DRC: DTM-ANCHORED RESIDUAL CORRECTION FOR LONG-HORIZON THERMOSPHERIC DENSITY FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate thermospheric density forecasting is essential for a growing range of low Earth orbit (LEO) applications, including satellite formation maintenance, space-debris collision avoidance, and reentry prediction. However, learning density dynamics directly from data remains challenging, particularly under strong and rapidly varying solar and geomagnetic activity. Rather than replacing established empirical models with neural predictors, we ask a different question: can a neural network learn to correct an empirical model while retaining it as a reliable prior? We propose **DRC**, a **D**TM-anchored **R**esidual **C**orrection framework, which uses the Drag Temperature Model (DTM) as a persistent prior and learns a bounded residual correction from historical observations and independently encoded space-weather context. A gated mechanism adaptively controls the contribution of the learned correction, while the bounded residual limits deviations from the empirical prior. We evaluate DRC on a curated dataset covering five satellites and periods of elevated solar and geomagnetic activity. DRC consistently reduces forecasting errors compared with representative deep learning baselines, demonstrating the potential of residual learning anchored in empirical models for long-horizon scientific forecasting.

1 INTRODUCTION

Thermospheric density prediction is essential for atmospheric drag estimation, orbit prediction, and space situational awareness for low Earth orbit (LEO) satellites (Emmert, 2015). Atmospheric drag is the dominant non-gravitational perturbation in LEO (Doornbos, 2012; Mutschler et al., 2026). Its estimation depends on the accuracy of thermospheric mass density estimates, which in turn affects orbit prediction, collision avoidance, debris warning, and reentry prediction. The rapid growth of the global LEO satellite population has increased the demand for accurate, timely, and robust drag modeling and density forecasting. These tasks are particularly challenging during periods of high solar activity and geomagnetic storms, when thermospheric density can change by more than 50% within a few hours (Sutton, 2008). Figure 1(a) illustrates this variability using GRACE-FO observations from 2024: changes in density level and transient enhancements accompany increased geomagnetic activity, superimposed on recurring along-orbit oscillations. Such variations pose substantial risks to satellite operations and control. In February 2022, enhanced atmospheric drag associated with geomagnetic storms led to the loss of 38 newly launched Starlink satellites (Berger et al., 2023). This event highlights the operational importance of reliable thermospheric density forecasting.

Thermospheric density modeling for operational orbit prediction and satellite control still relies primarily on semi-empirical models, including the Jacchia (Bowman et al., 2008), MSIS (Emmert et al., 2021), and DTM (Bruinsma & Boniface, 2021) families. These models are typically fitted to a limited set of historical observations. They struggle to fully capture the highly dynamic density variations driven by interactions among solar radiation, geomagnetic activity, altitude, latitude, and local time. Their errors often increase substantially during space-weather events (Wang et al., 2026; Zhang et al., 2025). Recent advances in deep learning for time series forecasting have motivated growing interest in applying these methods to thermospheric density prediction (Pan et al., 2024; Guo et al., 2026). However, directly applying general-purpose forecasting models to this task often yields limited performance. Figure 1(b) illustrates this performance gap at a 24-hour horizon, where

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

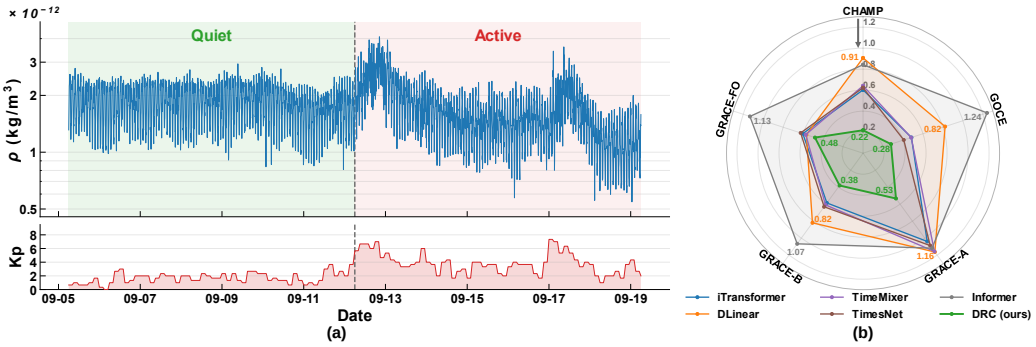

Figure 1: (a) GRACE-FO density and the geomagnetic index K_p during quiet and active periods in 2024. (b) Long-horizon forecasting performance of DRC and deep forecasting baselines across five satellites, measured by RSE at a 24-hour horizon. Lower values are better.

the selected deep forecasting baselines exhibit higher RSE than our proposed DRC across five satellite datasets. For long-horizon forecasting, deep models may rely on patterns specific to the historical training distribution and struggle to accommodate shifts in density levels across different solar and geomagnetic activity regimes. Moreover, purely data-driven models do not explicitly incorporate the domain knowledge encoded in established thermospheric models (Li et al., 2025; Sanchez-Hurtado et al., 2026). They may therefore exhibit systematic biases under distribution shifts, with prediction errors increasing as the forecast horizon extends.

Beyond these modeling challenges, data availability and evaluation practices remain major bottlenecks. The field lacks a unified thermospheric density dataset and a standardized benchmark for consistent training and systematic evaluation of deep forecasting models. Although researchers may have access to density estimates derived from onboard accelerometer measurements, these datasets often provide limited coverage of satellite platforms, orbital configurations, time periods, and solar and geomagnetic activity regimes (Siemes et al., 2023). Thermospheric density is closely linked to space environment indices (Emmert, 2015), yet few public datasets systematically align and integrate these sources (Zhu et al., 2026). Furthermore, accelerometer-derived density estimates exhibit substantial distribution shifts across time periods and satellites (Siemes et al., 2016; Bruinsma et al., 2023). Integrating observations from multiple satellites and time periods is therefore essential for evaluating model generalization under nonstationary space environment conditions.

To address these modeling and data challenges, we propose **D**TM-anchored **R**esidual **C**orrection (**DRC**), a framework that integrates empirical modeling with deep learning, and establish a unified benchmark for thermospheric density forecasting. DRC uses density estimates from DTM as a semi-empirical prior and trains a deep forecasting network to predict the residuals between observed densities and the DTM estimates. This formulation integrates the domain knowledge encoded in DTM with the capacity of deep networks to learn temporal patterns from observational data. Rather than learning thermospheric density entirely from data. The predicted corrections are then applied to the DTM estimates to produce the final density forecasts, preserving the contribution of empirical modeling while incorporating information learned from observations. Through this integration, DRC aims to improve prediction accuracy over long forecast horizons and strengthen generalization across diverse solar and geomagnetic activity regimes. Our main contributions are as follows:

- **DTM-anchored forecasting with residual correction.** We propose DRC, which incorporates prior knowledge from the semi-empirical DTM model into a Transformer-based forecasting network. By learning residual corrections, DRC improves long-horizon forecasting and mitigates error growth as the forecast horizon extends.

