REVISE, DON'T PREDICT: ENVIRONMENT RESPONSE FOR TIME SERIES FORECASTING WITH FUTURE EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

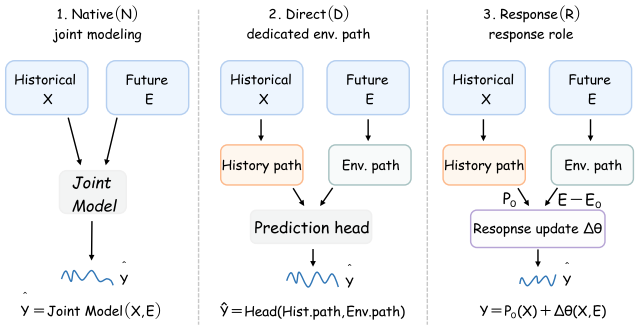
In time series forecasting, future exogenous variables describe the environment over the forecasting horizon, offering information that historical observations alone cannot provide. Existing forecasters exploit such variables through feature fusion, cross-variable interaction, or conditional generation, focusing mainly on *how* future exogenous information enters a forecaster. We study a complementary question: *what should a dedicated future-environment pathway predict—the full target, or the change to a forecast already supported by history?* We formalize the latter as ***Environment Response***: history first establishes a reference forecast, and the future environment guides how this reference should change. Under squared loss, this role corresponds to the gap between the history-only and environment-aware Bayes predictors, isolating what the future environment adds beyond history. Building on this formulation, we propose **Environment-Guided Flow (EGF)**, which constructs the reference from complementary historical views, encodes the future environment relative to a reference environment, and progressively refines the reference forecast through conditional flow. Beyond EGF, Environment Response can be integrated into existing TSF-X forecasters as a lightweight adapter: on three representative backbones, assigning a matched environment pathway the Response role lowers MSE relative to using it for full-target prediction on 9 of 12 datasets for each backbone, and target-aligned diagnostics show that it moves environment-induced forecast changes closer to the target-required correction. On 12 real-world benchmarks, EGF achieves the best result on 23 of 24 dataset–metric averages.

1 INTRODUCTION

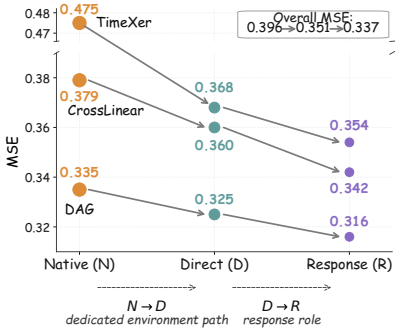
Forecasting is often framed as extrapolating an unknown future from the past. Recent forecasting models have substantially advanced the modeling of historical dynamics (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022; Nie et al., 2023; Zeng et al., 2023; Lin et al., 2024a;b; Chen et al., 2024; Xia et al., 2025; Liu et al., 2026). Yet in many operational systems, the future is not entirely unknown: calendar and holiday information may be known in advance, while weather, load, and generation conditions can be available over the prediction horizon (Lim et al., 2021; Olivares et al., 2023; Vagropoulos et al., 2016; Lago et al., 2021; Arango et al., 2025; Wu et al., 2024).

Forecasting with such horizon-available exogenous variables, termed time series forecasting with future exogenous variables (TSF-X) by Qiu et al. (2026), has consequently received increasing attention. We refer to these horizon-aligned variables as the *future environment*, as they describe the conditions under which the target will evolve. Existing methods have developed increasingly sophisticated mechanisms for incorporating such information, including feature conditioning and fusion (Olivares et al., 2023; Das et al., 2023; Lim et al., 2021; Arango et al., 2025; Tayal et al., 2024), endogenous–exogenous interaction modeling (Wang et al., 2024; Zhou et al., 2025; Qiu et al., 2026), and conditional generation (Li et al., 2026; Cheng et al., 2026). These advances largely address *how* future exogenous information enters a forecaster. We study a complementary question: *what should a dedicated future-environment pathway be responsible for predicting—the full target, or only the change to a forecast that history already supports?*

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



(a) **Predictive roles.** Native predicts jointly; Direct uses a dedicated environment pathway for full-target prediction; Response uses it for a reference-relative update.



(b) **Matched transfers.** MSE averaged over 12 datasets for each backbone; Overall MSE reports the mean over backbones.

Figure 1: **Predictive roles and matched-transfer results.**

To examine this question, we compare three configurations (Figure 1a). *Native* is the original future-aware forecaster, which models history and the future environment jointly without a dedicated environment pathway. *Direct* adds such a pathway but still uses it for full-target prediction. *Response* anchors prediction on a history-supported reference forecast and assigns the same pathway only to predicting its update. Direct and Response thus instantiate the two answers above: they use parameter- and receptive-field-matched adapters under two different predictive formulations. Moving from Native to Direct evaluates the benefit of adding a dedicated pathway, whereas moving from Direct to Response compares the two predictive formulations. With parameter- and receptive-field-matched temporal adapters on DAG (Qiu et al., 2026), CrossLinear (Zhou et al., 2025), and TimeXer (Wang et al., 2024), the first step reduces the average MSE from 0.396 to 0.351, and the second further reduces it to 0.337 (Figure 1b). This observation motivates the central hypothesis of this work: *future-environment information may be more effective when used to revise what history already supports than when asked to participate in predicting the full target.*

We refer to this predictive role as ***Environment Response***: history establishes a reference forecast, and future-environment information guides how this reference should change. The role admits a simple population interpretation. Under squared loss, the Bayes predictor decomposes as

$$\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{E}] = \underbrace{\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]}_{\text{History-supported forecast}} + \underbrace{\Delta^\star(\mathbf{X}, \mathbf{E})}_{\text{Oracle environment response}}, \tag{1}$$
$$\Delta^\star(\mathbf{X}, \mathbf{E}) \triangleq \mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{E}] - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}].$$

The response $\Delta^\star$ isolates what the future environment adds beyond history: it has zero conditional mean given $\mathbf{X}$, and $\mathbb{E}\|\Delta^\star(\mathbf{X}, \mathbf{E})\|_2^2$ equals the reduction in Bayes squared-error risk obtained by additionally conditioning on $\mathbf{E}$. The decomposition extends to general losses by replacing conditional means with conditional Bayes acts, such as the conditional medians associated with the absolute-error objective used in this work (Appendix A.1). This identity motivates a reference-plus-update inductive bias.

We realize this inductive bias in **Environment-Guided Flow (EGF)**, which proceeds in three stages. *History Belief* constructs the reference forecast from complementary historical hypotheses, namely multi-scale temporal views and a phase view built on regularly spaced anchors, combined by adaptive fusion at each future time step. Preserving multiple hypotheses matters because different future conditions can make different historical patterns relevant. *Environment Transport* encodes the future environment as a contrast to a reference environment, which holds each exogenous channel at its recent historical mean, and uses this contrast to reallocate weight among the historical hypotheses. *Flow Refinement* then evolves the reference forecast through a conditional flow guided by the transported response, and a learned gate controls how much of the proposed correction is applied.

Our main contributions are summarized as follows:

- We identify the predictive role of a dedicated future-environment pathway as a design choice in TSF-X, and formalize Environment Response, in which this pathway updates a history-supported forecast rather than predicting the full target.
- We propose EGF, which realizes Environment Response through complementary historical hypotheses, reference-relative environment transport, and gated flow refinement.
- Across three backbones, matched transfers show that assigning the Response role lowers MSE relative to full-target prediction on 9 of 12 datasets for each backbone. EGF achieves the best result on 23 of 24 dataset–metric averages over 12 TSF-X benchmarks, and replacing any of its three stages raises error on all six ablation datasets.

2 Related Work

Time Series Forecasting with Exogenous Variables. Forecasting with exogenous variables incorporates information beyond the target history through feature fusion, attention, and cross-variable modeling. Representative methods include TFT (Lim et al., 2021), TiDE (Das et al., 2023), TimeXer (Wang et al., 2024), and CrossLinear (Zhou et al., 2025). Additive treatments of exogenous effects have a long history: classical transfer-function models add a parametric exogenous term to a history-driven model (Box et al., 2015), and NBEATSx (Olivares et al., 2023) exposes trend, seasonal, and exogenous contributions through interpretable basis expansion. Environment Response is also additive, but defines the update relative to the history-only forecast and lets it depend on both history and the future environment.

When exogenous information is available over the forecasting horizon, temporal alignment and conditioning become central. ExoTST (Tayal et al., 2024) integrates exogenous signals with past context through cross-attention and cross-temporal fusion, while ChronosX (Arango et al., 2025) uses modular input/output injection, with future covariates adjusting pretrained output logits. DAG (Qiu et al., 2026) models temporal correlations between historical and future exogenous variables and channel correlations between exogenous and endogenous variables. GCGNet (Li et al., 2026) models joint temporal–channel dependencies through graph-consistent generation and graph-based refinement. KITE (Cheng et al., 2026) constructs a history-aware source distribution for exogenous-conditioned flow matching and regularizes covariate interactions with statistical priors. These works focus on how future covariates are exploited; we instead study the predictive role assigned to them.

Forecast Refinement and Correction. Forecast refinement spans post-processing, test-time adaptation, and residual learning. PIR (Liu et al., 2025) estimates instance-level forecast quality and revises predictions using historical and covariate context. δ-Adapter (Liang et al., 2026b) learns bounded input adjustments and output residual corrections for frozen forecasters. TAFAS (Kim et al., 2025) performs test-time adaptation through gated calibration, while COSA (Im & Kwon, 2026) conditions gated output corrections on statistics from recently observed targets. UEC-STD (Nguyen et al., 2026) corrects trend and seasonal forecast components to mitigate autoregressive error accumulation. DeepBooTS (Liang et al., 2026a) uses end-to-end residual boosting across network blocks. Environment Response differs by assigning horizon-available future exogenous information to reference-relative updating rather than generic forecast correction.

Generative Forecasting and Flow Matching. Among diffusion and flow forecasters (Rasul et al., 2021a;b; Li et al., 2022), TSDiff (Kollovieh et al., 2023) refines the predictions of base forecasters with a learned implicit density, TSFlow (Kollovieh et al., 2025) replaces the isotropic Gaussian source of flow matching with Gaussian-process priors, and KITE starts from a history-aware source. Flow Refinement also builds on flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Tong et al., 2024), but uses it at inference as a deterministic path that starts at the history-supported reference forecast, so that the gated displacement along this path is the environment-conditioned update.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND OVERVIEW

For a batch of size B, let $\mathbf{X} \in \mathbb{R}^{B\times L\times C_x}$, $\mathbf{E} \in \mathbb{R}^{B\times H\times C_e}$, and $\mathbf{Y} \in \mathbb{R}^{B\times H}$ denote the historical input, future environment, and future target, respectively, where L and H are the history and prediction lengths and $C_x = C_e + 1$ includes the target channel. Each sample is normalized using channel-wise statistics computed from its recent history, and the same transformation is applied to the future environment and target. In these normalized coordinates, the reference environment $\mathbf{E}_0 = \mathbf{0}$ corresponds to holding each future exogenous channel at its historical local mean.

EGF follows a reference-plus-update formulation:

$$\widehat{\mathbf{Y}} = \mathbf{P}_0(\mathbf{X}) + \Delta_\theta(\mathbf{X}, \mathbf{E}), \tag{2}$$

where $\mathbf{P}_0 \in \mathbb{R}^{B\times H}$ is the history-supported reference forecast and Δ_θ is its environment-conditioned update. As shown in Figure 2, EGF contains three stages. *History Belief* constructs $\mathbf{P}_0$ together with retained historical context H_X. *Environment Transport* interprets the future environment relative to $\mathbf{E}_0$ using the history-supported view information in H_X. *Flow Refinement* then evolves $\mathbf{P}_0$ conditioned on H_X and the transported response to obtain the final prediction.

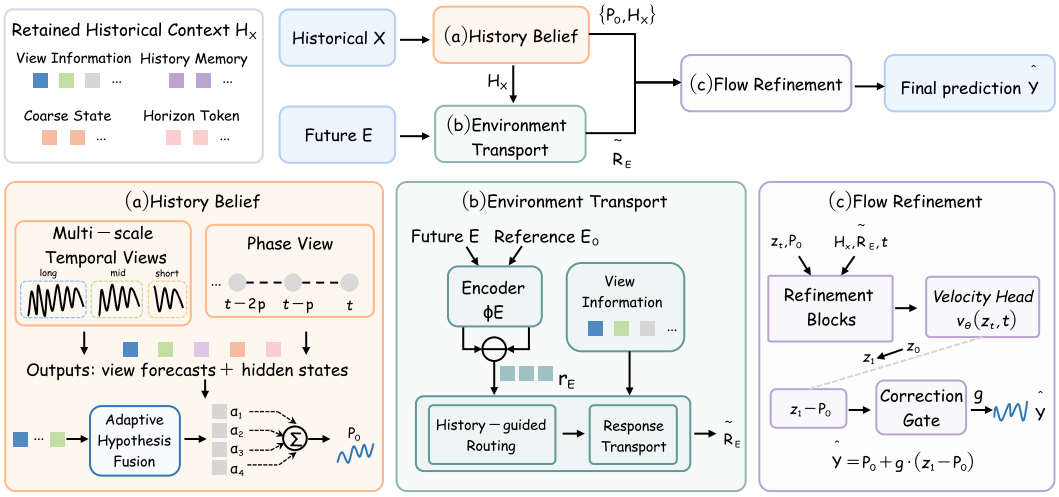


Figure 2: **Overview of Environment-Guided Flow (EGF).**

3.2 HISTORY BELIEF

History Belief builds the reference forecast $\mathbf{P}_0$ from history alone, retaining complementary hypotheses that different future conditions may make relevant: three multi-scale temporal views and one phase view, combined by adaptive fusion.

Multi-scale temporal views. The three temporal views model short-, mid-, and long-range history. For each view k, a strided patch projection converts the selected history into tokens, followed by a Transformer encoder. Position embeddings over the forecasting horizon query the encoded history through a view decoder, producing a horizon-aligned forecast $\mathbf{P}^{(k)} \in \mathbb{R}^{B\times H}$ and state $\mathbf{H}^{(k)} \in \mathbb{R}^{B\times H\times D}$, where D is the hidden dimension.

Phase view. The phase view ($v = 4$) models explicitly aligned historical changes. For each future time point, it retrieves valid historical anchor pairs separated by a fixed interval p. If $\mathbf{a}_j$ is the target level at an anchor and $\boldsymbol{\delta}_j$ is its change from the preceding anchor, the phase forecast is

$$\mathbf{P}^{(4)} = \sum_j \boldsymbol{\omega}_j \odot (\mathbf{a}_j + \boldsymbol{\gamma}_j \odot \boldsymbol{\delta}_j), \tag{3}$$

where $\boldsymbol{\omega}_j$ are softmax-normalized anchor weights and $\boldsymbol{\gamma}_j$ are tanh-bounded gains. The corresponding weighted anchor representations form the horizon-aligned state $\mathbf{H}^{(4)}$.

Adaptive hypothesis fusion. Let $\mathbf{P}^{(v)}$ and $\mathbf{H}^{(v)}$, $v = 1, \ldots, 4$, denote the four view forecasts and states. A fusion network produces softmax weights $\boldsymbol{\alpha}$ for each future time point and forms

$$\mathbf{P}_0 = \sum_{v=1}^{4} \boldsymbol{\alpha}^{(v)} \odot \mathbf{P}^{(v)}. \tag{4}$$

Besides $\mathbf{P}_0$, History Belief retains the view states and additional history representations required by later stages; we denote this retained historical context by H_X (Appendix B.1).

3.3 Environment Transport

Rather than fuse future exogenous variables directly into full-target prediction, Environment Transport represents the supplied environment relative to $\mathbf{E}_0$ and uses its contrast with view-level historical information in H_X to reallocate the history-supported hypotheses. The resulting transported response is $\widetilde{\mathbf{R}}_E$.

Reference-relative encoding. The supplied environment $\mathbf{E}$ and reference $\mathbf{E}_0$ are processed by the same horizon-aware encoder ϕ_E. Their encoded difference is

$$\mathbf{r}_E = \phi_E(\mathbf{E}) - \phi_E(\mathbf{E}_0), \qquad \mathbf{r}_E \in \mathbb{R}^{B \times H \times D}. \tag{5}$$

The shared encoder combines position embeddings over the forecasting horizon, temporal convolutions, and feed-forward transformations. Thus, $\mathbf{r}_E$ captures the trajectory-level contrast between the supplied and reference environments.

