DIVERSEFLOW: JOINTLY MODELING FORECASTING DIVERSITY AND UNCERTAINTY FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series forecasting is fundamental to many real-world applications, however, temporal observations often exhibit non-stationary dynamics and inherent uncertainty. Existing end-to-end forecasters typically learn a single deterministic mapping from historical observations to future values, which can limit forecast diversity and fails to characterize multiple plausible future outcomes. We propose DiverseFlow, it is a unified framework that jointly models forecast diversity and predictive uncertainty. DiverseFlow employs multiple structurally aligned predictors to capture complementary predictive views and introduces conditional residual flow matching to model flexible predictive distributions around deterministic forecasts. To stabilize distribution learning across heterogeneous temporal dynamics, we further develop conditional scale normalization, which decouples residual magnitude from distributional shape and allows the flow model to learn normalized residual dynamics more effectively. Crucially, DiverseFlow exploits complementary information from flow-generated forecasts to diversify deterministic predictors, enabling uncertainty modeling to directly improve point forecasting rather than serving only as an auxiliary probabilistic objective. Extensive experiments on ten real-world benchmark datasets show that DiverseFlow consistently improves point-forecasting accuracy over competitive baselines and achieves superior probabilistic forecasting performance over conditional Gaussian modeling across multiple metrics. Ablation studies further verify the effectiveness of forecast diversification, conditional scale normalization, and residual flow matching.

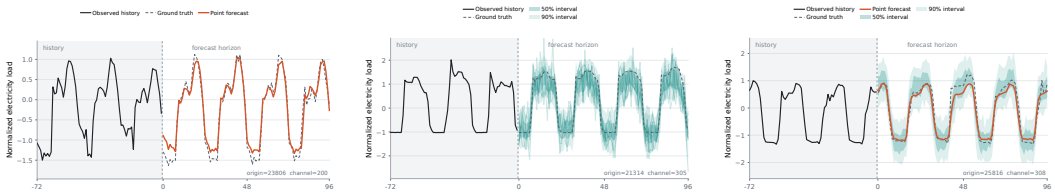
1 INTRODUCTION

Time series forecasting is fundamental to decision-making in a wide range of real-world applications, including finance, energy, transportation, and retail (Hao et al., 2026; Elliott & Timmermann, 2008). Recent deep learning advances have yielded powerful deterministic forecasting architectures, such as PatchTST (Nie et al., 2023), DLinear (Zeng et al., 2023), and iTransformer (Liu et al., 2024b), which capture complex temporal dependencies and achieve strong point-prediction performance. Nevertheless, such models learn a deterministic mapping from historical observations to a single future trajectory. They do not characterize the multiple plausible futures induced by intrinsic stochasticity in real-world time-series data, and cannot provide the calibrated uncertainty estimates critical for risk-aware downstream decision making.

To address uncertainty modeling, probabilistic forecasting methods parameterize conditional predictive distributions, while generative models enable sampling diverse plausible future trajectories. Beyond probabilistic modeling, forecast ensembling improves point-forecast accuracy by aggregating outputs from multiple predictors; the performance of ensembles hinges both on individual member accuracy and forecast diversity.

However, these two lines of work remain largely disconnected. Generative probabilistic models excel at sampling diverse trajectories(de Bézenac et al., 2020; Rasul et al., 2021b; Bilo et al., 2023), yet their learned distributional information is rarely exploited to boost point-forecasting performance. Recent hybrids couple deterministic prediction with conditional generation. D^3U models the uncertain component around a deterministic forecast with diffusion (Li et al., 2025), yet the generated

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) point forecasting **(b) probabilistic forecasting** **(c) joint forecasting**

Figure 1: Three output modes on Electricity, visualized with distinct test windows and independent DiverseFlow inferences. The dashed gray future is a reference. Point forecasting exposes one trajectory, probabilistic forecasting exposes conditional samples and intervals, and DiverseFlow exposes both.

distribution and the point forecast remain treated as separate outputs. Simply sampling more diverse trajectories does not by itself produce complementary point forecasts useful for ensembling. In parallel, conventional time-series ensembles aggregate independently trained deterministic predictors without leveraging generative residuals as additional ensemble candidates. Therefore, the challenge is to model forecasting uncertainty while harnessing the generated distribution to construct complementary forecasts that strengthen ensemble point prediction. Figure 1 contrasts point-only, probabilistic-only, and joint forecasting. The first two modes omit either uncertainty or a separately optimized point estimate. Our goal is to provide both and let the generated distribution contribute directly to point forecasting.

We propose DiverseFlow, a unified framework that bridges deterministic point forecasting, conditional residual flow matching, and validation-weighted forecast ensembling. Instead of generating full future trajectories from scratch, DiverseFlow learns residuals relative to a deterministic forecast center. DiverseFlow adopts conditional scale normalization to disentangle residual magnitude estimation from the learning of distributional shape, making generative training robust to highly heterogeneous residual scales across variables and time windows. The mean of the samples produced by our residual flow head serves as an additional forecasting candidate, combined with two independently trained PatchTST point predictors with different context lengths through convex weights optimized on the validation set. The point branch is trained first and then frozen, so the probabilistic extension requires no joint fine-tuning. This design lets the flow sample mean contribute to point forecasting while the sampled trajectories retain the ability to quantify predictive uncertainty.