- **A unified dataset covering five LEO satellites.** We curate accelerometer-derived thermospheric density observations from five LEO satellites and align them with space-weather indices and orbital state information to establish a standardized forecasting benchmark.

- **Strong long-horizon performance across satellite datasets.** DRC achieves the lowest 24-hour forecasting errors among the compared deep forecasting models on all five satellites, with more stable errors across forecast horizons.

2 Related Work

2.1 Time Series Forecasting Methods and Benchmarks

Deep time series forecasting encompasses attention-based, mixing-based, and convolutional architectures. Informer (Zhou et al., 2021) introduces efficient sparse attention; Autoformer (Wu et al., 2021) and FEDformer (Zhou et al., 2022) incorporate series decomposition and frequency-domain modeling, respectively. PatchTST (Nie et al., 2022) combines patching with channel-independent processing, while iTransformer (Liu et al., 2024) models cross-variable dependencies through variate tokens. TimeMixer (Wang et al., 2024) and TimesNet (Wu et al., 2022) capture multiscale and periodic variations. These methods are commonly evaluated on ETT, Weather, Electricity, and Traffic using multiple forecast horizons and MSE/MAE metrics (Zhou et al., 2021; Novati & Koumoutsakos, 2019; Trindade, 2015; Asuncion et al., 2007). However, benchmark performance does not establish robustness to domain-specific distribution shifts. DLinear (Zeng et al., 2023) highlights the importance of strong linear baselines, while studies on exogenous variables and nonstationarity motivate accounting for external drivers and changing distributions (Liu et al., 2022). These issues motivate us to explore thermospheric density prediction as a challenging scientific application of time series forecasting. To advance research in this domain, we seek to establish a new benchmark, providing a foundation for systematic evaluation.

2.2 Thermospheric Density Modeling and Prediction

Empirical and semi-empirical models, including NRLMSIS 2.0 (Emmert et al., 2021), DTM-2020 (Bruinsma & Boniface, 2021), and JB2008 (Bowman et al., 2008), provide efficient density estimates using location, time, and model-specific solar and geomagnetic proxies. Their accuracy can degrade during geomagnetic disturbances, with errors depending on the model, event, and orbital conditions (Berger et al., 2023; Wang et al., 2026). Machine-learning approaches complement these models through density estimation (Pan et al., 2024), LSTM-based forecasting (Wang et al., 2022), and global modeling (Wang & Bai, 2026). A key distinction is whether the inputs are available at forecast issuance: evaluation with contemporaneous observed drivers does not by itself demonstrate future forecasting capability. Satellite missions provide density estimates for model development (Siemes et al., 2023; 2016), supported by advances in retrieval and reprocessing (Pilinski et al., 2016; Licata & Mehta, 2019). These products include accelerometer-derived estimates and, for Swarm, GPS-derived estimates. Existing efforts such as STORM-AI (Sanchez-Hurtado et al., 2026) already support standardized multi-satellite forecasting. Our benchmark complements these resources by organizing records from five LEO satellites and evaluating multiple forecast horizons, activity regimes, and cross-satellite settings.

2.3 Hybrid Modeling and Residual Learning

Hybrid modeling combines mechanistic models with trainable components to incorporate domain knowledge while accounting for incomplete model descriptions. Early work integrates first-principles models with neural networks that estimate unknown process parameters (Psichogios & Ungar, 1992). Universal differential equations extend this approach by embedding trainable function approximators within differential equations to represent unknown mechanisms (Rackauckas et al., 2020). Residual learning offers a complementary formulation, in which a learned correction augments an imperfect baseline model. Yin et al. (2021) decomposes system dynamics into physical and learned components, minimizing the contribution of the learned component while maintaining agreement with observations. More broadly, Levine & Stuart (2022) formulate model-error learning in both continuous and discrete time, covering memoryless and memory-dependent corrections. Building on these perspectives, DRC uses DTM-2020 as a fixed semi-empirical prior for thermospheric density forecasting, retaining its domain knowledge while allowing the forecasting network to focus on residual corrections across forecast horizons.

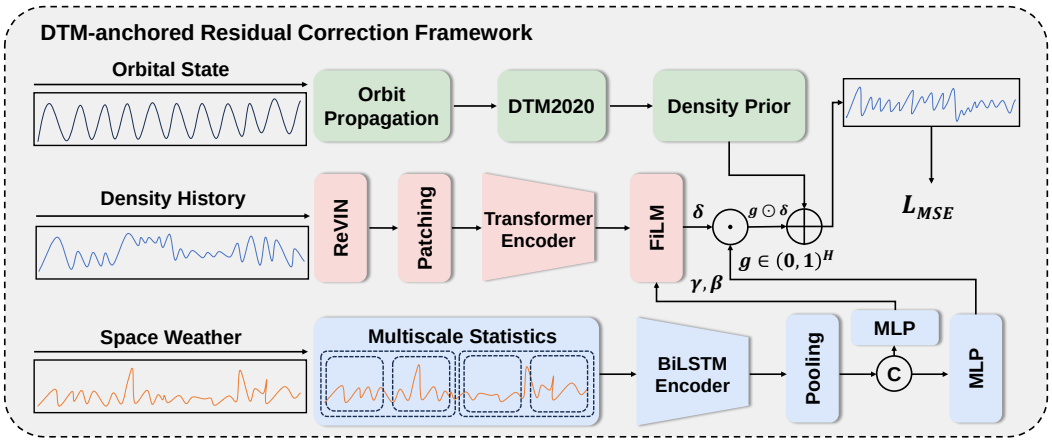

Figure 2: Overview of DRC. DTM-2020 provides density prior, while a patch-based Transformer learns residual corrections guided by multi-scale space-weather context through FiLM and gating. The corrections refine the DTM prior with end-to-end training using an MSE loss.

3 Methodology

3.1 Problem Definition

Thermospheric density forecasting is formulated as a multivariate-to-univariate prediction task. Given an input sequence $\mathbf{X} \in \mathbb{R}^{L\times D}$, the objective is to learn a mapping $\hat{\mathbf{y}} = f_\theta(\mathbf{X})$ that predicts log-density over the next H time steps, where L is the look-back window length, D is the number of input variables, and $\hat{\mathbf{y}} \in \mathbb{R}^H$ is the forecast. In our setting, $D = 12$, with each time step containing the log-density $y_t = \log_{10}\rho_t$, six space-weather indices characterizing solar and geomagnetic activity, and five orbital parameters comprising latitude, longitude, altitude, local solar time, and argument of latitude. The target sequence is $\mathbf{y} = (y_{L+1}, \ldots, y_{L+H})$.