History-guided routing. We compare $\mathbf{r}_E$ with the horizon-aligned view states retained in H_X using learned projections and cosine similarity. Starting from the history-only fusion weights $\boldsymbol{\alpha}$, the environment produces a new allocation over the four hypotheses:

$$\widetilde{\boldsymbol{\alpha}}^{(v)} = \mathrm{Softmax}_v\Big(\log \boldsymbol{\alpha}^{(v)} + \beta \,\mathrm{Cos}\Big(\mathrm{Proj}_E(\mathbf{r}_E), \mathrm{Proj}_H(\mathbf{H}^{(v)})\Big)\Big), \tag{6}$$

where β is the routing strength and softmax is taken over the four views. We further scale the redistribution by a KL-based trust weight with strength τ:

$$\Delta\mathbf{w} = \exp[-\tau D_{\mathrm{KL}}(\widetilde{\boldsymbol{\alpha}} \| \boldsymbol{\alpha})] \odot (\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}). \tag{7}$$

Positive and negative entries of $\Delta\mathbf{w}$ indicate which history-supported hypotheses become more or less relevant under the supplied environment.

Response transport. The redistribution induces a forecast-space shift $\mathbf{d}_P = \sum_v \Delta\mathbf{w}^{(v)} \odot \mathbf{P}^{(v)}$ and a feature-space shift $\mathbf{d}_H = \sum_v \Delta\mathbf{w}^{(v)} \odot \mathbf{H}^{(v)}$. Together with the encoded environment contrast and cross-view disagreement, these signals are processed by an adapter and a confidence head:

$$\widetilde{\mathbf{R}}_E = \mathbf{r}_E + s_{\mathrm{tr}} \|\Delta\mathbf{w}\|_1 \odot \mathbf{c} \odot \mathbf{A}, \tag{8}$$

where $\mathbf{A}$ is the feature update and $\mathbf{c}$ is the confidence weight for each future time point, and s_{tr} is a fixed transport scale.

3.4 Flow Refinement

Rather than predict the entire correction in one pass, Flow Refinement evolves $\mathbf{P}_0$ progressively, so that each step conditions on the partially corrected forecast together with the retained historical context H_X and the transported response $\widetilde{\mathbf{R}}_E$.

For the deterministic prediction path, let $\mathbf{z}_t \in \mathbb{R}^{B \times H}$ denote the prediction state at refinement time $t \in [0, 1]$. We define

$$\frac{d\mathbf{z}_t}{dt} = v_\theta\Big(\mathbf{z}_t, t \mid \mathbf{P}_0, H_X, \widetilde{\mathbf{R}}_E\Big), \qquad \mathbf{z}_0 = \mathbf{P}_0. \tag{9}$$

The vector field models dependencies across future time points, retrieves historical information from H_X, and attends to $\widetilde{\mathbf{R}}_E$, which also modulates the intermediate refinement state through scale and shift.

We integrate the vector field for S steps from $\mathbf{P}_0$ to obtain the terminal proposal $\mathbf{z}_1$, whose displacement $\mathbf{z}_1 - \mathbf{P}_0$ is the proposed correction. A correction gate $\mathbf{g}$ controls how much of it is applied:

$$\widehat{\mathbf{Y}} = \mathbf{P}_0 + \underbrace{\mathbf{g} \odot (\mathbf{z}_1 - \mathbf{P}_0)}_{\Delta_\theta(\mathbf{X},\mathbf{E})}. \tag{10}$$

The gate is conditioned on the proposed correction, the relevant historical states in H_X, and $\widetilde{\mathbf{R}}_E$. Detailed module inputs and the integration rule are provided in Appendix B.3.

3.5 Training Objective

We train all three stages jointly. The forecasting objective supervises the final prediction and the history-supported reference, while flow matching trains the refinement vector field:

$$\mathcal{L} = \mathrm{MAE}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_b \,\mathrm{MAE}(\mathbf{P}_0, \mathbf{Y}) + \lambda_f \mathcal{L}_{\mathrm{flow}}, \tag{11}$$

where λ_b and λ_f control reference supervision and flow matching, respectively.

For flow matching (Lipman et al., 2023), the training source is sampled around the history-supported reference: $\widetilde{\mathbf{z}}_0 = \mathbf{P}_0 + \sigma_0 \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For $t \sim \mathcal{U}(0,1)$, the linear path $\widetilde{\mathbf{z}}_t = (1-t)\widetilde{\mathbf{z}}_0 + t\mathbf{Y}$ has target velocity $\mathbf{Y} - \widetilde{\mathbf{z}}_0$, giving

$$\mathcal{L}_{\mathrm{flow}} = \mathrm{MSE}\Big(v_\theta\Big(\widetilde{\mathbf{z}}_t, t \mid \mathbf{P}_0, H_X, \widetilde{\mathbf{R}}_E\Big), \mathbf{Y} - \widetilde{\mathbf{z}}_0\Big). \tag{12}$$

The stochastic source is used only for flow-matching supervision. Prediction starts from the unperturbed $\mathbf{P}_0$ and follows the deterministic refinement path in Figure 2(c).

4 Experiments

We organize our experiments around four research questions: **RQ1:** How does EGF compare with existing TSF-X methods across datasets and application domains? **RQ2:** Does assigning matched environment adapters the Response role improve forecasting across different backbones? **RQ3:** Does Response bring environment-induced forecast changes closer to the target-required correction? **RQ4:** How do the three stages of EGF contribute to forecasting performance?

Setup. We evaluate EGF on 12 real-world TSF-X datasets. The five electricity-price datasets (NP, PJM, BE, FR, and DE) come from the EPF benchmark (Lago et al., 2021; Wang et al., 2024). The remaining datasets follow DAG (Qiu et al., 2026) and cover energy production (Energy), reservoir forecasting (Colbun and Rapel), and wind power (Sdwpfm1, Sdwpfm2, Sdwpfh1, and Sdwpfh2). The wind-power datasets use ERA5 meteorological covariates (Hersbach et al., 2020). Dataset splits and forecasting horizons follow DAG, and we report MSE and MAE.

We compare EGF with nine recent forecasting models: KITE (Cheng et al., 2026), DAG (Qiu et al., 2026), GCGNet (Li et al., 2026), CrossLinear (Zhou et al., 2025), TimeKAN (Huang et al., 2025), Amplifier (Fei et al., 2025), xPatch (Stitsyuk & Choi, 2025), DUET (Qiu et al., 2025), and TimeXer (Wang et al., 2024). Dataset and implementation details are provided in Appendix C.

4.1 Forecasting Performance (RQ1)

Table 1 compares EGF with nine forecasting baselines on 12 TSF-X datasets, with results averaged over two prediction horizons. EGF achieves the lowest MSE on 11 of 12 datasets and the lowest MAE on all 12 datasets, ranking first on 23 of the 24 dataset–metric averages. Relative to the strongest baseline, GCGNet, it lowers the dataset-averaged MSE and MAE by 10.2% and 9.1%, respectively. The exception is Rapel MSE, where CrossLinear is lower (0.196 vs. 0.237). At the horizon level, EGF ranks first, including ties, on 38 of 48 entries (Table 10).

Table 1: Forecasting results on 12 TSF-X datasets, averaged over two horizons; EGF also averages five seeds. † denotes our reproductions using the official DAG code and settings; other baseline results are paper-reported. Red and blue mark the best and second-best results. Lower is better.

Models	EGF (ours)		KITE (2026)		DAG† (2026)		GCGNet (2026)		CrossLinear† (2025)		TimeKAN (2025)		Amplifier (2025)		xPatch (2025)		DUET (2025)		TimeXer† (2024)	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	**0.299**	**0.307**	0.325	0.323	0.366	0.348	0.346	0.337	0.372	0.388	0.406	0.419	0.419	0.418	0.378	0.370	0.411	0.407	0.435	0.377
PJM	**0.084**	**0.174**	0.096	0.178	0.093	0.180	0.093	0.186	0.112	0.222	0.139	0.263	0.137	0.246	0.104	0.194	0.102	0.197	0.108	0.198
BE	**0.388**	**0.256**	0.427	0.286	0.423	0.280	0.423	0.287	0.481	0.338	0.548	0.407	0.558	0.413	0.495	0.333	0.514	0.354	0.455	0.291
FR	**0.378**	**0.206**	0.387	0.225	0.414	0.220	0.405	0.230	0.483	0.298	0.547	0.374	0.553	0.408	0.490	0.293	0.495	0.327	0.427	0.241
DE	**0.318**	**0.342**	0.350	0.370	0.377	0.373	0.387	0.387	0.484	0.451	0.473	0.446	0.473	0.441	0.474	0.433	0.482	0.430	0.475	0.418
Energy	**0.088**	**0.226**	0.111	0.257	0.124	0.268	0.122	0.262	0.239	0.402	0.218	0.381	0.233	0.389	0.148	0.303	0.202	0.367	0.167	0.319
Sdwpfm1	**0.381**	**0.417**	0.417	0.451	0.424	0.459	0.416	0.449	0.423	0.502	0.447	0.534	0.428	0.490	0.630	0.568	0.599	0.571	0.701	0.609
Sdwpfm2	**0.410**	**0.438**	0.475	0.483	0.477	0.483	0.458	0.479	0.533	0.573	0.497	0.564	0.490	0.512	0.516	0.495	0.514	0.490	0.803	0.653
Sdwpfh1	**0.365**	**0.427**	0.441	0.491	0.449	0.486	0.414	0.468	0.557	0.593	0.577	0.638	0.536	0.598	0.498	0.505	0.539	0.516	0.750	0.648
Sdwpfh2	**0.425**	**0.470**	0.500	0.526	0.533	0.537	0.472	0.501	0.537	0.574	0.646	0.671	0.521	0.581	0.610	0.551	0.647	0.566	0.891	0.719
Colbun	**0.080**	**0.117**	0.088	0.172	0.106	0.163	0.098	0.154	0.126	0.195	0.128	0.175	0.173	0.245	0.153	0.210	0.198	0.266	0.145	0.236
Rapel	0.237	**0.257**	0.244	0.294	0.234	0.303	0.212	0.259	**0.196**	0.274	0.249	0.310	0.257	0.322	0.333	0.381	0.269	0.326	0.344	0.362
1st Count	**11**	**12**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Table 2: Native–Direct–Response matched transfers across 12 datasets. Entries average two horizons per dataset. Red and blue mark the best and second-best values within each backbone, respectively. Ties are determined at the reported precision, and first-place counts include ties. Lower is better.

Models	DAG (2026)						CrossLinear (2025)						TimeXer (2024)					
Conditions	N		D		R		N		D		R		N		D		R	
Datasets	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	0.366	0.348	0.333	0.332	**0.312**	**0.323**	0.372	0.388	0.348	0.372	**0.311**	**0.336**	0.435	0.377	0.343	0.336	**0.331**	**0.335**
PJM	0.093	0.180	**0.092**	**0.177**	0.095	0.187	0.112	0.222	0.119	0.212	**0.102**	**0.201**	0.108	0.198	**0.101**	**0.192**	0.106	0.201
BE	0.423	0.280	0.419	**0.273**	**0.416**	0.286	0.481	0.338	0.452	0.328	**0.440**	**0.311**	0.455	0.291	0.429	0.278	**0.404**	**0.277**
FR	0.414	0.220	0.413	**0.218**	**0.399**	**0.218**	0.483	0.298	0.449	0.294	**0.425**	**0.250**	0.427	0.241	**0.422**	**0.236**	0.434	0.242
DE	0.377	0.373	0.369	0.366	**0.335**	**0.353**	0.484	0.451	0.424	0.417	**0.420**	**0.402**	0.475	0.418	0.416	**0.389**	**0.411**	0.396
Energy	0.124	0.268	0.111	0.256	**0.096**	**0.239**	0.239	0.402	0.295	0.444	**0.149**	**0.306**	0.167	0.319	0.122	0.274	**0.112**	**0.260**
Sdwpfm1	**0.424**	**0.459**	0.431	0.464	0.445	0.480	0.423	0.502	**0.375**	**0.475**	0.434	0.500	0.701	0.609	**0.481**	0.511	0.487	**0.501**
Sdwpfm2	0.477	0.483	0.466	**0.482**	**0.447**	0.491	0.533	0.573	0.513	0.560	**0.502**	**0.557**	0.803	0.653	0.552	0.546	**0.534**	**0.528**
Sdwpfh1	0.449	0.486	**0.411**	**0.450**	0.434	0.475	0.557	0.593	**0.461**	**0.537**	0.483	0.552	0.750	0.648	0.526	0.556	**0.486**	**0.525**
Sdwpfh2	0.533	0.537	0.464	**0.473**	**0.438**	0.491	0.537	0.574	**0.492**	**0.545**	0.499	0.559	0.891	0.719	0.604	0.598	**0.568**	**0.568**
Colbun	0.106	0.163	0.094	0.153	**0.082**	**0.142**	0.126	**0.195**	0.154	0.212	**0.123**	**0.195**	0.145	0.236	0.111	0.197	**0.085**	**0.152**
Rapel	**0.234**	**0.303**	0.294	0.347	0.288	0.331	**0.196**	**0.274**	0.242	0.309	0.211	0.282	0.344	0.362	0.308	0.338	**0.292**	**0.320**
1st Count	2	2	2	**6**	**8**	5	1	2	3	3	**8**	**8**	0	0	3	3	**9**	**9**

4.2 Transferability of Environment Response (RQ2)

To test whether Environment Response transfers beyond EGF, we compare three configurations on DAG, CrossLinear, and TimeXer. **Native (N)** is the original future-aware backbone. **Direct (D)** adds a temporal environment adapter before the backbone's final prediction mapping and uses it for full-target prediction. **Response (R)** anchors on a history-supported reference forecast $\mathbf{P}_r$, encodes the future environment relative to a reference trajectory, and uses the matched adapter to produce a gated update of $\mathbf{P}_r$. For each backbone, Direct and Response use adapters matched in parameter count and temporal receptive field, and the backbone and adapter are trained jointly (Appendix B.4).

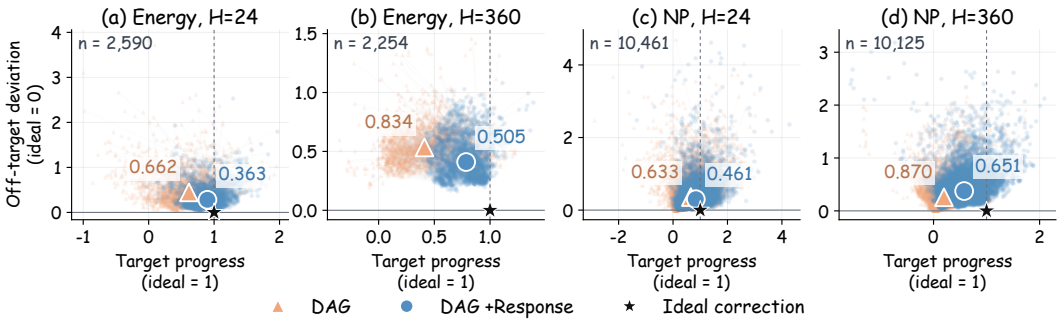


Figure 3: **Target-aligned forecast corrections.** DAG and DAG+Response on Energy and NP ($H = 24, 360$). Large markers show coordinate-wise medians; gray segments connect matched origins. Numbers report median normalized distances to the ideal correction $(1, 0)$.

Table 2 reports the results across all 12 datasets. Response achieves lower MSE than Direct on 9 of 12 datasets for each of the three backbones, and obtains the most MSE wins among the three configurations: 8 for DAG, 8 for CrossLinear, and 9 for TimeXer. Averaged over datasets, Response lowers MSE relative to Direct from 0.325 to 0.316 on DAG, from 0.360 to 0.342 on CrossLinear, and from 0.368 to 0.354 on TimeXer. MAE falls from 0.392 to 0.371 on CrossLinear and from 0.371 to 0.359 on TimeXer; on DAG, average MAE is on par (0.333 vs. 0.335).

