, Yan J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: *The eleventh international conference on learning representations*; 2023. .
- Zhu H, Yang Y, Jiang C, Nie W, Luo Z, Sun Z. Improving the accuracy of polar motion prediction by combining least squares and the iTransformer model. *Earth and Space Science*. 2026;13(5):e2025EA004962.