As is shown in Figure 2, DRC formulates forecasting as residual correction to a semi-empirical prior. Given the historical observations $\mathbf{X}$ and the DTM persistence prior $\mathbf{y}^{\mathrm{DTM}} \in \mathbb{R}^H$ in log-density space (Section 3.2), the forecast is

$$\hat{\mathbf{y}} = \mathbf{y}^{\mathrm{DTM}} + \mathbf{g}_\theta(\mathbf{X}, \mathbf{y}^{\mathrm{DTM}}) \odot \boldsymbol{\delta}_\theta(\mathbf{X}), \tag{1}$$

where $\odot$ denotes element-wise multiplication. A patch-based Transformer produces bounded residual corrections $\boldsymbol{\delta}_\theta(\mathbf{X}) \in \mathbb{R}^H$ in log-density space and per-step gates $\mathbf{g}_\theta(\mathbf{X}) \in (0,1)^H$ that control their strength. This formulation incorporates the domain knowledge encoded in DTM while allowing the network to learn bounded adjustments to prior predictions.

3.2 DTM Anchoring

We construct the prior using DTM-2020 (Bruinsma & Boniface, 2021) with persistent solar–geomagnetic drivers. Specifically, the drivers available at the final input step are held constant throughout the forecast horizon, while satellite positions are obtained through orbit propagation:

$$y_t^{\mathrm{DTM}} = \log_{10}\mathrm{DTM}\big(\mathbf{f}_L, \mathbf{r}_t, t\big),\ t = L+1, \ldots, L+H, \tag{2}$$

where $\mathbf{f}_L$ contains the latest available solar and geomagnetic indices required by DTM-2020, and $\mathbf{r}_t$ is the satellite position propagated from the orbital state at the forecast origin. These indices remain fixed throughout the forecast horizon, while the density varies with satellite position and time. The resulting sequence forms the prior $\mathbf{y}^{\mathrm{DTM}} \in \mathbb{R}^H$ in Eq. 1. The same procedure is used during training and evaluation, relying only on information available at the forecast origin.

3.3 Patch Encoding with Space-Weather Context

A univariate patch-based Transformer backbone is adopted to encode the log-density history $y_{1:L}$. The sequence is divided into N patches of length P with stride S. Each patch is linearly projected into a d-dimensional embedding and augmented with positional information. A Transformer encoder produces patch representations $\mathbf{Z} \in \mathbb{R}^{N \times d}$, where d is the hidden feature dimension. Patching summarizes local observations into tokens, shortening the attention sequence while enabling dependencies across patches to be modeled.

To condition these representations on space weather, we summarize the six historical indices over trailing windows $w \in \{12, 36, 72, 288\}$, corresponding to 1, 3, 6, and 24 hours at a 5-minute sampling interval. For each window, the maximum, mean, and standard deviation of each index form a summary vector $\mathbf{q}_w \in \mathbb{R}^{18}$. A BiLSTM (Graves & Schmidhuber, 2005) encodes the summary vectors ordered by window length, and average pooling aggregates the resulting scale-specific tokens:

$$\mathbf{c} = \mathrm{Pool}\big(\mathrm{Enc}_s(\mathbf{q}_{12}, \mathbf{q}_{36}, \mathbf{q}_{72}, \mathbf{q}_{288})\big) \in \mathbb{R}^d. \tag{3}$$

This representation captures peaks, background levels, and variability across time scales, complementing the latest index values used by the DTM persistence prior. The context vector $\mathbf{c}$ modulates each patch representation $\mathbf{z}_i$ through FiLM (Perez et al., 2018):

$$\tilde{\mathbf{z}}_i = \gamma \odot \mathbf{z}_i + \beta, \ (\gamma, \beta) = \mathrm{MLP}_\phi(\mathbf{c}), \tag{4}$$

where $\gamma, \beta \in \mathbb{R}^d$ are context-dependent scaling and shift vectors shared across patches, and $\odot$ denotes element-wise multiplication. The modulated representations form $\tilde{\mathbf{Z}}$. The same context vector also conditions the residual gates described in Section 3.4, allowing space-weather information to guide both feature extraction and correction strength.

3.4 Gated Residual Correction

A MLP maps the flattened patch representations to bounded corrections in log-density space:

$$\boldsymbol{\delta}_\theta(\mathbf{X}) = b \, \tanh\big(\mathrm{MLP}_\delta(\mathrm{Flat}(\tilde{\mathbf{Z}}))\big), \tag{5}$$

where $b > 0$ is the satellite-specific correction bound in log-density units. The output layer's weights and biases are initialized to zero, so the initial correction is zero and DRC reproduces the DTM prior. To adapt the correction strength at each forecast step, a gating network combines the space-weather context with summary statistics of the prior:

$$\mathbf{g}_\theta(\mathbf{X}, \mathbf{y}^{\mathrm{DTM}}) = \mathrm{sigmoid}\big(\mathrm{MLP}_g(\mathrm{Concat}(\mathbf{c}, \bar{p}, \Delta p))\big) \in (0, 1)^H, \tag{6}$$

where Concat denotes concatenation. The prior is standardized using training-set statistics, and $\bar{p}$ and Δp denote its mean and last-minus-first difference over the forecast horizon, respectively. These features summarize the prior level and overall change. The gate scales each residual before it is added to the prior in Eq. 1; smaller gate values retain forecasts closer to DTM. The bounded head and sigmoid gate ensure $|\hat{y}_t - y_t^{\mathrm{DTM}}| \le b$ at every forecast step.

3.5 Training Objective

All trainable modules are jointly optimized using the final density predictions $\hat{\rho}_{i,t} = 10^{\hat{y}_{i,t}}$, where i indexes samples in a mini-batch $\mathcal{B}$ and t indexes time steps. We define the normalized prediction error as $e_{i,t} = (\hat{\rho}_{i,t} - \rho_{i,t})/\sigma_\rho$, where σ_ρ is the training-set standard deviation of density. The training objective is the mean squared error in standardized density space:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{|\mathcal{B}|H} \sum_{i \in \mathcal{B}} \sum_{t=L+1}^{L+H} e_{i,t}^2, \tag{7}$$

where $|\mathcal{B}|$ is the batch size and H is the forecast horizon.