4.3 Target Alignment of Forecast Corrections (RQ3)

Quantitative correction geometry. We compare DAG and DAG+Response on Energy and NP at prediction horizons $H = 24$ and 360. Both models are evaluated using a persistence environment, which repeats the last observed exogenous state, as a common diagnostic input. For each model and test origin, the *target-required correction* is the vector from its persistence forecast to the observed target. The *environment-induced correction* is the forecast change obtained by replacing the persistence environment with the supplied future environment.

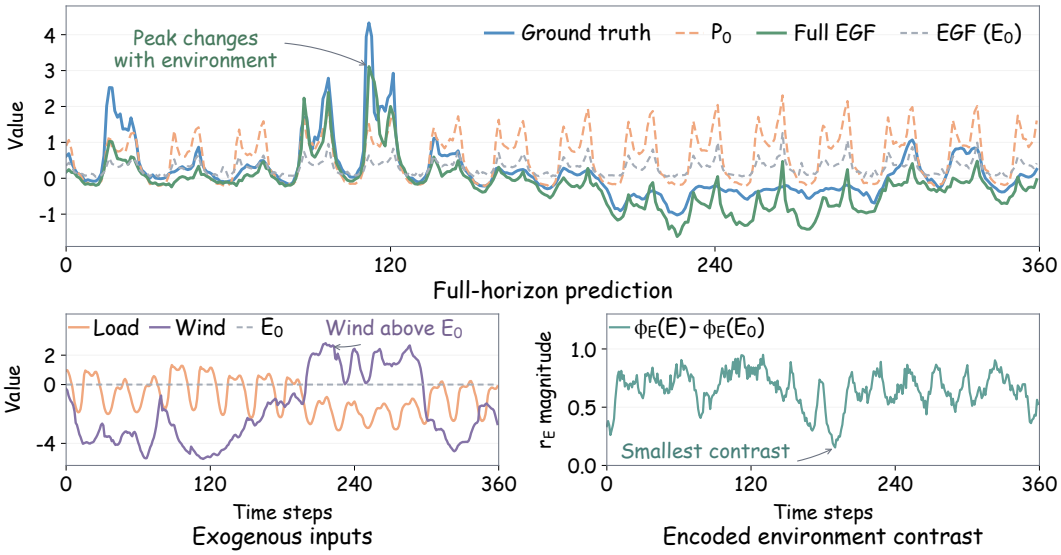


Figure 4: **Environment Response in EGF** on NP ($H = 360$). Bottom: supplied environment, reference environment $\mathbf{E}_0$, and the magnitude of $\mathbf{r}_E$. Top: reference forecast $\mathbf{P}_0$, EGF predictions under the supplied environment and under $\mathbf{E}_0$, and the observed target.

Table 3: Stage-level ablations on six datasets. Each entry averages the two horizons and five seeds. EGF gain is the relative error reduction over each replacement.

Model	NP		PJM		FR		DE		Energy		Sdwpfh1		Average		EGF gain	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
EGF (Ours)	**0.299**	**0.307**	**0.084**	**0.174**	**0.378**	**0.206**	**0.318**	**0.342**	**0.088**	**0.226**	**0.365**	**0.427**	**0.255**	**0.280**	–	–
Single-view History Belief	0.354	0.334	0.093	0.181	0.442	0.228	0.348	0.368	0.114	0.266	0.382	0.446	0.289	0.304	11.8%	7.9%
Direct Environment Fusion	0.310	0.314	0.087	0.178	0.390	0.209	0.335	0.349	0.111	0.253	0.431	0.466	0.277	0.295	7.9%	5.1%
One-step MLP Refinement	0.345	0.325	0.089	0.177	0.388	0.212	0.325	0.344	0.106	0.247	0.410	0.443	0.277	0.291	7.9%	3.8%

Figure 3 decomposes the environment-induced correction into *target progress* along the required direction and normalized *off-target deviation* orthogonal to it. The point $(1, 0)$ represents the ideal correction. Its distance from each measured correction equals the ratio of the Euclidean forecast error under the supplied environment to that under persistence (Appendix A.3). Response reduces the median normalized distance in all four settings, by 25–45%, bringing environment-induced forecast changes closer to the corrections required by the observed targets.

Qualitative response behavior. Figure 4 illustrates how the future environment changes an EGF forecast on an NP example with $H = 360$. The lower panels show the supplied load and wind trajectories, the reference environment $\mathbf{E}_0$, and the magnitude of their encoded contrast $\mathbf{r}_E$; the upper panel compares the history-only reference $\mathbf{P}_0$ with EGF predictions under the supplied environment and under $\mathbf{E}_0$. The reference forecast $\mathbf{P}_0$ primarily follows a recurring historical pattern. With the supplied environment, EGF adjusts several price peaks and the level of the forecast later in the horizon, and replacing the supplied environment with $\mathbf{E}_0$ attenuates several of these adjustments. The difference between these two EGF predictions isolates the effect of changing the environment while holding the history and trained model fixed; $\mathbf{P}_0$ provides the history-only anchor. Appendix E further evaluates EGF with incomplete future environments, alternative reference environments, and test-time environment interventions.

4.4 Contributions of EGF Stages (RQ4)

We replace each EGF stage in turn while keeping the other two unchanged. As shown in Table 3, replacing History Belief with a single temporal view, Environment Transport with direct MLP fusion, or Flow Refinement with a one-step MLP correction raises both MSE and MAE on all six datasets; relative to these replacements, EGF lowers average MSE by 11.8%, 7.9%, and 7.9%, respectively. Finer-grained ablations (Appendix D) show that the phase view accounts for most of the History Belief gain (0.024 of the 0.034 average-MSE gap), and that reference-relative environment encoding outperforms absolute encoding on all six datasets and both metrics.

5 Conclusion

We introduced Environment Response for TSF-X, in which a dedicated future-environment pathway updates a history-supported forecast rather than predicting the full target. Our matched transfers separate two choices that are usually entangled: adding a dedicated environment pathway lowers average MSE from 0.396 to 0.351, and letting the same pathway revise a history-supported forecast lowers it further to 0.337, with both reductions holding on each of the three backbones. EGF realizes this role and achieves the best result on 23 of 24 dataset–metric averages across four application domains.

A limitation of EGF is that it treats the future environment as a given point trajectory, whereas in deployment such trajectories may themselves be uncertain forecasts; extending Environment Response to uncertain environments and probabilistic targets is a natural next step (Appendix I).

Ethics Statement

This work focuses on time series forecasting using existing benchmark datasets and does not involve human subjects or private personal data. We are not aware of specific ethical concerns introduced by the proposed method.

Reproducibility Statement

We provide the model formulation and training objectives in the main text, with detailed model specifications, algorithms, dataset descriptions, experimental settings, and complete quantitative results in the Appendix. The anonymous supplementary material includes the implementation of EGF and code for the main forecasting, matched-transfer, ablation, target-aligned correction, and efficiency experiments.

AI Use Statement

Generative AI tools were used to assist with experimental design, including the target-aligned correction analysis; mathematical derivation and proof checking; language polishing and manuscript consistency checks; and refinement of methodological and implementation details, including the handling of future exogenous information across different backbone models. All AI-assisted mathematical arguments, methodological choices, code, experimental results, and manuscript claims were independently reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

References

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Henri Joseph Senetaire, Ali Caner Turkmen, Oleksandr Shchur, Danielle C. Maddix, Michael Bohlke-Schneider, Bernie Wang, and Syama Sundar Rangapuram. Chronosx: Adapting pretrained time series models with exogenous variables. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2242–2250. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/arango25a.html.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

Peng Chen, Yingying ZHANG, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 10527–10545, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2be6705de7412adf107900add727a795-Paper-Conference.pdf.

Hanyin Cheng, Jingrong Zhou, Yang Shu, and Chenjuan Guo. KITE: Knowledge-guided probabilistic modeling for time series forecasting with exogenous variables. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FYkXNAqq2D.

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

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: bringing attention to neglected low-energy components in time series forecasting. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i11.33267. URL https://doi.org/10.1609/aaai.v39i11.33267.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Songtao Huang, Zhen Zhao, Can Li, and LEI BAI. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 28514–28529, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/46362971bfc3a97e6a271f2eb90fba17-Paper-Conference.pdf.

Jeonghwan Im and Hyuk-Yoon Kwon. Cosa: Context-aware output-space adapter for test-time adaptation in time series forecasting. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 24845–24874, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2a8ce71baac4c89bf9ff479d8240c7d9-Paper-Conference.pdf.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i17.33965. URL https://doi.org/10.1609/aaai.v39i17.33965.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang (Bernie) Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 28341–28364. Curran Associates, Inc., 2023. doi: 10.52202/075280-1232. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5a1a10c2c2c9b9af1514687bc24b8f3d-Paper-Conference.pdf.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 95435–95457, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ee1a1ecc92f35702b5c29dad3dc909ea-Paper-Conference.pdf.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2021.116983. URL https://www.sciencedirect.com/science/article/pii/S0306261921004529.

Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. Generative time series forecasting with diffusion, denoise, and disentanglement. In S. Koyejo, S. Mohamed, A. Agar-

wal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23009–23022. Curran Associates, Inc., 2022. doi: 10.52202/068431-1672. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/91a85f3fb8f570e6be52b333b5ab017a-Paper-Conference.pdf.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. GCGNet: Graph-consistent generative network for time series forecasting with exogenous variables. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EO5jwQ5NCw.

Daojun Liang, Jing Chen, Xiao Wang, Yinglong Wang, and Shuo Li. Deepboots: dual-stream residual boosting for drift-resilient time-series forecasting. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026a. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i28.39509. URL https://doi.org/10.1609/aaai.v40i28.39509.

Daojun Liang, Qi Li, Yinglong Wang, Jing Chen, Hu Zhang, Xiaoxiao Cui, Qizheng Wang, and Shuo Li. The forecast after the forecast: A post-processing shift in time series. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 68159–68189, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6f7d90b1198fec96defd80b5ebd5bc81-Paper-Conference.pdf.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.03.012. URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 106315–106345. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3373. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bfe7998398779dde03cad7a73b1f81b6-Paper-Conference.pdf.

Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang. SparseTSF: Modeling long-term time series forecasting with *1k* parameters. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30211–30226. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/lin24n.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yangyou Liu, Zezhi Shao, Xinyu Chen, Hu Chen, Fei Wang, and Yuankai Wu. PULSE: Generative phase evolution for non-stationary time series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JJIqZzujgE.

Zhiding Liu, Mingyue Cheng, Guanhao Zhao, Jiqian Yang, Qi Liu, and Enhong Chen. Improving time series forecasting via instance-aware post-hoc revision. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 35604–35629. Curran Associates, Inc., 2025. doi: 10.52202/085713-1196. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/331c41353b053683e17f7c88a797701d-Paper-Conference.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Minh Hoang Nguyen, Van Dai Do, Huu Hiep Nguyen, Dung Nguyen, Kien Do, and Hung Le. Reviving error correction in modern deep time-series forecasting. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZlDndRDMHm.

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

Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2022.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207022000413.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 1185–1196, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709325. URL https://doi.org/10.1145/3690624.3709325.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. DAG: A dual correlation network for time series forecasting with exogenous variables. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=jntRRU4zcb.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/rasul21a.html.

Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs M Bergmann, and Roland Vollgraf. Multivariate probabilistic time series forecasting via conditioned normalizing flows. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=WiGQBFuVRv.

Artyom Stitsyuk and Jaesik Choi. xpatch: dual-stream time series forecasting with exponential seasonal-trend decomposition. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i19.34270. URL https://doi.org/10.1609/aaai.v39i19.34270.

Kshitij Tayal, Arvind Renganathan, Xiaowei Jia, Vipin Kumar, and Dan Lu. Exotst: Exogenous-aware temporal sequence transformer for time series prediction. In *2024 IEEE International Conference on Data Mining (ICDM)*, pp. 857–862, 2024. doi: 10.1109/ICDM59182.2024.00105.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Stylianos I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *2016 IEEE International Energy Conference (ENERGYCON)*, pp. 1–6, 2016. doi: 10.1109/ENERGYCON.2016.7514029.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498. Curran Associates, Inc., 2024. doi: 10.52202/079017-0015. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0113ef4642264adc2e6924a3cbbdf532-Paper-Conference.pdf.

Binqing Wu, Weiqi Chen, Wengwei Wang, Bingqing Peng, Liang Sun, and Ling Chen. Weathergnn: Exploiting meteo- and spatial-dependencies for local numerical weather prediction bias-correction. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 2433–2441. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/269. URL https://doi.org/10.24963/ijcai.2024/269. Main Track.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf.

Mingyuan Xia, Chunxu Zhang, Zijian Zhang, Hao Miao, Qidong Liu, Yuanshao Zhu, and Bo Yang. Timeemb: A lightweight static-dynamic disentanglement framework for time series forecasting. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 76227–76258. Curran Associates, Inc., 2025. doi: 10.52202/085713-2560. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6e79ae23fe8a8bb15a2b2972053f4fc3-Paper-Conference.pdf.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

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.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 4120–4131, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736899. URL https://doi.org/10.1145/3711896.3736899.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.),

Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Appendix for Environment-Guided Flow (EGF)

CONTENTS

A Mathematical Properties **17**
- A.1 Population Interpretation of Environment Response . . . 17
- A.2 Reference Environments . . . 18
- A.3 Target-Aligned Correction Geometry . . . 18

B Details of EGF **18**
- B.1 History Belief . . . 19
- B.2 Environment Transport . . . 19
- B.3 Flow Refinement . . . 19
- B.4 Matched Transfer Design . . . 20

C Implementation Details **21**
- C.1 Datasets and Evaluation . . . 21
- C.2 Training and Inference . . . 22

D Additional Ablation Studies **23**

E Response Analysis **24**
- E.1 Incomplete Future Environments . . . 24
- E.2 Reference Environment Sensitivity . . . 24
- E.3 Environment Interventions . . . 25
- E.4 Internal Response Diagnostics . . . 25

F Efficiency Analysis **26**

G Sensitivity Analysis **27**
- G.1 Hyperparameter Sensitivity . . . 27
- G.2 Refinement-Step Sensitivity . . . 28
- G.3 Length Sensitivity . . . 28

H Additional Results and Visualizations **29**

I Limitations and Future Work **29**

A Mathematical Properties

We give a population interpretation of Environment Response and derive the error identity used by the target-aligned analysis in Section 4.3.

A.1 Population Interpretation of Environment Response

Proposition. Let $\ell(\mathbf{y}, \mathbf{a})$ be a prediction loss for which measurable conditional Bayes acts exist. Let

$$\begin{aligned} \mathbf{a}^\star_X &\in \arg\min_{\mathbf{a}} \mathbb{E}[\ell(\mathbf{Y}, \mathbf{a}) \mid \mathbf{X}], \\ \mathbf{a}^\star_{X,E} &\in \arg\min_{\mathbf{a}} \mathbb{E}[\ell(\mathbf{Y}, \mathbf{a}) \mid \mathbf{X}, \mathbf{E}] \end{aligned} \tag{13}$$

be measurable selections from the corresponding sets of conditional Bayes acts. Define the associated population Environment Response as

$$\Delta^\star_\ell(\mathbf{X}, \mathbf{E}) = \mathbf{a}^\star_{X,E} - \mathbf{a}^\star_X. \tag{14}$$

Then

$$\mathbf{a}^\star_{X,E} = \mathbf{a}^\star_X + \Delta^\star_\ell(\mathbf{X}, \mathbf{E}), \tag{15}$$

and

$$\mathbb{E}[\ell(\mathbf{Y}, \mathbf{a}^\star_{X,E})] \le \mathbb{E}[\ell(\mathbf{Y}, \mathbf{a}^\star_X)]. \tag{16}$$

Proof. Equation 15 follows directly from Equation 14. Since $\mathbf{a}^\star_X$ is also $(\mathbf{X}, \mathbf{E})$-measurable,

$$\mathbb{E}[\ell(\mathbf{Y}, \mathbf{a}^\star_{X,E}) \mid \mathbf{X}, \mathbf{E}] \le \mathbb{E}[\ell(\mathbf{Y}, \mathbf{a}^\star_X) \mid \mathbf{X}, \mathbf{E}].$$

Taking expectations gives Equation 16. □

For the horizon-wise MAE used by EGF,

$$\ell_{\mathrm{MAE}}(\mathbf{y}, \mathbf{a}) = H^{-1}\|\mathbf{y} - \mathbf{a}\|_1, \tag{17}$$

the conditional Bayes acts are coordinate-wise conditional medians. If a median is not unique, any measurable selection in Equation 13 gives a valid associated response.

Under squared loss, $\ell_2(\mathbf{y}, \mathbf{a}) = \|\mathbf{y} - \mathbf{a}\|_2^2$, the Bayes acts are conditional means. Hence

$$\begin{aligned} \Delta^\star_\ell(\mathbf{X}, \mathbf{E}) &= \mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{E}] - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] \\ &= \Delta^\star(\mathbf{X}, \mathbf{E}), \end{aligned} \tag{18}$$

recovering Equation 1. In this case,

$$\mathbb{E}[\Delta^\star(\mathbf{X}, \mathbf{E}) \mid \mathbf{X}] = \mathbf{0}, \tag{19}$$

and

$$\begin{aligned} &\mathbb{E}\,\|\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}]\|_2^2 - \mathbb{E}\,\|\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{E}]\|_2^2 \\ &\quad = \mathbb{E}\,\|\Delta^\star(\mathbf{X}, \mathbf{E})\|_2^2 . \end{aligned} \tag{20}$$

Proof. Equation 19 follows from

$$\begin{aligned} \mathbb{E}[\Delta^\star \mid \mathbf{X}] &= \mathbb{E}[\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{E}] \mid \mathbf{X}] - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] \\ &= \mathbf{0}. \end{aligned}$$

Let

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{E}].$$

Since $\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{X}, \mathbf{E}] = \mathbf{0}$,

$$\mathbb{E}\langle \boldsymbol{\varepsilon}, \Delta^\star \rangle = 0.$$

Using

$$\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \boldsymbol{\varepsilon} + \Delta^\star$$

and expanding the squared norm gives Equation 20. □

A.2 Reference Environments

EGF uses two distinct references. The learned forecast reference $\mathbf{P}_0(\mathbf{X})$ anchors the final prediction, while $\mathbf{E}_0$ anchors the contrast used by Environment Transport:

$$\mathbf{r}_E = \phi_E(\mathbf{E}) - \phi_E(\mathbf{E}_0), \qquad \mathbf{r}_E(\mathbf{E}_0) = \mathbf{0}. \tag{21}$$

The equality $\mathbf{r}_E(\mathbf{E}_0) = \mathbf{0}$ concerns the environment representation and does not require $\Delta_\theta(\mathbf{X}, \mathbf{E}_0) = \mathbf{0}$.

For two environments $\mathbf{E}$ and $\mathbf{E}'$ under the same history, Equation 2 gives

$$\widehat{\mathbf{Y}}(\mathbf{X}, \mathbf{E}) - \widehat{\mathbf{Y}}(\mathbf{X}, \mathbf{E}') = \Delta_\theta(\mathbf{X}, \mathbf{E}) - \Delta_\theta(\mathbf{X}, \mathbf{E}'). \tag{22}$$

A.3 Target-Aligned Correction Geometry

The following quantities are computed separately for each model and test origin; we omit those two indices for clarity. Let f denote the corresponding model and $\mathbf{E}_p$ the diagnostic persistence environment, and define

$$\begin{aligned} \mathbf{P}_p &= f(\mathbf{X}, \mathbf{E}_p), \\ \mathbf{d}^\star &= \mathbf{Y} - \mathbf{P}_p, \\ \mathbf{d} &= f(\mathbf{X}, \mathbf{E}) - \mathbf{P}_p. \end{aligned} \tag{23}$$

Here $\mathbf{d}^\star$ is the target-required correction and $\mathbf{d}$ is the correction induced by the supplied future environment.

For $\|\mathbf{d}^\star\|_2 > 0$, define

$$\begin{aligned} x &= \frac{\langle \mathbf{d}, \mathbf{d}^\star \rangle}{\|\mathbf{d}^\star\|_2^2}, \\ y &= \frac{\|\mathbf{d} - x\mathbf{d}^\star\|_2}{\|\mathbf{d}^\star\|_2}. \end{aligned} \tag{24}$$

Proposition. The distance from (x, y) to the ideal correction $(1, 0)$ satisfies

$$\sqrt{(x-1)^2 + y^2} = \frac{\|\mathbf{Y} - f(\mathbf{X}, \mathbf{E})\|_2}{\|\mathbf{Y} - \mathbf{P}_p\|_2}. \tag{25}$$

Proof. Let

$$\mathbf{e} = \mathbf{d} - x\mathbf{d}^\star.$$

By the definition of x, $\langle \mathbf{e}, \mathbf{d}^\star \rangle = 0$, while

$$\mathbf{Y} - f(\mathbf{X}, \mathbf{E}) = (1 - x)\mathbf{d}^\star - \mathbf{e}.$$

Therefore,

$$\begin{aligned} \|\mathbf{Y} - f(\mathbf{X}, \mathbf{E})\|_2^2 &= (1-x)^2 \|\mathbf{d}^\star\|_2^2 + \|\mathbf{e}\|_2^2 \\ &= \big((1-x)^2 + y^2\big) \|\mathbf{d}^\star\|_2^2. \end{aligned}$$

Equation 25 follows immediately. □

The identity is independent of the training loss. We compute the coordinates in double precision for model–origin pairs satisfying $\|\mathbf{d}^\star\|_2^2 > 10^{-12}$. Figure 3 reports coordinate-wise medians and median normalized distances.

B Details of EGF

This section details the three stages of EGF: History Belief, Environment Transport, and Flow Refinement. It also describes the matched transfer design used to evaluate Environment Response beyond EGF. Prediction and training procedures are given in Appendix C.2.

B.1 History Belief

Let $\mathbf{Q} \in \mathbb{R}^{B\times H\times D}$ denote the learned Horizon Tokens shown in Figure 2. For each temporal view k, the strided patch encoder produces $\mathbf{T}^{(k)}$, and the View Decoder maps these tokens to the prediction horizon:

$$\left(\mathbf{P}^{(k)}, \mathbf{H}^{(k)}\right) = \mathrm{ViewDecoder}_k\left(\mathbf{Q}, \mathbf{T}^{(k)}\right). \tag{26}$$

Here $\mathbf{P}^{(k)} \in \mathbb{R}^{B\times H}$ is the view forecast and $\mathbf{H}^{(k)} \in \mathbb{R}^{B\times H\times D}$ is its horizon-aligned state. The projected $\mathbf{Q}$ and pooled history form the decoder queries, while $\mathbf{T}^{(k)}$ supplies the keys and values. A feed-forward block updates the attended representation, and the forecast head predicts a residual over the last observed target.

The three temporal views use at most 96, 336, and 720 observations, with patch strides 2, 6, and 24. For future time point $h \in \{0, \ldots, H-1\}$, the j-th phase anchor pair uses positions $L+h-jp$ and $L+h-(j+1)p$. Only pairs with both positions in $[0, L-1]$ enter the phase softmax. Each valid pair combines its projected anchor level and historical change with future-time and anchor-index embeddings. The bounded gain head is initialized at zero.

The four view forecasts and states are stacked internally along the view dimension. We collect the view-level information retained for Environment Transport as

$$\begin{aligned} \mathcal{V} &= \left\{\{\mathbf{P}^{(v)}, \mathbf{H}^{(v)}\}_{v=1}^{4}, \boldsymbol{\alpha}\right\}, \\ \mathbf{M} &= \mathrm{Concat}\left(\{\mathbf{T}^{(k)}\}_{k=1}^{3}, \mathbf{H}^{(4)}\right), \\ H_X &= \{\mathcal{V}, \mathbf{M}, \mathbf{C}, \mathbf{Q}\}. \end{aligned} \tag{27}$$

In the terminology of Figure 2, $\mathcal{V}$, $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{Q}$ denote the retained View Information, History Memory, Coarse State, and Horizon Tokens, respectively, with $\mathbf{C} \in \mathbb{R}^{B\times H\times D}$. Thus, H_X is a compact notation for the retained history context rather than a single tensor consumed in full by every later module. Environment Transport uses $\mathcal{V}$. The refinement vector field uses $\mathbf{C}$ and $\mathbf{Q}$ for input conditioning and $\mathbf{M}$ for History Cross-Attention. The correction gate uses $\mathbf{C}$ and the cross-view disagreement derived from $\mathcal{V}$.

B.2 Environment Transport

Environment Transport reads $\mathcal{V} = \{\{\mathbf{P}^{(v)}, \mathbf{H}^{(v)}\}_{v=1}^{4}, \boldsymbol{\alpha}\}$. The shared environment encoder contains an input projection, horizon-position embeddings, residual depthwise temporal convolutions, and feed-forward blocks. The supplied and reference environments share all encoder parameters.

The cross-view disagreement at each future time point is

$$\mathbf{s} = \mathrm{Std}_v\left(\mathbf{P}^{(v)}\right). \tag{28}$$

Given the routing change $\Delta\mathbf{w}$, the forecast- and state-level shifts are

$$\mathbf{d}_P = \sum_v \Delta\mathbf{w}^{(v)} \odot \mathbf{P}^{(v)}, \qquad \mathbf{d}_H = \sum_v \Delta\mathbf{w}^{(v)} \odot \mathbf{H}^{(v)}. \tag{29}$$

The Confidence Head and Adapter Head receive $\mathbf{r}_E$, $\mathbf{d}_H$, $\mathbf{d}_P$, and $\mathbf{s}$. Both use input width $2D+2$ and hidden width D. The Confidence Head outputs one sigmoid weight per future time point. The Adapter Head outputs a D-dimensional update, with its final layer initialized to zero. Their outputs form $\widetilde{\mathbf{R}}_E$ in Equation 8.

B.3 Flow Refinement

The vector field first embeds $(\mathbf{z}_t, \mathbf{P}_0)$ at each future time point. Projected $\mathbf{C}$, $\widetilde{\mathbf{R}}_E$, $\mathbf{Q}$, and the flow-time embedding are added to this representation.

The result passes through a configuration-dependent stack of Refinement Blocks. Each block applies Horizon Self-Attention, History Cross-Attention over $\mathbf{M}$, and Environment Cross-Attention over $\widetilde{\mathbf{R}}_E$. The transported response also produces lead-wise scale and shift terms. A residual depthwise

temporal convolution and feed-forward block complete the refinement layer. Layer normalization and a linear Velocity Head produce the velocity.

Prediction starts from $\mathbf{z}^{(0)} = \mathbf{P}_0$. With $\Delta t = 1/S$ and $t_{n+\frac{1}{2}} = (n + \frac{1}{2})\Delta t$, the fixed-step update is

$$\mathbf{z}^{(n+1)} = \mathbf{z}^{(n)} + \Delta t\, v_\theta \left(\mathbf{z}^{(n)}, t_{n+\frac{1}{2}} \mid \mathbf{P}_0, \mathbf{M}, \mathbf{C}, \mathbf{Q}, \widetilde{\mathbf{R}}_E\right), \qquad n = 0, \ldots, S-1. \tag{30}$$

Each step requires one velocity evaluation. After S steps, the terminal state in the continuous-time notation of Section 3.4 is $\mathbf{z}_1 \equiv \mathbf{z}^{(S)}$.

The Adaptive Correction Gate embeds $(\mathbf{z}_1 - \mathbf{P}_0, \mathbf{P}_0)$ and combines it with $\mathbf{C}$, cross-view disagreement $\mathbf{s}$, $\widetilde{\mathbf{R}}_E$, and the fixed source scale σ_0. Its MLP has shape $(3D + 2) \to D \to 1$, followed by a temperature-scaled sigmoid. The final bias is initialized to -1.

B.4 Matched Transfer Design

We instantiate the matched transfer design on DAG (Qiu et al., 2026), CrossLinear (Zhou et al., 2025), and TimeXer (Wang et al., 2024). Direct and Response use parameter-matched temporal environment adapters with hidden width 32 and matched temporal receptive field. Each adapter contains a residual depthwise temporal convolution with kernel size five, a feed-forward block with expansion ratio two, and a two-layer GELU gate. Direct assigns this pathway to full-target prediction. Response assigns the matched pathway to a reference-relative forecast update. Response additionally uses auxiliary MSE losses on the reference and proposal forecasts; Direct does not use these auxiliary losses. Following the official DAG benchmark implementation, all matched-transfer experiments retain its original single-run protocol with the fixed random seed 2021. Native, Direct, and Response are trained and evaluated under exactly the same seed and backbone-specific settings, preserving direct comparability with the official benchmark protocol.

Reference construction. The reference forecast is constructed according to the native structure of each backbone. DAG already provides a history-only temporal prediction, which we directly use as $\mathbf{P}_r$. CrossLinear and TimeXer do not expose an analogous history-only prediction path; for these backbones, we obtain $\mathbf{P}_r$ from the native forecasting path under a persistence environment that repeats the last observed exogenous value over the forecasting horizon. Thus, the implementations are backbone-specific while sharing the same role: $\mathbf{P}_r$ is supported only by information available from history.

Let $\widehat{\mathbf{Y}}_N$ denote the Native forecast, $\mathbf{P}_r$ the history-supported reference, and s_y the historical target scale. The adapter predicts a normalized proposal $\mathbf{u}$ and gate logits $\mathbf{G}$. The Response forecast is

$$\widehat{\mathbf{Y}}_R = \mathbf{P}_r + 2\ \mathrm{Sigmoid}(\mathbf{G}) \odot \left(\widehat{\mathbf{Y}}_N + s_y\mathbf{u} - \mathbf{P}_r\right). \tag{31}$$

The term $\widehat{\mathbf{Y}}_N + s_y\mathbf{u}$ is the proposal forecast. The proposal and gate output layers are initialized to zero. Thus, $\mathbf{u} = \mathbf{0}$ and $\mathbf{G} = \mathbf{0}$ at initialization, giving $\widehat{\mathbf{Y}}_R = \widehat{\mathbf{Y}}_N$.

Auxiliary-supervision control. Response additionally uses auxiliary MSE losses on the reference and proposal forecasts. To control for this difference, we further evaluate the *Response w/o Aux.* variant, which removes only these two losses while keeping the Response architecture and the matched-transfer protocol unchanged. The results are reported in Table 4.

Without the Response-specific auxiliary losses, Response w/o Aux. still lowers average MSE relative to Direct on all three backbones, with MSE wins on 8/12, 8/12, and 11/12 datasets for DAG, CrossLinear, and TimeXer, respectively. This shows that the improvement of the reference-relative formulation is not solely explained by the additional auxiliary supervision.

Table 4: Direct–Response w/o Aux.–Response matched transfers across 12 datasets. Entries average two horizons per dataset. Red and blue mark the best and second-best values within each backbone, respectively. Ties are determined at the reported precision, and first-place counts include ties. Lower is better.