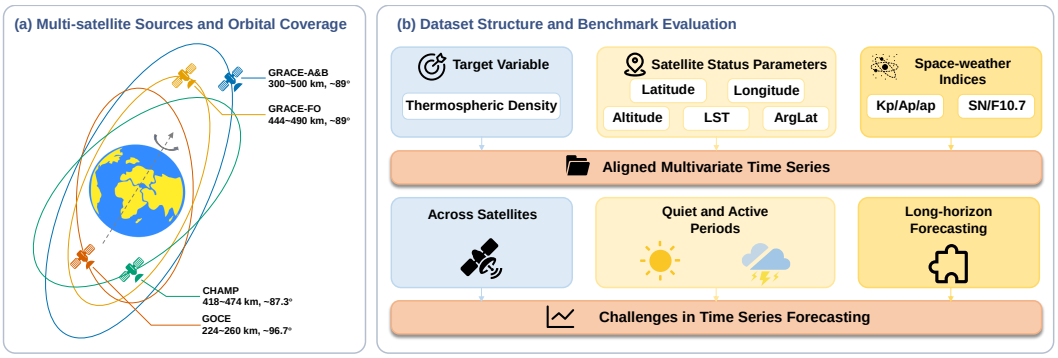

Figure 3: Overview of the thermospheric density dataset. (a) Satellite sources and schematic orbital coverage. (b) Density observations, orbital variables, and space-weather indices are aligned into multivariate time series for forecasting evaluation across satellite datasets.

4 Dataset

We construct a thermospheric density dataset by integrating satellite observations from the TOLEOS project at Delft University of Technology (TU Delft) (Siemes et al., 2023) with space-weather indices from the GFZ German Research Centre for Geosciences (Bartels, 1957; Matzka et al., 2021). As illustrated in Figure 3(a), the dataset integrates observations from five LEO satellites (CHAMP, GRACE-A, GRACE-B, GRACE-FO, and GOCE), which have complementary orbital configurations and collectively provide temporal coverage from 2000 to 2025. Each record combines an accelerometer-derived density estimate with five orbital variables and six space-weather indices aligned by timestamp, as summarized in Figure 3(b). We remove invalid observations, resample the records to a 5-minute interval, and organize approximately 4.9 million retained observations.

The diversity of orbital configurations and solar activity conditions introduces substantial differences in density levels and temporal patterns, challenging models to maintain forecasting accuracy across satellite datasets and prediction horizons. We divide each satellite record into uninterrupted segments and construct input and target windows entirely within individual segments. Training, validation, and test sets are partitioned chronologically at segment boundaries with a target ratio of 80:10:10. The dataset encompasses observations from multiple satellites across active and quiet eras, capturing a broad range of orbital conditions and thermospheric density variations, as outlined in Figure 3(b). This diversity provides a challenging setting for forecasting over long horizons, requiring models to capture complex temporal dynamics under varying solar and geomagnetic activity. More preprocessing details and satellite-level statistics are provided in Appendix A.

5 Experiments

5.1 Experimental Setup

Baselines. We compare DRC with six widely used deep time series forecasting methods: PatchTST (Nie et al., 2022), DLinear (Zeng et al., 2023), iTransformer (Liu et al., 2024), Informer (Zhou et al., 2021), TimeMixer (Wang et al., 2024), and TimesNet (Wu et al., 2022). These methods represent diverse modeling approaches, including linear decomposition, Transformer-based forecasting, multi-scale mixing, and periodic pattern modeling. All models use multivariate inputs comprising density history, orbital parameters, and space-weather indices to forecast the density.

Implementation Details. DRC uses a three-layer patch-based Transformer backbone with hidden dimension 64, a multi-scale space-weather encoder, and gated residual correction. All models use a fixed look-back window of 288 steps (24 hours) to forecast 1, 3, 6, 12, and 24 hours ahead, with particular emphasis on long-horizon performance at 24 hours. Trainable models are optimized using Adam with a one-cycle learning-rate schedule, a peak learning rate of 10^{-4}, batch size 128, and a maximum of 100 epochs. Detailed architecture settings and early-stopping criteria are provided in Appendix B.1. We report MAE, MSE, and RSE on density predictions in the original physical scale,

Table 1: Performance comparison for 288-step (24-hour) forecasting. Lower values are better. MAE and MSE are reported in units of $10^{-12}\,\mathrm{kg/m^3}$ and $10^{-24}\,(\mathrm{kg/m^3})^2$; RSE is dimensionless. **Bold** and underlined values indicate the best and second-best results, respectively.

Satellite	Metric	DRC (ours)	PatchTST	DLinear	iTransformer	TimeMixer	TimesNet	Informer
CHAMP	MAE	**0.5882**	2.3070	2.1670	1.4549	1.5452	1.5997	1.8991
	MSE	**1.0133**	20.4009	17.4850	7.7286	8.7496	8.2612	15.0816
	RSE	**0.2184**	0.9798	0.9071	0.6031	0.6417	0.6235	0.8424
GOCE	MAE	**4.4735**	9.3793	13.6734	7.9894	7.8610	6.4702	21.8800
	MSE	**34.0258**	145.3264	293.0888	102.0037	100.9775	72.1835	669.3582
	RSE	**0.2799**	0.5784	0.8214	0.4846	0.4821	0.4076	1.2413
GRACE-A	MAE	**0.4234**	1.0836	1.0187	0.8937	1.0117	0.9523	0.9061
	MSE	**0.3731**	2.0627	1.7734	1.4154	1.7540	1.5514	1.6388
	RSE	**0.5335**	1.2543	1.1631	1.0391	1.1567	1.0878	1.1181
GRACE-B	MAE	**0.3018**	0.5975	0.6365	0.4417	0.4769	0.4833	0.8296
	MSE	**0.1859**	0.7764	0.8629	0.4409	0.4757	0.5111	1.4617
	RSE	**0.3804**	0.7774	0.8196	0.5859	0.6085	0.6308	1.0667
GRACE-FO	MAE	**0.2119**	0.2631	0.2660	0.2842	0.2704	0.2949	0.5496
	MSE	**0.1052**	0.1457	0.1497	0.1692	0.1531	0.1815	0.5971
	RSE	**0.4429**	0.5595	0.5671	0.6030	0.5735	0.6245	1.1327

Table 2: Performance comparison averaged equally over five forecast horizons (1, 3, 6, 12, and 24 hours). Lower values are better. MAE and MSE are reported in units of $10^{-12}\,\mathrm{kg/m^3}$ and $10^{-24}\,(\mathrm{kg/m^3})^2$; RSE is dimensionless. **Bold** and underlined values indicate the best and second-best results, respectively.