Models	DAG (2026)						CrossLinear (2025)						TimeXer (2024)					
Conditions	D		R w/o Aux.		R		D		R w/o Aux.		R		D		R w/o Aux.		R	
Datasets	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	0.333	0.332	0.314	**0.318**	**0.312**	0.323	0.348	0.372	0.328	0.349	**0.311**	**0.336**	0.343	0.336	**0.323**	**0.326**	0.331	0.335
PJM	**0.092**	**0.177**	0.095	0.178	0.095	0.187	0.119	0.212	**0.099**	**0.200**	0.102	0.201	0.101	0.192	**0.099**	**0.190**	0.106	0.201
BE	0.419	**0.273**	0.420	0.292	**0.416**	0.286	0.452	0.328	0.454	0.331	**0.440**	**0.311**	0.429	0.278	**0.400**	**0.269**	0.404	0.277
FR	0.413	**0.218**	0.420	0.223	**0.399**	**0.218**	0.449	0.294	0.427	**0.246**	**0.425**	0.250	**0.422**	**0.236**	0.440	0.237	0.434	0.242
DE	0.369	0.366	0.339	**0.353**	**0.335**	**0.353**	0.424	0.417	**0.412**	0.411	0.420	**0.402**	0.416	0.389	**0.375**	**0.378**	0.411	0.396
Energy	0.111	0.256	0.097	0.241	**0.096**	**0.239**	0.295	0.444	0.159	0.319	**0.149**	**0.306**	0.122	0.274	0.115	0.263	**0.112**	**0.260**
Sdwpfm1	0.431	0.464	**0.422**	**0.457**	0.445	0.480	**0.375**	**0.475**	0.388	0.478	0.434	0.500	0.481	0.511	**0.480**	**0.491**	0.487	0.501
Sdwpfm2	0.466	0.482	0.449	**0.468**	**0.447**	0.491	0.513	0.560	**0.500**	**0.555**	0.502	0.557	0.552	0.546	**0.531**	**0.524**	0.534	0.528
Sdwpfh1	**0.411**	**0.450**	0.438	0.459	0.434	0.475	**0.461**	**0.537**	0.478	0.548	0.483	0.552	0.526	0.556	**0.486**	**0.511**	**0.486**	0.525
Sdwpfh2	0.464	**0.473**	0.452	0.475	**0.438**	0.491	0.492	0.545	**0.478**	**0.543**	0.499	0.559	0.604	0.598	0.571	**0.554**	**0.568**	0.568
Colbun	0.094	0.153	0.086	0.145	**0.082**	**0.142**	0.154	0.212	0.204	0.216	**0.123**	**0.195**	0.111	0.197	**0.082**	**0.148**	0.085	0.152
Rapel	0.294	0.347	**0.266**	**0.314**	0.288	0.331	0.242	0.309	0.230	0.290	**0.211**	**0.282**	0.308	0.338	0.303	0.329	**0.292**	**0.320**
1st Count	2	**5**	2	**5**	**8**	4	2	2	4	4	**6**	**6**	1	1	**8**	**9**	4	2

C Implementation Details

C.1 Datasets and Evaluation

We evaluate EGF on 12 real-world TSF-X datasets whose future exogenous variables are available over the forecasting horizon. The datasets cover electricity-price forecasting, power generation, hydropower, and wind-power forecasting.

Electricity-price datasets. NP, PJM, BE, FR, and DE come from the EPF benchmark (Lago et al., 2021; Wang et al., 2024). **NP** records the Nord Pool electricity market at hourly resolution. The electricity price is the endogenous target, while grid-load and wind-power forecasts provide the two future exogenous variables. **PJM** records the Pennsylvania–New Jersey–Maryland Interconnection market. We forecast the zonal electricity price in the Commonwealth Edison (COMED) area using system-load and COMED zonal-load forecasts as exogenous variables. **BE** contains the hourly Belgian electricity price as the target, with system-load and generation forecasts as future exogenous inputs. **FR** uses the French electricity price as the target and corresponding generation and load forecasts as exogenous variables. **DE** records the German electricity market, where the hourly electricity price is forecast from wind-power and Amprion zonal-load forecasts.

Energy and reservoir datasets. The remaining datasets follow the TSF-X collection introduced by DAG (Qiu et al., 2026). **Energy** contains hourly power-generation measurements from Chile. Thermoelectric generation is the endogenous target, while battery storage, geothermal, hydroelectric, solar, and wind generation constitute the five exogenous variables. **Colbun** and **Rapel** are daily-resolution hydropower datasets for the corresponding reservoirs in Chile. Water level is the forecasting target, and precipitation and tributary inflow provide the two exogenous variables.

Wind-power datasets. **Sdwpfh1**, **Sdwpfh2**, **Sdwpfm1**, and **Sdwpfm2** are wind-power datasets derived from two turbines in the Longyuan wind farm. Sdwpfh1 and Sdwpfh2 are sampled hourly, whereas Sdwpfm1 and Sdwpfm2 use 30-minute sampling. The endogenous target is turbine active power (Patv). Each dataset contains six ERA5 meteorological covariates (Hersbach et al., 2020): temperature, surface pressure, relative humidity, wind speed, wind direction, and total precipitation.

Table 5: Statistics of the 12 TSF-X datasets. #Exo denotes the number of future exogenous variables; each dataset has one endogenous target. Sdwpfh and Sdwpfm each contain two turbine datasets (h1/h2 and m1/m2), with the listed length applying to each turbine.

Dataset	#Exo	Frequency	Length	Split	Exogenous variables	Target
NP	2	1 hour	52,416	7:1:2	Grid load, wind power	Electricity price
PJM	2	1 hour	52,416	7:1:2	System load, COMED zonal load	Electricity price
BE	2	1 hour	52,416	7:1:2	Generation, system load	Electricity price
FR	2	1 hour	52,416	7:1:2	Generation, system load	Electricity price
DE	2	1 hour	52,416	7:1:2	Wind power, zonal load	Electricity price
Energy	5	1 hour	13,064	7:1:2	Battery, geothermal, hydro, solar, wind	Thermoelectric generation
Colbun	2	1 day	2,958	7:1:2	Precipitation, tributary inflow	Water level
Rapel	2	1 day	3,366	7:1:2	Precipitation, tributary inflow	Water level
Sdwpfh	6	1 hour	14,641	7:1:2	ERA5 climate features	Active power
Sdwpfm	6	30 minutes	29,281	7:1:2	ERA5 climate features	Active power

Forecasting protocol. Following DAG (Qiu et al., 2026), all datasets are split chronologically into training, validation, and test sets with a 7:1:2 ratio. NP, PJM, BE, FR, DE, Energy, and the four wind-power datasets use $(L, H) = (168, 24)$ for short-term forecasting and $(720, 360)$ for long-term forecasting. For both Colbun and Rapel, we use $(L, H) = (60, 10)$ and $(180, 30)$ for the short- and long-horizon settings, respectively. For EGF, training origins use strides 4 and 8 for the short- and long-horizon settings, while validation and test origins use stride one. Colbun and Rapel use stride one throughout.

Preprocessing and evaluation. All channels are first standardized using statistics fitted on the training partition. At each forecasting origin, EGF computes channel-wise mean and scale statistics from the configured recent-history window and applies the same transformation to the historical input, future environment, and target. In these normalized coordinates, the reference environment is $\mathbf{E}_0 = \mathbf{0}$. Predictions are inverse-transformed from the origin-relative normalization before MSE and MAE are evaluated on the dataset-standardized scale.

DAG (Qiu et al., 2026), CrossLinear (Zhou et al., 2025), and TimeXer (Wang et al., 2024) are reproduced using the official DAG code and settings. Following DAG, incomplete final test batches are retained (`drop_last=False`). KITE (Cheng et al., 2026) and the non-reproduced common baselines use paper-reported results, while GCGNet (Li et al., 2026) uses its reported results. Complete horizon-level results are provided in Appendix H.

C.2 Training and Inference

Algorithms 1 and 2 summarize the training and deterministic prediction procedures. Both follow the three-stage forward path in Figure 2; training additionally constructs the stochastic path used by the flow-matching objective in Section 3.5.

Optimization and hyperparameters. We train all EGF modules jointly using AdamW (Loshchilov & Hutter, 2019). The maximum number of training epochs is 12, with an early-stopping patience of 4. The checkpoint with the lowest validation $\frac{1}{2}(\mathrm{MSE} + \mathrm{MAE})$ is used for evaluation, and gradient-norm clipping is applied during training.

Across dataset–horizon configurations, $D \in \{64, 96, 128\}$ and the batch size is chosen from $\{32, 64, 128\}$. The initial learning rates typically range from 1.0×10^{-4} to 7.5×10^{-4}, with a few configuration-specific exceptions. Appendix G further studies the hidden width D, source scale σ_0, transport scale s_{tr}, correction-gate temperature, and the number of integration steps S.

Software and hardware. EGF is implemented in PyTorch (Paszke et al., 2019) with Python 3.10. The released environment uses PyTorch 2.12.1 and CUDA 13.0. Random generators are fixed and deterministic PyTorch execution is enabled for the formal EGF experiments. The efficiency measurements in Appendix F are performed on an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU.

Algorithm 1 EGF training

Require: Training minibatch $(\mathbf{X}, \mathbf{E}, \mathbf{Y})$, integration steps S
Ensure: Updated parameters θ
Normalize $(\mathbf{X}, \mathbf{E}, \mathbf{Y})$ using history-based channel statistics
Set the normalized reference environment $\mathbf{E}_0 \leftarrow \mathbf{0}$
$(\mathbf{P}_0, \mathcal{V}, \mathbf{M}, \mathbf{C}, \mathbf{Q}) \leftarrow \textsc{HistoryBelief}(\mathbf{X})$
$\mathbf{r}_E \leftarrow \phi_E(\mathbf{E}) - \phi_E(\mathbf{E}_0)$
$\widetilde{\mathbf{R}}_E \leftarrow \textsc{EnvironmentTransport}(\mathbf{r}_E, \mathcal{V})$
$\mathbf{z}^{(0)} \leftarrow \mathbf{P}_0$
for $n = 0, \dots, S-1$ **do**
    $t_{n+\frac{1}{2}} \leftarrow (n + \frac{1}{2})/S$
    $\mathbf{z}^{(n+1)} \leftarrow \mathbf{z}^{(n)} + S^{-1} v_\theta(\mathbf{z}^{(n)}, t_{n+\frac{1}{2}} \mid \mathbf{P}_0, \mathbf{M}, \mathbf{C}, \mathbf{Q}, \widetilde{\mathbf{R}}_E)$
end for
$\mathbf{z}_1 \leftarrow \mathbf{z}^{(S)}$
$\mathbf{s} \leftarrow \mathrm{Std}_v(\mathbf{P}^{(v)})$
$\mathbf{g} \leftarrow \textsc{CorrectionGate}(\mathbf{z}_1 - \mathbf{P}_0, \mathbf{P}_0, \mathbf{C}, \widetilde{\mathbf{R}}_E, \mathbf{s}, \sigma_0)$
$\widehat{\mathbf{Y}} \leftarrow \mathbf{P}_0 + \mathbf{g} \odot (\mathbf{z}_1 - \mathbf{P}_0)$
$\mathcal{L}_{\text{pred}} \leftarrow \mathrm{MAE}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_b \,\mathrm{MAE}(\mathbf{P}_0, \mathbf{Y})$
Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}(0, 1)$
$\widetilde{\mathbf{z}}_0 \leftarrow \mathbf{P}_0 + \sigma_0 \boldsymbol{\epsilon}$
$\widetilde{\mathbf{z}}_t \leftarrow (1-t)\widetilde{\mathbf{z}}_0 + t\mathbf{Y}$
$\mathcal{L}_{\text{flow}} \leftarrow \mathrm{MSE}\Big(v_\theta(\widetilde{\mathbf{z}}_t, t \mid \mathbf{P}_0, \mathbf{M}, \mathbf{C}, \mathbf{Q}, \widetilde{\mathbf{R}}_E), \mathbf{Y} - \widetilde{\mathbf{z}}_0\Big)$
Update θ using $\mathcal{L}_{\text{pred}} + \lambda_f \mathcal{L}_{\text{flow}}$

Algorithm 2 EGF prediction

Require: Historical input $\mathbf{X}$, future environment $\mathbf{E}$, integration steps S
Ensure: Forecast $\widehat{\mathbf{Y}}$
Normalize $(\mathbf{X}, \mathbf{E})$ using history-based channel statistics
Set $\mathbf{E}_0 \leftarrow \mathbf{0}$
$(\mathbf{P}_0, \mathcal{V}, \mathbf{M}, \mathbf{C}, \mathbf{Q}) \leftarrow \textsc{HistoryBelief}(\mathbf{X})$
$\mathbf{r}_E \leftarrow \phi_E(\mathbf{E}) - \phi_E(\mathbf{E}_0)$
$\widetilde{\mathbf{R}}_E \leftarrow \textsc{EnvironmentTransport}(\mathbf{r}_E, \mathcal{V})$
$\mathbf{z}^{(0)} \leftarrow \mathbf{P}_0$
for $n = 0, \dots, S-1$ **do**
    $t_{n+\frac{1}{2}} \leftarrow (n + \frac{1}{2})/S$
    $\mathbf{z}^{(n+1)} \leftarrow \mathbf{z}^{(n)} + S^{-1} v_\theta(\mathbf{z}^{(n)}, t_{n+\frac{1}{2}} \mid \mathbf{P}_0, \mathbf{M}, \mathbf{C}, \mathbf{Q}, \widetilde{\mathbf{R}}_E)$
end for
$\mathbf{z}_1 \leftarrow \mathbf{z}^{(S)}$
$\mathbf{s} \leftarrow \mathrm{Std}_v(\mathbf{P}^{(v)})$
$\mathbf{g} \leftarrow \textsc{CorrectionGate}(\mathbf{z}_1 - \mathbf{P}_0, \mathbf{P}_0, \mathbf{C}, \widetilde{\mathbf{R}}_E, \mathbf{s}, \sigma_0)$
$\widehat{\mathbf{Y}} \leftarrow \mathbf{P}_0 + \mathbf{g} \odot (\mathbf{z}_1 - \mathbf{P}_0)$
return inverse-normalized $\widehat{\mathbf{Y}}$

D Additional Ablation Studies

The stage-level ablations in Table 3 replace complete EGF stages. We next isolate the main design choices within each stage on the same six datasets.

History Belief. w/o Phase View removes the phase-aligned historical view while retaining the three temporal views and their adaptive fusion. **w/o History-Belief Supervision** keeps the architecture unchanged but sets $\lambda_b = 0$, removing the auxiliary $\mathrm{MAE}(\mathbf{P}_0, \mathbf{Y})$ loss.

Environment Transport. Absolute Environment Encoding removes the reference contrast, replacing $\phi_E(\mathbf{E}) - \phi_E(\mathbf{E}_0)$ with $\phi_E(\mathbf{E})$. **w/o KL Trust** removes the KL-based attenuation applied to environment-driven redistribution of the historical hypotheses. **w/o Confidence** removes confidence modulation from the transported environment update.

Table 6: Fine-grained ablations on the six datasets in Table 3. Each entry first averages the two horizons within each seed and then averages over five seeds. EGF gain is the relative error reduction over each variant.

Model	NP		PJM		FR		DE		Energy		Sdwpfh1		Average		EGF gain	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
EGF (Ours)	0.299	0.307	**0.084**	**0.174**	**0.378**	**0.206**	**0.318**	0.342	0.088	0.226	**0.365**	**0.427**	**0.255**	**0.280**	–	–
Absolute Environment Encoding	0.310	0.312	0.086	0.175	0.380	0.207	0.324	0.344	0.093	0.232	0.382	0.433	0.262	0.284	2.7%	1.4%
w/o Phase View	0.336	0.323	0.089	0.179	0.425	0.223	0.323	0.352	0.115	0.265	0.387	0.441	0.279	0.297	8.6%	5.7%
w/o KL Trust	0.299	0.309	0.087	0.175	0.382	0.209	0.321	0.342	**0.086**	**0.224**	0.378	0.433	0.259	0.282	1.5%	0.7%
w/o Confidence	0.299	0.309	0.086	0.175	0.429	0.208	0.323	0.341	0.090	0.229	0.376	0.432	0.267	0.282	4.5%	0.7%
w/o Correction Gate	**0.296**	0.307	0.087	0.175	0.385	0.207	0.322	0.343	0.091	0.230	0.396	0.449	0.263	0.285	3.0%	1.8%
w/o Flow-Matching Loss	0.302	**0.305**	0.088	0.175	0.379	0.210	0.320	0.341	0.094	0.235	0.398	0.432	0.263	0.283	3.0%	1.1%
w/o History-Belief Supervision	0.303	0.311	0.089	0.177	0.385	0.209	0.319	**0.340**	0.087	0.225	0.368	0.431	0.258	0.282	1.2%	0.7%

Flow Refinement. w/o Correction Gate applies the refinement displacement directly without the learned terminal gate. **w/o Flow-Matching Loss** keeps the deterministic refinement path but sets $\lambda_f = 0$, removing $\mathcal{L}_{\text{flow}}$ during training.

In History Belief (Table 6), removing the phase view raises the average MSE/MAE from $0.255/0.280$ to $0.279/0.297$, giving EGF gains of $8.6\%/5.7\%$. In contrast, removing the auxiliary supervision on $\mathbf{P}_0$ changes the average only to $0.258/0.282$. The main benefit of History Belief therefore comes from the phase-aligned representation rather than from its auxiliary loss alone.

Within Environment Transport, Absolute Environment Encoding increases the average error to $0.262/0.284$ and is worse on both metrics across all six datasets. Since both variants observe the same future covariates, this comparison isolates how the environment is represented: encoding its change from $\mathbf{E}_0$ is more effective than encoding the future trajectory in isolation. Removing confidence raises average MSE to 0.267, mostly through FR ($0.378 \rightarrow 0.429$), while removing KL trust gives 0.259. Confidence therefore has the larger aggregate effect on the transported update, whereas KL trust provides a lighter constraint on environment-driven view redistribution.