Satellite	Metric	DRC (ours)	PatchTST	DLinear	iTransformer	TimeMixer	TimesNet	Informer
CHAMP	MAE	**0.6170**	1.1871	1.1997	0.9824	1.0011	1.0455	1.4185
	MSE	**1.2119**	7.2389	6.7425	4.0494	4.1204	4.4485	10.2905
	RSE	**0.2352**	0.5061	0.4976	0.4164	0.4131	0.4321	0.6805
GOCE	MAE	**4.2221**	6.6708	7.2621	5.0800	5.8734	5.2412	18.1562
	MSE	**30.8047**	77.4040	102.3500	46.6698	58.6757	48.2969	497.0496
	RSE	**0.2662**	0.4096	0.4434	0.3136	0.3587	0.3238	1.0473
GRACE-A	MAE	**0.4165**	0.5837	0.5865	0.5448	0.5814	0.6358	0.7687
	MSE	**0.3710**	0.7670	0.7370	0.6188	0.7085	0.7929	1.2470
	RSE	**0.5109**	0.6717	0.6681	0.6186	0.6580	0.7108	0.9273
GRACE-B	MAE	0.2901	0.3655	0.3759	**0.2815**	0.3120	0.3066	0.7157
	MSE	**0.1659**	0.3218	0.3455	0.1941	0.2194	0.2207	1.1709
	RSE	**0.3605**	0.4670	0.4800	0.3668	0.3951	0.3912	0.9356
GRACE-FO	MAE	0.2227	0.1877	**0.1849**	0.2015	0.1943	0.2113	0.4702
	MSE	0.1006	0.0790	**0.0779**	0.0906	0.0838	0.0972	0.4854
	RSE	0.4638	0.3992	**0.3952**	0.4265	0.4115	0.4429	1.0102

using identical test windows within each setting. Metric definitions and DTM prior computation are described in Appendices B.2 and B.3, respectively.

5.2 Main Results

Long-horizon forecasting performance. Table 1 compares DRC with six general-purpose models at a 288-step (24-hour) horizon. DRC achieves the lowest MAE, MSE, and RSE across all five satellite datasets, demonstrating strong long-horizon forecasting performance. The largest improvements occur on CHAMP and GRACE-A, where iTransformer is the second-best model for all three metrics. On CHAMP, DRC reduces MAE, MSE, and RSE by 59.57%, 86.89%, and 63.79%, respec-

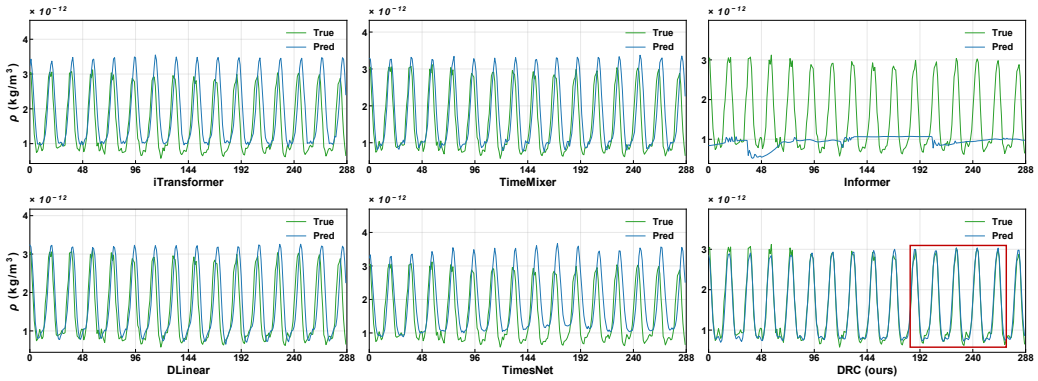

Figure 4: Qualitative comparison of 288-step (24-hour) forecasting on the same test sequence. The red box highlights DRC's closer agreement with the true densities in both peak and trough magnitudes toward the end of the forecast horizon.

tively, relative to iTransformer. On GRACE-A, the corresponding reductions are 52.62%, 73.64%, and 48.66%. DRC also outperforms the strongest deep forecasting baseline on each of the remaining satellite datasets. These results support the effectiveness of incorporating DTM prior knowledge through residual correction for long-horizon thermospheric density forecasting.

Average performance across five different forecast horizons. Table 2 reports results averaged equally over five forecast horizons (1, 3, 6, 12, and 24 hours). DRC ranks first in 11 of 15 satellite–metric comparisons, achieving the lowest MAE, MSE, and RSE on CHAMP, GOCE, and GRACE-A, and the lowest MSE and RSE on GRACE-B. The largest gains occur on CHAMP: MAE and MSE decrease by 37.19% and 70.07% relative to iTransformer, while RSE decreases by 43.06% relative to TimeMixer, the respective second-best models. MSE reductions reach 40.05% on GRACE-A and 33.99% on GOCE relative to iTransformer.

The five-satellite dataset challenges long-horizon forecasting through diverse orbital conditions, space-weather backgrounds, and density variations. The substantially larger 24-hour errors of general-purpose models suggest limitations in forecasting from historical patterns without a semi-empirical prior. DRC uses DTM knowledge to anchor future density levels and learns residual corrections, supporting stronger long-horizon performance. Its competitive horizon-averaged results further support the value of this prior across forecast scales.

5.3 Qualitative Analysis

Temporal alignment and amplitude accuracy. Figure 4 compares 288-step (24-hour) predictions on the same test sequence. Most baselines capture the recurring oscillations but exhibit phase offsets, overestimated peaks, or mismatched trough levels. Informer produces a comparatively flat trajectory that misses most of the variation. DRC more closely follows the true density in both oscillation timing and amplitude, maintaining agreement across successive peaks and troughs. This alignment persists toward the end of the forecast horizon, as highlighted by the red box, without pronounced phase drift or amplitude distortion. The example complements the quantitative results by illustrating DRC's consistent tracking of density variations over the 24-hour forecast.

Error stability across forecast horizons. Figure 5 shows that DRC maintains relatively stable MSE from 1 to 24 hours and achieves the lowest 24-hour errors on CHAMP, GOCE, and GRACE-A. Its MSE decreases on CHAMP and varies only modestly on GOCE and GRACE-A. In comparison, most deep forecasting baselines exhibit substantial error growth at longer horizons, while Informer retains comparatively high errors, especially on GOCE. Although some baselines perform better at short horizons, DRC's advantage becomes more pronounced as the horizon extends. Results for the remaining satellites are provided in Appendix C.

Together, these comparisons highlight two challenges in thermospheric density forecasting: preserving the timing and amplitude of recurring variations and limiting error growth over longer horizons. General-purpose deep models often struggle with one or both. DRC addresses these challenges by

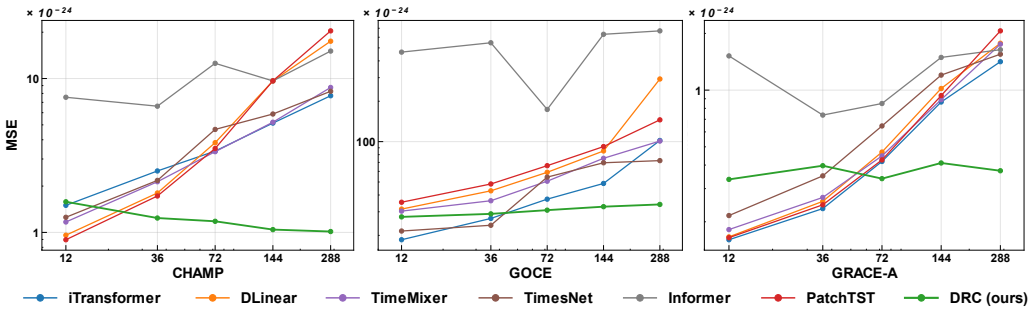

Figure 5: MSE across forecast horizons on CHAMP, GOCE, and GRACE-A. Each step represents 5 minutes, with horizons spanning 1–24 hours. MSE is shown on logarithmic axes in units of $10^{-24}\,(\mathrm{kg/m^3})^2$. Results for GRACE-B and GRACE-FO are provided in Figure 6 in the appendix.