For Flow Refinement, removing the correction gate gives $0.263/0.285$, showing that the proposed displacement benefits from being adjusted before it is added to $\mathbf{P}_0$. Removing flow-matching supervision gives $0.263/0.283$. Its effect is larger on MSE than MAE (3.0% versus 1.1%), which is consistent with the MSE velocity objective used to train the vector field.

E Response Analysis

We analyze Environment Response from four perspectives: incomplete future inputs, alternative environment references, test-time environment interventions, and the internal response process.

E.1 Incomplete Future Environments

We evaluate EGF when only part of the future environment is available. On NP, PJM, BE, and FR, we randomly mask 10%, 30%, or 50% of the future-covariate entries and replace each missing entry with the last observed value of the corresponding exogenous channel. EGF is retrained for each missing rate, while the 0% condition uses the complete future environment.

Figure 5 traces a gradual increase in forecasting error as the missing rate grows. At 50% missingness, the four-dataset average MSE and MAE increase by 7.8% and 5.9%, respectively, relative to the complete-environment setting.

E.2 Reference Environment Sensitivity

EGF uses $\mathbf{E}_0 = \mathbf{0}$ as the default reference after origin-relative normalization. Each exogenous channel is centered using recent-history statistics, so the zero reference corresponds in the original scale to the recent-history local mean.

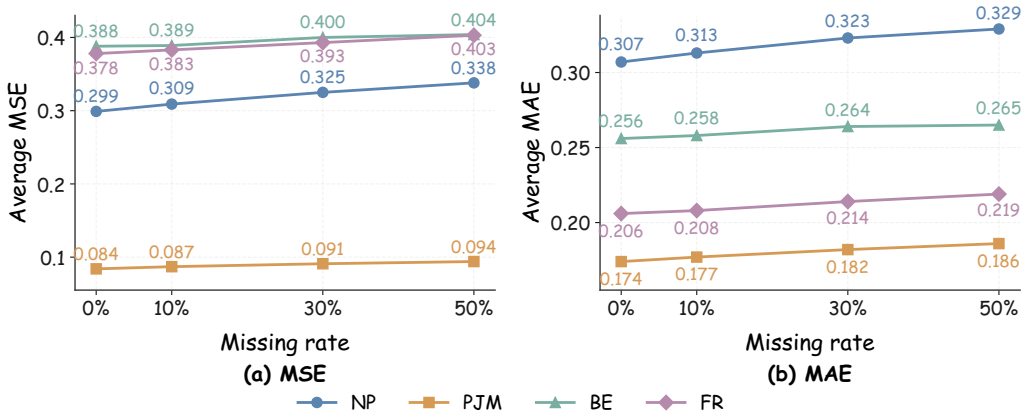


Figure 5: Forecasting with incomplete future environments on NP, PJM, BE, and FR. We report MSE and MAE under 0%, 10%, 30%, and 50% missing future covariates, averaged over two forecasting horizons and five seeds. Lower is better.

We compare it with a *persistence reference*, which repeats the last observed exogenous value over the prediction horizon. EGF is retrained under this reference with all other settings unchanged, while the supplied future environment $\mathbf{E}$ remains the same.

The zero reference achieves lower MSE than persistence in all eight settings and lower MAE in seven (Table 7), with the largest gaps on Sdwpfh1 and on Colbun at $H = 30$; the exception is Colbun MAE at $H = 10$ (0.084 vs. 0.083). The main experiments therefore use $\mathbf{E}_0 = \mathbf{0}$.

E.3 Environment Interventions

We modify the supplied future environment at inference while keeping the trained checkpoint fixed.

Three conditions are considered: *Actual* uses the supplied future environment; *Shift* $+1$ shifts the future trajectory by one time step; and *Persistence* repeats the last observed exogenous value over the full prediction horizon.

Table 7: Reference environment sensitivity across eight dataset–horizon settings.

Dataset	H	Zero		Persistence	
		MSE	MAE	MSE	MAE
NP	24	**0.172**	**0.213**	0.176	0.214
NP	360	**0.426**	**0.401**	0.446	0.407
Energy	24	**0.062**	**0.190**	0.066	0.195
Energy	360	**0.114**	**0.262**	0.115	0.264
Sdwpfh1	24	**0.352**	**0.412**	0.439	0.470
Sdwpfh1	360	**0.377**	**0.441**	0.415	0.460
Colbun	10	**0.050**	0.084	0.053	**0.083**
Colbun	30	**0.111**	**0.150**	0.162	0.182

In Figure 6, Persistence increases both MSE and MAE on all four datasets, with larger changes on NP, Energy, and Sdwpfh1. Shift $+1$ produces smaller changes: both metrics increase on NP, Energy, and Sdwpfh1, while Colbun remains unchanged at the reported precision. Persistence changes the future environment values, whereas Shift $+1$ mainly changes their temporal alignment with the prediction horizon. EGF responds to both the supplied values and their temporal positions.

E.4 Internal Response Diagnostics

The previous ablations measure the effect of individual components on forecasting accuracy. Here, we inspect routing, refinement, and terminal gating within the same prediction path. Starting from the history-based forecast $\mathbf{P}_0$, Environment Transport adjusts the allocation among historical hypotheses, Flow Refinement proposes a forecast-space correction, and the terminal gate controls the magnitude of the applied correction. We report the horizon-averaged routing change, refinement proposal magnitude, and applied correction magnitude. Confidence records the confidence modulation in Environment Transport, while Gate records the terminal correction gate.

For each dataset, we fix one seed-2024 evaluation setting: $H = 360$ for NP, Energy, and Sdwpfh1, and $H = 30$ for Colbun. Within the corresponding evaluation, we compute the MSE of every

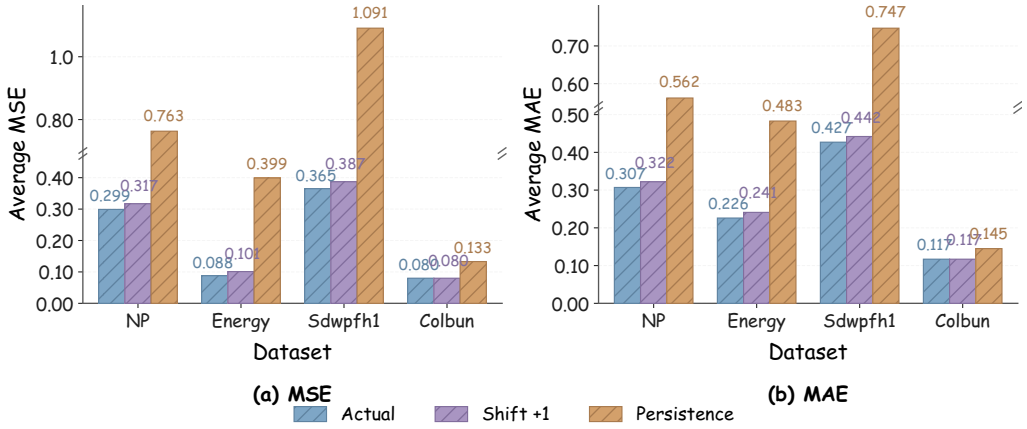


Figure 6: Test-time environment interventions on NP, Energy, Sdwpfh1, and Colbun. We compare the supplied future environment, a one-step temporal shift, and persistence using fixed trained checkpoints. Results are averaged over two forecasting horizons and five seeds. Lower is better.

test origin, sort the origins by per-origin MSE, and select the origin at the lower median, with ties resolved by origin index. Table 8 reports the internal quantities at the selected origin. Gate, Confidence, and the three response magnitudes are averaged over the corresponding prediction horizon, and Case MSE reports the error at that origin.

Table 8: Internal response diagnostics for selected test cases from the seed-2024 evaluations. Proposal and Applied are the horizon-averaged absolute magnitudes of $\mathbf{z}_1 - \mathbf{P}_0$ and $\widehat{\mathbf{Y}} - \mathbf{P}_0$, respectively; Routing averages $\|\Delta \mathbf{w}\|_1$ over the horizon. Gate and Confidence are also horizon averages. Case MSE reports the error at the selected test origin.

Dataset	H	Gate	Confidence	Proposal	Applied	Routing	Case MSE
NP	360	0.917	0.751	0.828	0.806	0.202	0.232
Energy	360	0.951	0.749	0.236	0.227	0.015	0.112
Sdwpfh1	360	0.418	0.407	0.399	0.262	0.036	0.360
Colbun	30	0.555	0.734	0.079	0.045	0.210	0.009

Routing versus correction. Energy shows little routing change (0.015) but retains most of its refinement proposal (0.236 → 0.227), whereas Colbun has the largest routing change (0.210) and only a small final correction (0.045). Routing change therefore has no simple correspondence with forecast-space correction. **Refinement versus gating.** NP combines the largest proposal (0.828) with a high Gate value of 0.917, giving a final correction of 0.806. Sdwpfh1 proposes a correction of 0.399, while its lower Gate value of 0.418 reduces the applied correction to 0.262.

F Efficiency Analysis

We evaluate computational efficiency on Energy with $L = 720$ and $H = 360$. We compare EGF with KITE (Cheng et al., 2026), DAG (Qiu et al., 2026), GCGNet (Li et al., 2026), CrossLinear (Zhou et al., 2025), and TimeXer (Wang et al., 2024). All models use their official implementations and recommended inference settings. Profiling is performed in FP32 on an NVIDIA RTX PRO 6000 Blackwell Server Edition. Each model is evaluated in five independent processes with batch size 32. For each run, we perform 50 warm-up iterations, reset the CUDA peak-memory statistics, and then execute 200 synchronized inference iterations. FLOPs are reported per sample. For memory, we report the median peak allocated GPU memory across the five runs. For latency, we first take the median over the 200 iterations within each run and then report the median across the five runs.

Table 9: Computational efficiency on Energy ($L = 720$, $H = 360$). #Para denotes the number of learnable parameters. #FLOPs denotes floating-point operations per sample. Mem denotes peak allocated GPU memory (MiB) during inference. Inf. denotes inference latency per batch (ms). M and G indicate 10^6 and 10^9, respectively. Memory and latency are measured with batch size fixed to 32.

Model	MAE	#Para	#FLOPs	Mem	Inf.	Model	MAE	#Para	#FLOPs	Mem	Inf.
EGF	**0.262**	**4.19 M**	10.20 G	588.51	37.76	DAG	0.321	6.12 M	20.71 M	67.11	6.26
KITE	0.305	6.92 M	74.58 G	1412.75	133.67	+ **Response**	**0.300**	6.14 M	34.60 M	82.47	6.65
DAG	0.321	6.12 M	**20.71 M**	**67.11**	6.26	CrossLinear	0.385	24.49 M	96.93 M	150.66	**0.72**
GCGNet	0.330	17.20 M	322.64 M	319.33	27.67	+ **Response**	**0.336**	24.51 M	111.15 M	154.42	2.82
CrossLinear	0.385	24.49 M	96.93 M	150.66	**0.72**	TimeXer	0.365	5.31 M	39.53 M	71.95	1.28
TimeXer	0.365	5.31 M	39.53 M	71.95	1.28	+ **Response**	**0.303**	5.33 M	92.95 M	82.16	4.17

Table 9 reports the empirical efficiency on Energy with batch size fixed to 32. **Full EGF.** EGF achieves the lowest MAE (0.262) with the fewest learnable parameters (4.19M). It requires 10.20G FLOPs and 37.76 ms per batch, more than the lightweight backbones, mainly because of attention over the history, prediction horizon, and future environment and the iterative Flow Refinement; KITE requires 74.58G FLOPs and 133.67 ms. **Matched transfers.** Adding Response to DAG, CrossLinear, and TimeXer increases the parameter count by about 0.02M while improving MAE from 0.321 to 0.300, from 0.385 to 0.336, and from 0.365 to 0.303, respectively. The additional FLOPs are 13.89M, 14.22M, and 53.42M, with inference latencies of 6.65 ms, 2.82 ms, and 4.17 ms, respectively.

G Sensitivity Analysis

We analyze the sensitivity of EGF to model hyperparameters, refinement steps, prediction horizon, and history length.

G.1 Hyperparameter Sensitivity

We conduct a one-factor-at-a-time sensitivity analysis on NP, PJM, Energy, and Sdwpfh1 with $L = 720$, $H = 360$, and seed 2024 fixed. The tested factors are hidden dimension D, flow-source scale σ_0, environment-transport scale s_{tr}, and correction-gate temperature. For the last three, we scale the corresponding dataset-specific value by $\{0.5, 0.75, 1.0, 1.25, 1.5\}$.

Figure 7 summarizes the effects of the four hyperparameters. **Model width.** NP and Energy attain their lowest MAE at $D = 160$, while PJM and Sdwpfh1 favor smaller hidden dimensions. **Relative scales.** The effect of s_{tr} is small on NP, Energy, and Sdwpfh1, while PJM shows a larger variation; σ_0 has a more visible effect on NP and Energy. **Gate temperature.** The largest variation appears on Sdwpfh1. NP and PJM change less across the tested range, while Energy reaches lower error at larger values. EGF is less sensitive to s_{tr} than to model width and gate temperature.

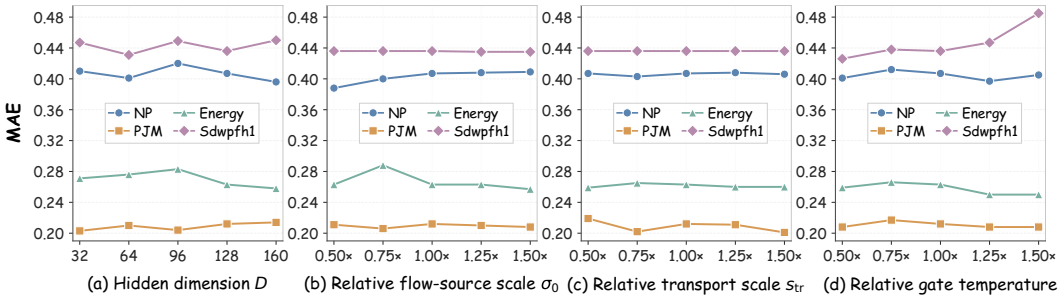


Figure 7: **Hyperparameter sensitivity of EGF.** Panel (a) varies the hidden dimension D. Panels (b)–(d) vary the relative flow-source scale σ_0, environment-transport scale s_{tr}, and correction-gate temperature. Lower MAE is better.

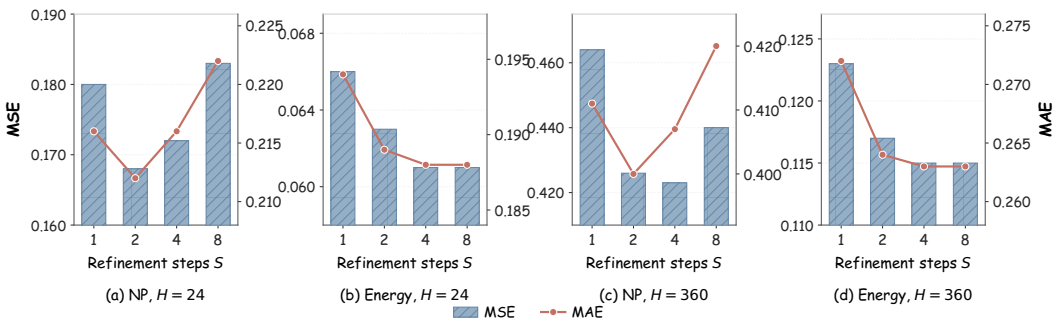


Figure 8: **Refinement-step sensitivity.** NP and Energy are evaluated at $H = 24$ and $H = 360$. Bars report MSE and curves report MAE. Lower is better.

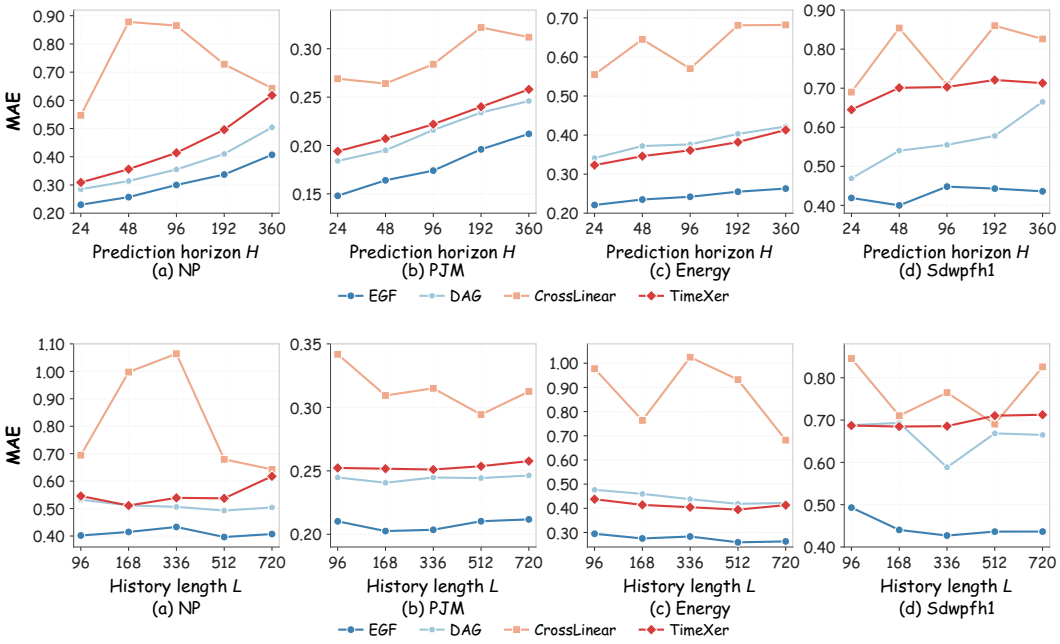


Figure 9: **Length sensitivity.** Top: prediction-horizon analysis with $L = 720$. Bottom: history-length analysis with $H = 360$. MAE is reported for EGF, DAG, CrossLinear, and TimeXer. Lower is better.

G.2 Refinement-Step Sensitivity

We vary the number of velocity evaluations $S \in \{1, 2, 4, 8\}$ at inference while keeping each trained checkpoint fixed. Experiments are conducted on NP and Energy at $H = 24$ and $H = 360$.

On NP (Figure 8), increasing S from one to two reduces both MSE and MAE at the two prediction horizons, while $S = 8$ produces higher error. On Energy, both metrics improve up to four steps and change little afterward. Across the four settings, the best results are concentrated around two to four velocity evaluations.

G.3 Length Sensitivity

We evaluate two length-related factors. For prediction-horizon sensitivity, we fix $L = 720$ and vary $H \in \{24, 48, 96, 192, 360\}$ on NP, PJM, Energy, and Sdwpfh1 with seed 2024. All methods retain their long-horizon architecture settings, with only the horizon-dependent output shape changed. For

history-length sensitivity, we fix $H = 360$ and vary $L \in \{96, 168, 336, 512, 720\}$ on the same four datasets. Each model is retrained for every history length with seed 2024.

Figure 9 compares the two length sweeps. **Prediction horizon.** EGF achieves the lowest MAE in all 20 dataset–horizon settings. Error generally increases with H on NP, PJM, and Energy, while Sdwpfh1 shows a less monotonic pattern. **History length.** EGF also achieves the lowest MAE in all 20 dataset–history-length settings. The best L is 512 on NP and Energy, 168 on PJM, and 336 on Sdwpfh1. The preferred history length varies across datasets, and longer histories do not consistently improve forecasting accuracy.

H Additional Results and Visualizations

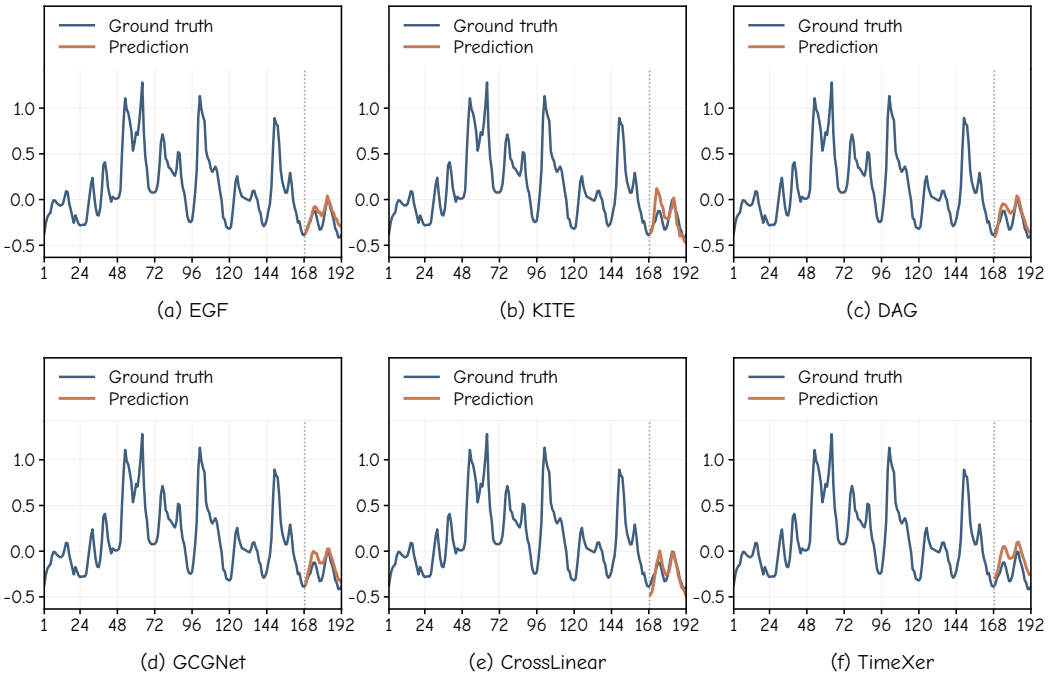
We provide additional forecasting visualizations together with the complete quantitative results. Figures 10 and 11 compare EGF with KITE, DAG, GCGNet, CrossLinear, and TimeXer on four representative datasets. Each visualization shows the predicted trajectory against the ground truth for the corresponding historical input and prediction horizon.

Figure 12 further shows representative Response-transfer cases for DAG, CrossLinear, and TimeXer on NP, PJM, Energy, and Colbun. For each backbone, we compare the original Native forecast with the corresponding +Response forecast and the ground truth. Across these examples, Response adjusts the forecast trajectory toward the observed future while preserving the overall behavior of the underlying backbone.

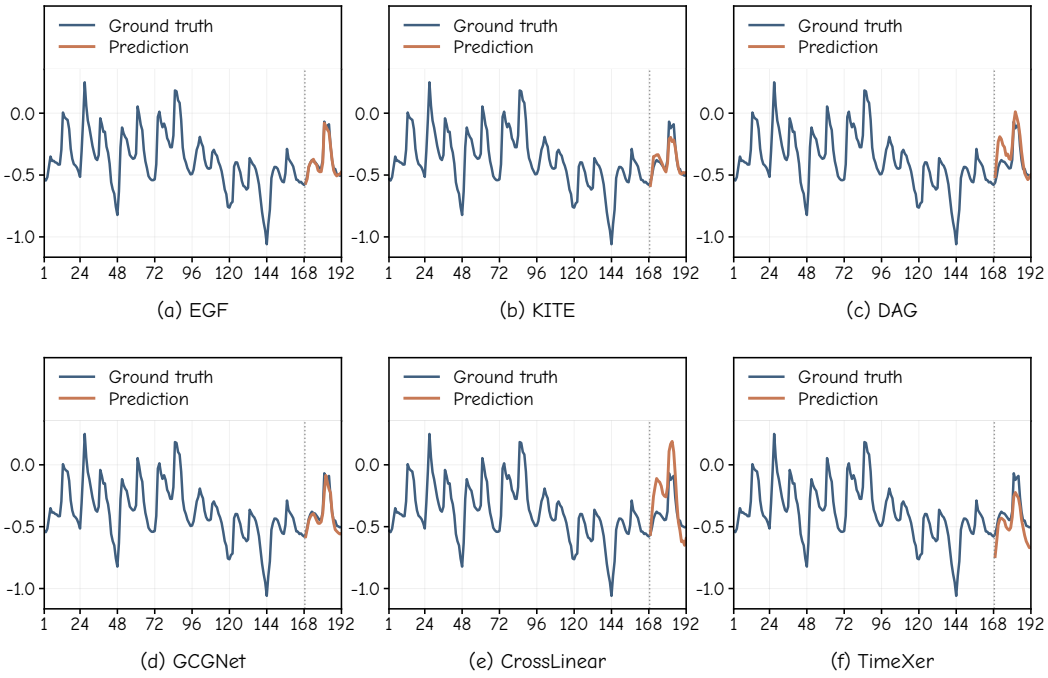
Table 10 reports the complete horizon-wise results on all 12 datasets. EGF results are averaged over five seeds (2024–2028), and † marks our reproduced baselines. Table 11 further reports the five individual EGF runs used to compute the mean and sample standard deviation.

I Limitations and Future Work

EGF targets point forecasting with future exogenous trajectories supplied over the prediction horizon. In the electricity-price datasets these trajectories are operational forecasts, whereas the other datasets supply observed or reanalysis values (ERA5 for wind power), which are more accurate than the environment forecasts available at deployment; extending Environment Response to account for uncertainty in the environment and in the resulting forecast update is a natural next step. Iterative refinement also makes EGF more expensive than lightweight TSF-X backbones (Appendix F), which motivates adapting the number of refinement steps to each forecasting instance. Another direction is to learn the reference environment from context.

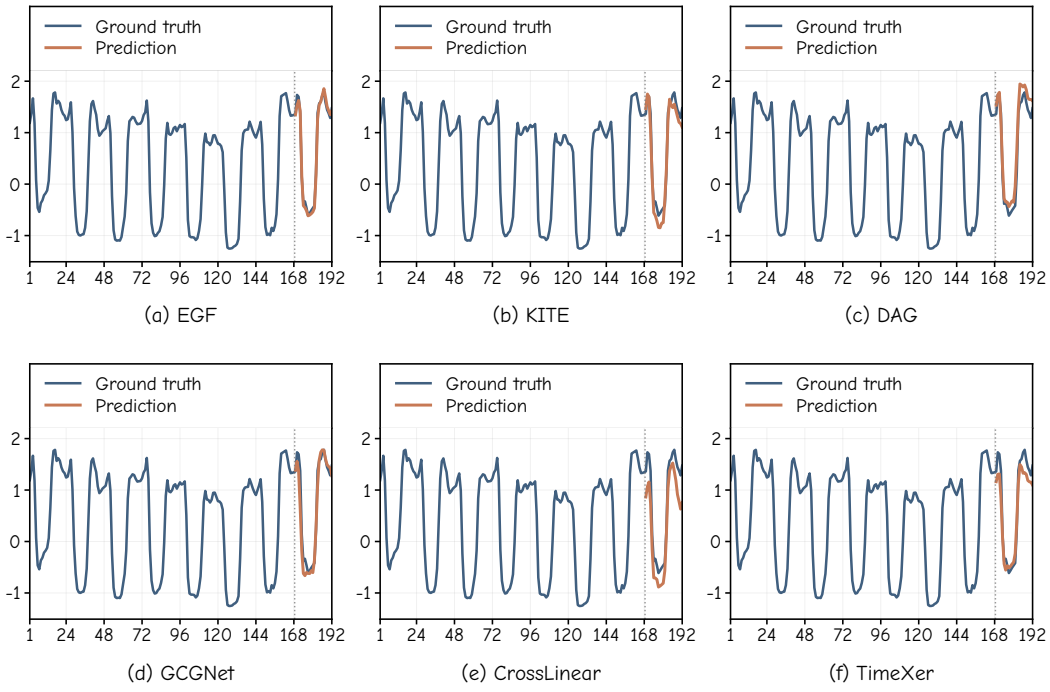


NP: $L = 168$, $H = 24$.

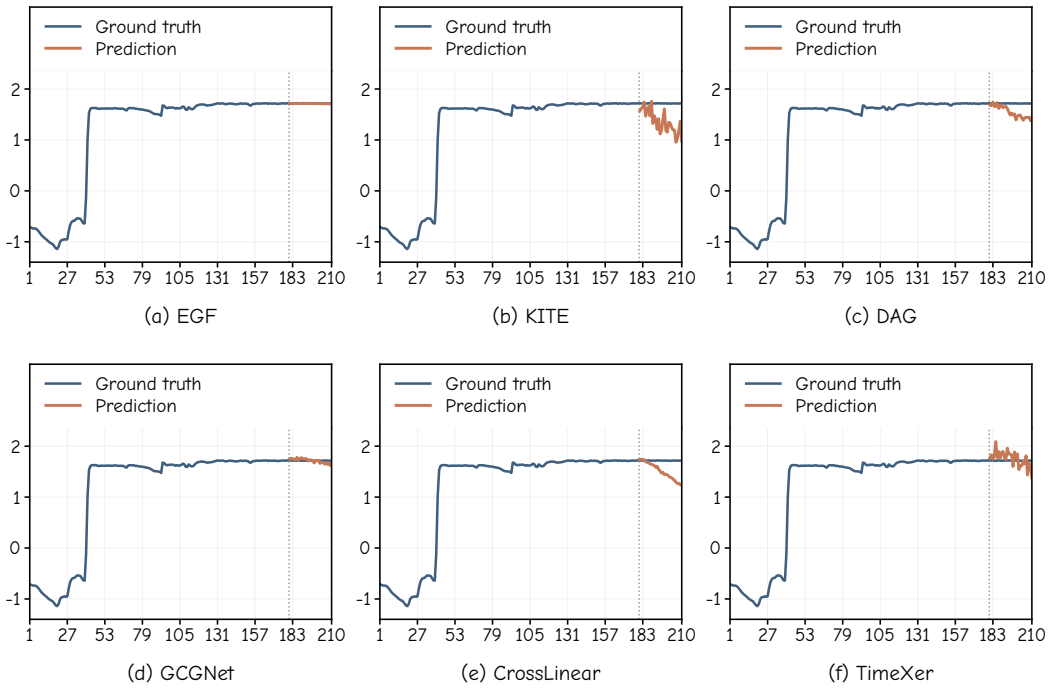


PJM: $L = 168$, $H = 24$.

Figure 10: Prediction cases on NP and PJM by different models under input-168-predict-24 settings.

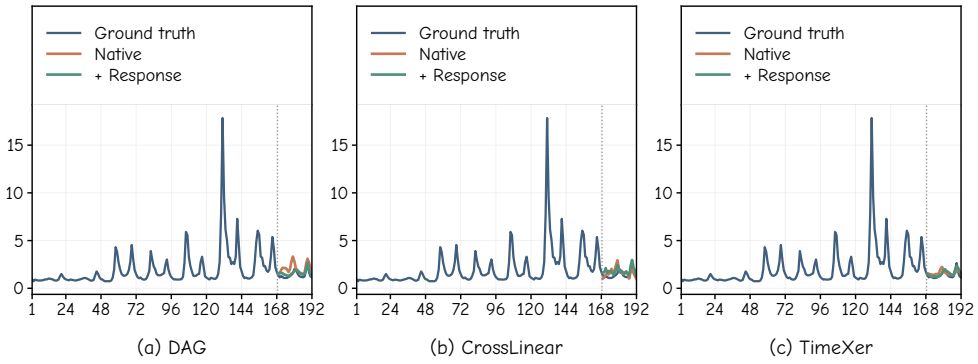


Energy: $L = 168$, $H = 24$.

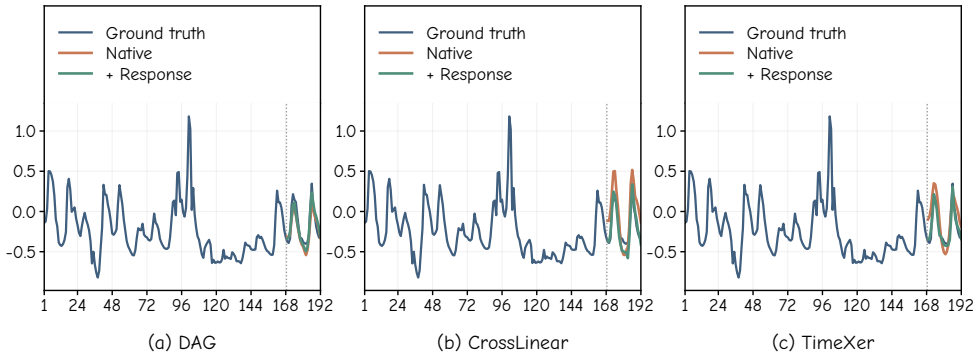


Colbun: $L = 180$, $H = 30$.