Table 3: Component ablation on GRACE-FO in the full-data setting at a 288-step (24-hour) forecast horizon. MAE and MSE are reported in units of $10^{-12}\,\mathrm{kg/m^3}$ and $10^{-24}\,(\mathrm{kg/m^3})^2$, respectively; RSE is dimensionless. Lower values are better, and the best results are highlighted in **bold**.

Metric	Transformer-based backbone	Backbone + space-weather context	Backbone + DTM anchoring	DRC (full model)
MAE ↓	0.2631	0.2629	0.2297	**0.2119**
MSE ↓	0.1457	0.1454	0.1079	**0.1052**
RSE ↓	0.5595	0.5589	0.4815	**0.4429**

using DTM prior knowledge to anchor density predictions and space-weather context to guide residual corrections. The closer agreement with the true density in the illustrated sequence and the more stable errors across horizons support the effectiveness of this design for long-horizon forecasting.

5.4 Ablation Studies

Table 3 examines the contributions of space-weather context and DTM anchoring on GRACE-FO at a 288-step (24-hour) forecast horizon. Adding space-weather context to the Transformer-based backbone yields only marginal improvements, reducing MAE, MSE, and RSE by approximately 0.08%, 0.21%, and 0.11%, respectively. This suggests that contextual information alone provides limited benefits for this long-horizon forecasting setting. Incorporating DTM anchoring produces substantially larger gains, with reductions of 12.69% in MAE, 25.94% in MSE, and 13.94% in RSE relative to the backbone. These results support the value of the semi-empirical prior in guiding density predictions over an extended forecast horizon. The full DRC model further reduces MAE, MSE, and RSE by 7.75%, 2.50%, and 8.02%, respectively, compared with the variant using DTM anchoring alone. This improvement suggests that space-weather conditioning is more effective when combined with the anchored correction formulation than when added directly to the backbone. Overall, DRC achieves reductions of 19.46% in MAE, 27.80% in MSE, and 20.84% in RSE relative to the backbone. The results indicate complementary benefits from DTM prior knowledge and context-dependent residual correction for 24-hour forecasting on this dataset.

6 Conclusion

We present DRC, a thermospheric density forecasting framework that combines DTM prior knowledge with context-dependent residual correction, and constructed a unified dataset covering five LEO satellites. DRC outperforms the compared general-purpose deep forecasting models on all five satellite datasets at the 24-hour horizon and ranks first in 11 of the 15 comparisons averaged across forecast horizons. Further analyses show relatively stable errors across horizons and accurate tracking of density variations, although performance varies across satellites. These results highlight the value of integrating semi-empirical knowledge with deep learning for long-horizon thermospheric density forecasting.

REFERENCES

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

J Bartels. The geomagnetic measures for the time-variations of solar corpuscular radiation, described for use in correlation studies in other geophysical fields. *Ann. Intern. Geophys.*, 4:227–236, 1957.

TE Berger, M Dominique, G Lucas, M Pilinski, V Ray, R Sewell, EK Sutton, JP Thayer, and E Thiemann. The thermosphere is a drag: The 2022 starlink incident and the threat of geomagnetic storms to low earth orbit space operations. *Space Weather*, 21(3):e2022SW003330, 2023.

Bruce Bowman, W Kent Tobiska, Frank Marcos, Cheryl Huang, Chin Lin, and William Burke. A new empirical thermospheric density model jb2008 using new solar and geomagnetic indices. In *AIAA/AAS astrodynamics specialist conference and exhibit*, pp. 6438, 2008.

Sean Bruinsma and Claude Boniface. The operational and research dtm-2020 thermosphere models. *Journal of space weather and space climate*, 11:47, 2021.

Sean Bruinsma, Christian Siemes, John T Emmert, and Martin G Mlynczak. Description and comparison of 21st century thermosphere data. *Advances in Space Research*, 72(12):5476–5489, 2023.

Eelco Doornbos. *Thermospheric density and wind determination from satellite dynamics*. Springer Science & Business Media, 2012.

John T Emmert. Thermospheric mass density: A review. *Advances in Space Research*, 56(5): 773–824, 2015.

John T Emmert, Douglas Patrick Drob, J Michael Picone, David E Siskind, M Jones Jr, Martin G Mlynczak, Peter F Bernath, X Chu, E Doornbos, Bernd Funke, et al. Nrlmsis 2.0: A whole-atmosphere empirical model of temperature and neutral species densities. *Earth and Space Science*, 8(3):e2020EA001321, 2021.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

Yu Guo, Xiaohong Zhang, Fei Guo, Zhiyu Zhang, and Xiaodong Ren. Upper thermospheric mass density forecasting using deep learning model and leo satellite observations. *GPS Solutions*, 30 (2):73, 2026.

Matthew Levine and Andrew Stuart. A framework for machine learning of model error in dynamical systems. *Communications of the American Mathematical Society*, 2(7), 2022.

Ling Li, Changyong He, Dunyong Zheng, Shaoning Li, and Dong Zhao. A prior knowledge-enhanced deep learning framework for improved thermospheric mass density prediction. *Atmosphere*, 16(5):539, 2025.

Richard Licata and Piyush M Mehta. Physics-informed machine learning for probabilistic space weather modeling and forecasting: Thermosphere and satellite drag. In *AGU Fall Meeting Abstracts*, volume 2019, pp. SM31E–3198, 2019.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *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 *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Jürgen Matzka, Claudia Stolle, Yosuke Yamazaki, Oliver Bronkalla, and Achim Morschhauser. The geomagnetic kp index and derived indices of geomagnetic activity. *Space weather*, 19(5): e2020SW002641, 2021.

Shaylah Mutschler, Marcin Pilinski, Sean Bruinsma, Eric Sutton, W Kent Tobiska, Delores Knipp, Tzu-Wei Fang, Steve Casali, Vishnuu Mallik, Brandon DiLorenzo, et al. A survey of current operations-ready thermospheric density models for drag modeling in leo operations. *The Journal of the Astronautical Sciences*, 73(2):32, 2026.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Guido Novati and Petros Koumoutsakos. Remember and forget for experience replay. In *International Conference on Machine Learning*, pp. 4851–4860. PMLR, 2019.

Qian Pan, Chao Xiong, Hermann Lühr, Artem Smirnov, Yuyang Huang, Chunyu Xu, Xu Yang, Yunliang Zhou, and Yang Hu. Machine learning based modeling of thermospheric mass density. *Space Weather*, 22(5):e2023SW003844, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Marcin D Pilinski, Geoff Crowley, Eric Sutton, and Mihail Codrescu. Improved orbit determination and forecasts with an assimilative tool for satellite drag specification. In *Advanced Maui optical and space surveillance technologies conference*, volume 104, 2016.

Dimitris C Psichogios and Lyle H Ungar. A hybrid neural network-first principles approach to process modeling. *AIChE Journal*, 38(10):1499–1511, 1992.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

Sergio Sanchez-Hurtado, Haley E Solera, William E Parker, Mia Tian, Ruoxi Qian, Edenna Chen, Enrico M Zucchelli, Rasyid Ridha, Dominic Bruno, Christopher Yeung, et al. Ai challenge for satellite tracking and orbit resilience modeling (storm-ai): Data set, design, and results. *Space Weather*, 24(8):e2026SW004994, 2026.

Christian Siemes, João de Teixeira da Encarnação, Eelco Doornbos, Jose Van Den Ijssel, Jiří Kraus, Radek Pereštỳ, Ludwig Grunwaldt, Guy Apelbaum, Jakob Flury, and Poul Erik Holmdahl Olsen. Swarm accelerometer data processing from raw accelerations to thermospheric neutral densities. *Earth, Planets and Space*, 68(1):92, 2016.

Christian Siemes, Claudia Borries, Sean Bruinsma, Isabel Fernandez-Gomez, Natalia Hładczuk, Josevan den IJssel, Timothy Kodikara, Kristin Vielberg, and Pieter Visser. New thermosphere neutral mass density and crosswind datasets from champ, grace, and grace-fo. *Journal of space weather and space climate*, 13:16, 2023.

Eric K Sutton. *Effects of solar disturbances on the thermosphere densities and winds from CHAMP and GRACE satellite accelerometer data*. PhD thesis, University of Colorado at Boulder, 2008.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

Jack C Wang, Jia Yue, Sean Bruinsma, Masha Kuznetsova, Joseph Sypal, Richard Mullinix, Chiu Wiegand, Paul Dimarzio, Min-Yang Chou, Maksym Petrenko, et al. Comprehensive and open assessment of thermospheric models during geomagnetic storm times within ccmc framework. *Space Weather*, 24(4):e2025SW004782, 2026.

Peian Wang, Zhou Chen, Xiaohua Deng, Jingsong Wang, Rongxing Tang, Haimeng Li, Sheng Hong, and Zhiping Wu. The prediction of storm-time thermospheric mass density by lstm-based ensemble learning. *Space Weather*, 20(3):e2021SW002950, 2022.

Ruochen Wang and Xiaoli Bai. A machine-learning-based global thermospheric density forecasting model. *Space Weather*, 24(6):e2026SW004968, 2026.

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 *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel De Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Kun Zhang, Yanshi Huang, Fengsi Wei, Pingbing Zuo, Hao Yang, Jinlong Ji, and Zehao Chen. Thermospheric density variations during extreme geomagnetic storms and their potential impact on the orbit of the china space station. *Frontiers in Astronomy and Space Sciences*, 12:1644152, 2025.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

Kemin Zhu, Jizhang Sang, Xianjie Li, Lei Liu, Yangbo Qian, and Shengfu Xia. Applying hpo indices to empirical thermospheric density models during geomagnetic storms. *Earth and Space Science*, 13(2):e2025EA004454, 2026.

TABLE OF CONTENTS IN APPENDIX

A Dataset Construction **13**

B Additional Experimental Details **13**

B.1 Model Configuration and Optimization 13

B.2 Evaluation Metrics 14

B.3 Computation of the DTM Prior 14

C Additional Results across Forecast Horizons **15**

D Broader Impacts **15**

E Limitations and Future Work **15**

F AI Use Statement **16**

G Ethics Statement **16**

H Reproducibility Statement **16**

A DATASET CONSTRUCTION

Data sources and diversity. The dataset integrates satellite records from the TU Delft TOLEOS project (Siemes et al., 2023) with solar and geomagnetic indices obtained from GFZ (Bartels, 1957; Matzka et al., 2021). The satellite records contain thermospheric density estimates derived from onboard accelerometer measurements and five orbital variables: latitude, longitude, altitude, local solar time, and argument of latitude. The six space-weather indices are K_p, ap, Ap, sunspot number, and observed and adjusted $F_{10.7}$, aligned with the satellite records by timestamp. Collectively spanning 2000–2025, the observations cover quiet and active periods across multiple phases of solar activity and varying geomagnetic conditions. Together with differences in satellite altitude and orbital geometry, this coverage captures recurring along-orbit oscillations, shifts in background density levels, and transient enhancements, posing challenges for accurate long-horizon forecasting.

Data cleaning and resampling. Each satellite is processed independently. We remove records with missing required observation fields or nonpositive density, then retain the first valid observation in each occupied 5-minute bin. Timestamps are aligned to the bin starts, and density observations are neither averaged nor interpolated. Missing $F_{10.7}$ values are forward-filled from earlier available records. The processed dataset contains 4,895,496 observations before forecasting-window selection. Table 4 summarizes the temporal coverage and observation counts.

Temporal segmentation and chronological splits. To preserve temporal continuity, we start a new segment whenever adjacent retained timestamps differ by more than 5 minutes and construct forecasting windows entirely within uninterrupted segments. Segments are sorted chronologically and assigned to training, validation, and test sets with a target ratio of 80:10:10. Each segment belongs to a single split, preventing windows from crossing split boundaries. The realized proportions may differ from the target ratio because split boundaries are restricted to segment boundaries.

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 MODEL CONFIGURATION AND OPTIMIZATION

Table 5 summarizes the default DRC configuration, shared across satellites and forecast horizons. The space-weather encoder aggregates the maximum, mean, and standard deviation of each index over four trailing windows. A two-layer BiLSTM captures cross-scale dependencies, and average pooling produces a context vector that guides FiLM modulation and residual gating. The correction head is initialized to output zero, so the initial predictions coincide with the DTM-2020 prior estimates. Training then learns context-dependent corrections to refine these estimates.

Table 4: Dataset coverage and observation counts after preprocessing and before forecasting-window selection. Temporal coverage includes periods with missing observations.

Satellite	Period	Observations
CHAMP	2000–2010	981,418
GOCE	2009–2013	366,747
GRACE-A	2002–2017	1,439,680
GRACE-B	2002–2017	1,313,978
GRACE-FO	2018–2025	793,673
Total	2000–2025	4,895,496

Table 5: Default DRC configuration.