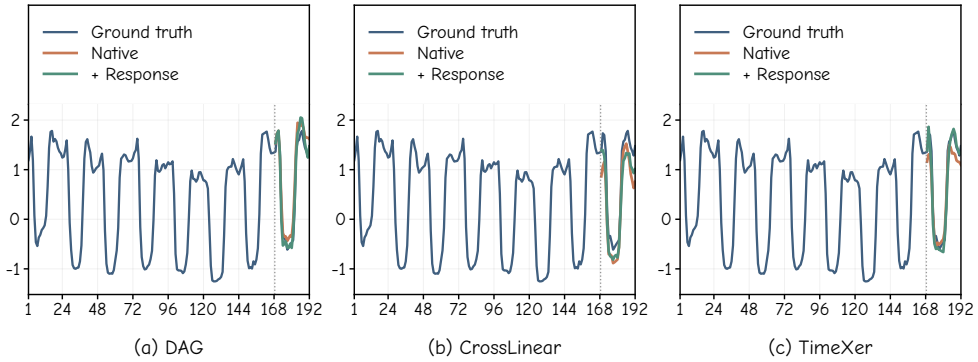
Figure 11: Prediction cases on Energy and Colbun by different models. Energy uses input-168-predict-24, while Colbun uses input-180-predict-30.



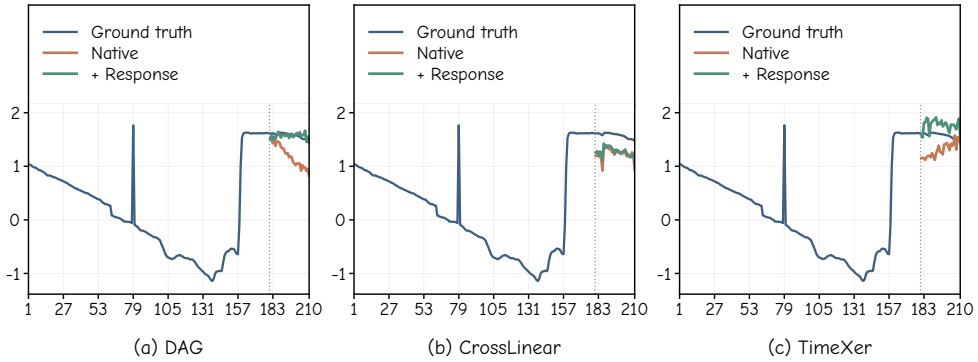
NP: $L = 168$, $H = 24$.



PJM: $L = 168$, $H = 24$.



Energy: $L = 168$, $H = 24$.



Colbun: $L = 180$, $H = 30$.

Figure 12: **Qualitative examples of Response transfer.** Forecast trajectories of DAG, CrossLinear, and TimeXer on NP, PJM, Energy, and Colbun.

Table 10: Full horizon-wise forecasting results on 12 TSF-X datasets. Avg averages the two forecasting horizons. Red and blue mark the best and second-best results, respectively. The 1st Count summarizes horizon-level first places, including ties. Lower is better.

Models		EGF (ours)		KITE (2026)		DAG† (2026)		GCGNet (2026)		CrossLinear† (2025)		TimeKAN (2025)		Amplifier (2025)		xPatch (2025)		DUET (2025)		TimeXer† (2024)	
Datasets	Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	24	**0.172**	**0.213**	0.179	0.223	0.201	0.236	0.197	0.233	0.211	0.266	0.273	0.310	0.252	0.303	0.234	0.278	0.246	0.287	0.233	0.264
	360	**0.426**	**0.401**	0.471	0.422	0.530	0.460	0.496	0.442	0.532	0.509	0.538	0.529	0.587	0.534	0.523	0.461	0.576	0.528	0.637	0.490
	Avg	**0.299**	**0.307**	0.325	0.323	0.366	0.348	0.346	0.337	0.372	0.388	0.406	0.419	0.419	0.418	0.378	0.370	0.411	0.407	0.435	0.377
PJM	24	**0.056**	**0.142**	**0.056**	0.143	0.057	0.143	0.058	0.148	0.089	0.192	0.115	0.244	0.096	0.208	0.077	0.168	0.072	0.166	0.075	0.166
	360	**0.113**	**0.206**	0.135	0.214	0.130	0.217	0.128	0.223	0.135	0.252	0.162	0.281	0.177	0.285	0.131	0.220	0.131	0.228	0.140	0.231
	Avg	**0.084**	**0.174**	0.096	0.178	0.093	0.180	0.093	0.186	0.112	0.222	0.139	0.263	0.137	0.246	0.104	0.194	0.102	0.197	0.108	0.198
BE	24	0.330	**0.226**	0.348	0.240	0.361	0.229	**0.323**	0.227	0.393	0.259	0.451	0.319	0.471	0.339	0.411	0.256	0.432	0.272	0.397	0.253
	360	**0.447**	**0.287**	0.507	0.332	0.485	0.330	0.524	0.347	0.569	0.416	0.645	0.495	0.646	0.487	0.579	0.410	0.597	0.436	0.513	0.330
	Avg	**0.388**	**0.256**	0.427	0.286	0.423	0.280	0.423	0.287	0.481	0.338	0.548	0.407	0.558	0.413	0.495	0.333	0.514	0.354	0.455	0.291
FR	24	**0.304**	**0.165**	0.311	0.178	0.354	0.171	0.332	0.179	0.390	0.226	0.454	0.296	0.459	0.348	0.424	0.235	0.384	0.251	0.366	0.208
	360	**0.451**	**0.247**	0.462	0.271	0.473	0.269	0.478	0.280	0.575	0.370	0.641	0.452	0.648	0.468	0.556	0.350	0.607	0.403	0.489	0.273
	Avg	**0.378**	**0.206**	0.387	0.225	0.414	0.220	0.405	0.230	0.483	0.298	0.547	0.374	0.553	0.408	0.490	0.293	0.495	0.327	0.427	0.241
DE	24	**0.224**	**0.294**	0.262	0.327	0.287	0.327	0.282	0.328	0.387	0.396	0.399	0.412	0.394	0.407	0.399	0.394	0.376	0.378	0.339	0.362
	360	**0.412**	**0.390**	0.437	0.413	0.468	0.419	0.491	0.445	0.581	0.506	0.547	0.479	0.551	0.474	0.550	0.472	0.589	0.482	0.610	0.474
	Avg	**0.318**	**0.342**	0.350	0.370	0.377	0.373	0.387	0.387	0.484	0.451	0.473	0.446	0.473	0.441	0.474	0.433	0.482	0.430	0.475	0.418
Energy	24	**0.062**	**0.190**	0.071	0.208	0.079	0.215	0.065	0.194	0.242	0.419	0.135	0.298	0.138	0.306	0.092	0.240	0.117	0.283	0.122	0.273
	360	**0.114**	**0.262**	0.151	0.305	0.170	0.321	0.179	0.330	0.236	0.385	0.302	0.464	0.328	0.472	0.204	0.367	0.288	0.452	0.212	0.365
	Avg	**0.088**	**0.226**	0.111	0.257	0.124	0.268	0.122	0.262	0.239	0.402	0.218	0.381	0.233	0.389	0.148	0.303	0.202	0.367	0.167	0.319
Sdwpfm1	24	0.377	0.405	0.384	0.410	0.351	**0.396**	0.375	0.409	0.355	0.473	0.418	0.503	**0.346**	0.445	0.669	0.580	0.551	0.564	0.558	0.533
	360	**0.386**	**0.429**	0.450	0.491	0.497	0.523	0.458	0.490	0.491	0.531	0.476	0.566	0.510	0.535	0.592	0.556	0.646	0.577	0.845	0.684
	Avg	**0.381**	**0.417**	0.417	0.451	0.424	0.459	0.416	0.449	0.423	0.502	0.447	0.534	0.428	0.490	0.630	0.568	0.599	0.571	0.701	0.609
Sdwpfm2	24	0.418	0.440	0.443	0.450	**0.373**	**0.412**	0.410	0.443	0.477	0.536	0.474	0.538	0.394	0.462	0.508	0.478	0.445	0.452	0.627	0.570
	360	**0.402**	**0.435**	0.506	0.515	0.582	0.555	0.505	0.514	0.589	0.611	0.520	0.589	0.587	0.563	0.523	0.513	0.584	0.528	0.978	0.736
	Avg	**0.410**	**0.438**	0.475	0.483	0.477	0.483	0.458	0.479	0.533	0.573	0.497	0.564	0.490	0.512	0.516	0.495	0.514	0.490	0.803	0.653
Sdwpfh1	24	**0.352**	**0.412**	0.424	0.464	0.410	0.439	0.381	0.441	0.548	0.585	0.511	0.582	0.576	0.627	0.524	0.505	0.527	0.513	0.651	0.587
	360	**0.377**	**0.441**	0.457	0.517	0.489	0.533	0.446	0.495	0.566	0.601	0.643	0.694	0.497	0.569	0.472	0.504	0.551	0.519	0.849	0.708
	Avg	**0.365**	**0.427**	0.441	0.491	0.449	0.486	0.414	0.468	0.557	0.593	0.577	0.638	0.536	0.598	0.498	0.505	0.539	0.516	0.750	0.648
Sdwpfh2	24	0.439	0.475	0.477	0.498	0.452	0.475	**0.426**	**0.474**	0.467	0.539	0.580	0.614	0.473	0.533	0.640	0.554	0.629	0.563	0.820	0.677
	360	**0.411**	**0.464**	0.524	0.555	0.613	0.599	0.519	0.529	0.607	0.609	0.713	0.729	0.569	0.628	0.579	0.547	0.665	0.569	0.962	0.761
	Avg	**0.425**	**0.470**	0.500	0.526	0.533	0.537	0.472	0.501	0.537	0.574	0.646	0.671	0.521	0.581	0.610	0.551	0.647	0.566	0.891	0.719
Colbun	10	**0.050**	**0.084**	0.055	0.101	0.061	0.098	0.057	**0.084**	0.071	0.102	0.061	0.101	0.071	0.121	0.090	0.104	0.089	0.134	0.113	0.172
	30	**0.111**	**0.150**	0.121	0.243	0.150	0.229	0.138	0.224	0.181	0.288	0.195	0.249	0.275	0.370	0.217	0.315	0.307	0.397	0.176	0.299
	Avg	**0.080**	**0.117**	0.088	0.172	0.106	0.163	0.098	0.154	0.126	0.195	0.128	0.175	0.173	0.245	0.153	0.210	0.198	0.266	0.145	0.236
Rapel	10	0.181	0.194	0.192	0.231	0.145	0.194	**0.140**	**0.191**	0.163	0.209	0.174	0.231	0.181	0.227	0.161	0.216	0.174	0.219	0.301	0.308
	30	0.293	**0.321**	0.296	0.358	0.323	0.412	0.284	0.327	**0.229**	0.339	0.325	0.390	0.333	0.416	0.505	0.546	0.365	0.432	0.387	0.416
	Avg	0.237	**0.257**	0.244	0.294	0.234	0.303	0.212	0.259	**0.196**	0.274	0.249	0.310	0.257	0.322	0.333	0.381	0.269	0.326	0.344	0.362
1st Count		**18**	**20**	1	0	1	2	3	3	1	0	0	0	1	0	0	0	0	0	0	0

Table 11: EGF results over five random seeds. Mean and Std denote the five-seed mean and sample standard deviation. Lower is better.

Setup		Random Seed													
Dataset	Horizon	2024		2025		2026		2027		2028		Mean		Std	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	24	0.1685	0.2116	0.1710	0.2127	0.1710	0.2122	0.1729	0.2144	0.1743	0.2158	0.1715	0.2133	0.0022	0.0017
	360	0.4233	0.4073	0.4283	0.4037	0.4392	0.3969	0.4167	0.3948	0.4205	0.4019	0.4256	0.4009	0.0087	0.0051
PJM	24	0.0561	0.1421	0.0567	0.1422	0.0579	0.1421	0.0557	0.1416	0.0552	0.1407	0.0563	0.1418	0.0010	0.0006
	360	0.1176	0.2117	0.1055	0.2035	0.1156	0.2084	0.1126	0.2019	0.1116	0.2048	0.1126	0.2061	0.0046	0.0040
BE	24	0.3220	0.2175	0.3200	0.2198	0.3277	0.2277	0.3408	0.2353	0.3374	0.2283	0.3296	0.2257	0.0092	0.0071
	360	0.4567	0.2911	0.4501	0.2865	0.4548	0.2894	0.4348	0.2819	0.4396	0.2871	0.4472	0.2872	0.0096	0.0035
FR	24	0.2995	0.1641	0.3052	0.1656	0.3094	0.1674	0.3035	0.1640	0.3019	0.1644	0.3039	0.1651	0.0037	0.0014
	360	0.4584	0.2570	0.4417	0.2472	0.4448	0.2393	0.4663	0.2503	0.4455	0.2418	0.4513	0.2471	0.0105	0.0070
DE	24	0.2239	0.2932	0.2199	0.2932	0.2253	0.2944	0.2240	0.2940	0.2274	0.2945	0.2241	0.2939	0.0028	0.0006
	360	0.4132	0.3878	0.4237	0.3943	0.4067	0.3876	0.4120	0.3923	0.4040	0.3856	0.4119	0.3895	0.0076	0.0036
Energy	24	0.0609	0.1876	0.0610	0.1899	0.0597	0.1861	0.0613	0.1895	0.0650	0.1958	0.0616	0.1898	0.0020	0.0037
	360	0.1153	0.2631	0.1275	0.2791	0.1088	0.2569	0.1074	0.2529	0.1102	0.2564	0.1139	0.2617	0.0082	0.0104
Sdwpfm1	24	0.3820	0.4015	0.3757	0.4060	0.3777	0.4103	0.3743	0.4005	0.3760	0.4061	0.3771	0.4049	0.0030	0.0040
	360	0.3587	0.4137	0.3798	0.4269	0.4022	0.4488	0.4193	0.4356	0.3681	0.4191	0.3856	0.4288	0.0249	0.0139
Sdwpfm2	24	0.4037	0.4353	0.4200	0.4428	0.4242	0.4420	0.4238	0.4399	0.4204	0.4382	0.4184	0.4396	0.0085	0.0030
	360	0.4143	0.4570	0.3887	0.4212	0.3946	0.4246	0.3601	0.4113	0.4509	0.4631	0.4017	0.4355	0.0337	0.0231
Sdwpfh1	24	0.3553	0.4200	0.3452	0.4122	0.3480	0.3981	0.3590	0.4225	0.3545	0.4095	0.3524	0.4124	0.0057	0.0097
	360	0.3684	0.4365	0.3545	0.4556	0.4271	0.4414	0.3837	0.4238	0.3506	0.4477	0.3769	0.4410	0.0310	0.0120
Sdwpfh2	24	0.4251	0.4737	0.4503	0.4757	0.4575	0.4848	0.4072	0.4543	0.4559	0.4877	0.4392	0.4752	0.0221	0.0131
	360	0.3502	0.4262	0.3718	0.4670	0.4062	0.4885	0.4138	0.4435	0.5108	0.4971	0.4105	0.4645	0.0617	0.0298
Colbun	10	0.0446	0.0777	0.0476	0.0788	0.0553	0.0953	0.0520	0.0884	0.0483	0.0782	0.0496	0.0837	0.0041	0.0078
	30	0.0953	0.1372	0.0974	0.1477	0.1085	0.1605	0.1293	0.1501	0.1221	0.1522	0.1105	0.1495	0.0149	0.0084
Rapel	10	0.1783	0.1897	0.1966	0.2119	0.1715	0.1852	0.1812	0.1912	0.1779	0.1911	0.1811	0.1938	0.0093	0.0104
	30	0.2944	0.3198	0.2941	0.3200	0.2672	0.3154	0.3105	0.3256	0.3010	0.3249	0.2935	0.3211	0.0161	0.0042