Parameter	Value
Transformer layers / attention heads	3 / 4
Hidden / feed-forward dimension	64 / 128
Patch length / stride	16 / 8
Dropout	0.05
Space-weather aggregation windows	1, 3, 6, 24 hours
Statistics per index and window	Maximum, mean, standard deviation
Context encoder	Two-layer BiLSTM + average pooling
Anchor	DTM-2020
Optimizer	Adam
Learning-rate schedule	One-cycle
Peak learning rate	10^{-4}
Warm-up fraction	0.2
Batch size	128
Maximum training epochs	100
Early-stopping patience	20 epochs
Random seed	2021

B.2 Evaluation Metrics

We evaluate forecasts in the original density scale using mean absolute error (MAE), mean squared error (MSE), and root relative squared error (RSE). Let ρ_j and $\hat{\rho}_j$ denote the true and predicted densities for the jth evaluated value, respectively. Given n values across all test samples and forecast steps, we define

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{n}\sum_{j=1}^{n}|\hat{\rho}_j - \rho_j|, \\ \mathrm{MSE} &= \frac{1}{n}\sum_{j=1}^{n}(\hat{\rho}_j - \rho_j)^2, \\ \mathrm{RSE} &= \sqrt{\frac{\sum_{j=1}^{n}(\hat{\rho}_j - \rho_j)^2}{\sum_{j=1}^{n}(\rho_j - \bar{\rho})^2}}, \end{aligned} \tag{8}$$

where $\bar{\rho} = \frac{1}{n}\sum_{j=1}^{n}\rho_j$ is the mean true density. MAE and MSE have units of $\mathrm{kg/m^3}$ and $(\mathrm{kg/m^3})^2$, respectively, while RSE is dimensionless. Lower values indicate better performance for all metrics.

B.3 Computation of the DTM Prior

DRC incorporates prior knowledge from DTM-2020 through density estimates computed over the forecast horizon. At each forecast origin, the latest available solar and geomagnetic index values required by DTM-2020 are held constant, while future satellite positions are obtained by propagating the orbital state available at that time. DTM-2020 is then evaluated at each future time and position. Thus, the prior density varies along the predicted orbit even though the index values remain fixed.

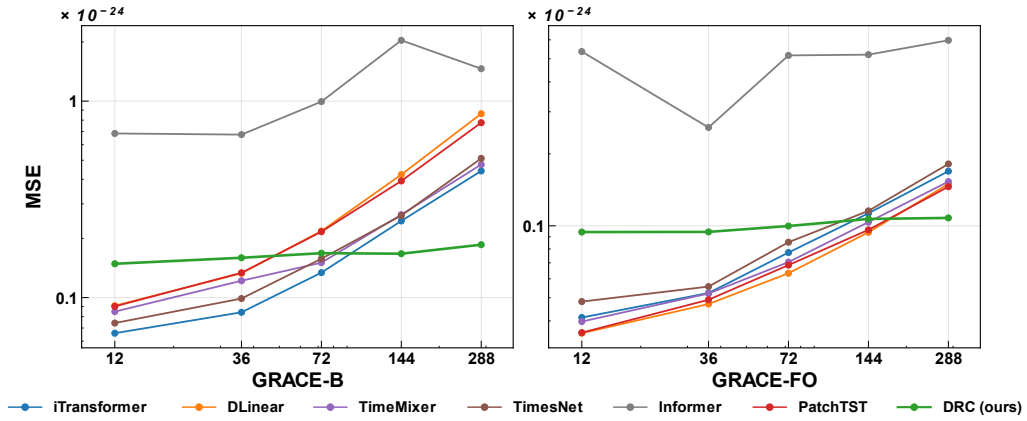

Figure 6: MSE across forecast horizons on GRACE-B and GRACE-FO, complementing Figure 5. Each step represents 5 minutes, with horizons spanning 1–24 hours. MSE is shown on logarithmic axes in units of $10^{-24}\,(\mathrm{kg/m^3})^2$. DRC maintains relatively stable errors and achieves the lowest MSE at 24 hours, while some baselines perform better at shorter horizons.

The resulting density estimates are transformed into $\log_{10}$ space and combined with the learned residual corrections. The same procedure is used during training and evaluation, without accessing future density observations or future observed space-weather indices.

C Additional Results across Forecast Horizons

Figure 6 shows that DRC maintains relatively stable MSE across forecast horizons on GRACE-B and GRACE-FO. Although several deep forecasting baselines achieve lower errors at short horizons, their MSE increases substantially as the horizon extends. DRC exhibits only modest error growth and achieves the lowest MSE among all compared methods at 288 steps (24 hours) on both satellites. This contrast is particularly evident on GRACE-FO, where DRC initially trails several baselines but outperforms them at the longest horizon. These results complement the main experiments by demonstrating that DRC's strength lies in sustaining forecasting accuracy over extended horizons, even when it offers limited advantages for short-term prediction.

D Broader Impacts

Improved thermospheric density forecasting could support accurate atmospheric drag estimation, orbit prediction, and satellite operations. The curated dataset facilitate research on forecasting under changing space-weather conditions. However, density prediction improvements do not directly establish operational reliability. Applications to collision avoidance or reentry prediction require additional validation, uncertainty assessment, and integration with orbit propagation systems.

E Limitations and Future Work

DRC's benefits vary across satellites and forecast horizons, and some deep forecasting models perform better at shorter horizons. The DTM prior holds solar and geomagnetic indices constant during forecasting, limiting its ability to anticipate abrupt future disturbances. Evaluation is limited to five satellite datasets and horizons up to 24 hours. Future work will examine longer horizons, additional missions, uncertainty quantification, and forecasted space-weather inputs. Repeated runs with matched model-selection budgets would further strengthen the empirical assessment.

F AI USE STATEMENT

We used large language model tools solely for translation, language polishing, and LaTeX typesetting. These tools were not used for experimental design, data collection or processing, result interpretation, numerical calculation, figure or table generation, or any other activity that could affect the authenticity of the experimental results.

G ETHICS STATEMENT

This study uses satellite observations and space-weather indices from publicly accessible scientific sources and does not involve human participants or personal data. We acknowledge the original data providers, and any redistribution must respect the applicable source licenses and usage conditions. Reported forecasting improvements should be interpreted within the evaluated settings and should not be treated as guarantees of safety in operational satellite systems.

H REPRODUCIBILITY STATEMENT

The main text describes the DRC architecture and training objective. Appendix A documents the data sources, preprocessing, temporal segmentation, and chronological splits. Experimental settings and model configurations are detailed in Appendix B.1 and Table 5. Metric definitions and the computation of the DTM prior are provided in Appendices B.2 and B.3, respectively. Together, these descriptions specify the principal components of the forecasting and evaluation pipeline.