Our contributions are as follows: Our main contributions are summarized as follows:

- We propose DiverseFlow, a unified framework for modeling forecast diversity and predictive uncertainty. It integrates deterministic predictors with conditional residual flow matching to produce point forecasts and characterize multiple plausible futures around a deterministic forecast center.
- We leverage the Flow sample mean as an additional forecasting candidate alongside independently trained predictors with different historical contexts. Through validation-fitted convex ensembling, the learned residual distribution contributes directly to point prediction, extending the role of uncertainty modeling beyond probabilistic sampling.
- We introduce conditional scale normalization for residual flow matching, decoupling residual magnitude from distributional shape through an elementwise positive scale. This enables the flow model to learn normalized residual dynamics across variables and time windows, with ablations supporting its effectiveness in probabilistic forecasting.

2 Related Work

Deterministic forecasting. PatchTST uses temporal patches and shared channel-independent parameters to preserve local structure while reducing the cost of attention over long histories (Nie et al., 2023). DLinear demonstrates that decomposition and linear mappings can be competitive on long-term forecasting benchmarks (Zeng et al., 2023). Other approaches model temporal variation in two dimensions (Lai et al., 2018; Wu et al., 2023; 2021) or use variate tokens to capture mul-

tivariate relationships (Liu et al., 2024b; Wang et al., 2024). Language-informed methods include Time-LLM (Jin et al., 2024), UniTime (Liu et al., 2024a), TimeCMA (Liu et al., 2025), and T3Time (Chowdhury et al., 2026). These approaches motivate diverse representations of forecasting context; a deterministic output alone, however, does not describe predictive uncertainty.

Conditional generative forecasting. Probabilistic forecasting can either parameterize a chosen predictive distribution or learn to generate conditional samples(Wu et al., 2026; Zhao et al., 2024; Yoon et al., 2019). TimeGrad applies denoising diffusion to autoregressive forecasting (Rasul et al., 2021a); CSDI learns conditional distributions of unobserved entries and can be applied to forecasting (Tashiro et al., 2021). TimeDiff develops non-autoregressive conditional diffusion (Shen & Kwok, 2023), and TMDM integrates Transformer information into diffusion forecasting (Li et al., 2024). D^3U combines deterministic prediction with conditional diffusion to model uncertainty (Li et al., 2025). Flow matching instead trains a continuous vector field by regression along probability paths (Lipman et al., 2023). Recent RDIT, CGFM, and WrapFlow study residual conditional diffusion, auxiliary-guided residual flow matching, and continuous-time residual vector fields, respectively (Lai et al., 2025; Xu et al., 2025; Shen et al., 2026). These works establish residual generation as an active direction: RDIT emphasizes residual diffusion and scale selection, CGFM emphasizes the auxiliary source distribution and conditional path, and WrapFlow targets continuous-time irregular series. Our focus is narrower: conditional elementwise scale normalization in residual space and explicit evaluation of the generated mean as a point-forecast candidate. The distinction is therefore about these interfaces and their matched evaluation, not the use of flow matching alone.

Complementarity in ensembles. An ensemble's squared error depends on both member errors and their relationships. Different contexts or architectures can produce useful disagreement, but inaccurate disagreement can also hurt. In DiverseFlow, the flow mean is a separately evaluated candidate rather than an automatically beneficial consequence of stochastic sampling. We use a convex combination fitted on validation data and compare two- and three-member ensembles with the same deterministic members to assess the contribution of the flow output.

3 METHOD

Figure 3 summarizes the three components of DiverseFlow: external and internal PatchTST point predictors and a flow-matching residual predictor. The point predictors use separate initialization and historical contexts. The residual branch transports Gaussian noise conditioned on sequence features, while a separate scale network restores residual magnitude. Validation-fitted convex weights combine the point predictions and the Flow sample mean.

3.1 PROBLEM DEFINITION AND FRAMEWORK OVERVIEW

Given a multivariate history $X \in \mathbb{R}^{B\times L\times C}$ before the forecast origin, the objective is to predict $Y \in \mathbb{R}^{B\times H\times C}$, where B is the batch size, L the available history length, H the forecast horizon, and C the number of variables. The framework provides a point forecast $\widehat{Y}$ and K probabilistic forecast samples $\{\widetilde{Y}^{(k)}\}_{k=1}^{K}$, each with the same shape as Y.

The internal predictor f_θ and external predictor g_ω use historical windows $X_f \in \mathbb{R}^{B\times L_f\times C}$ and $X_g \in \mathbb{R}^{B\times L_g\times C}$, respectively. Let $X_c \in \mathbb{R}^{B\times L_c\times C}$ denote the historical window used by the probabilistic branch. These windows are trailing segments of X, with $L_f, L_g, L_c \le L$, and end at the same forecast origin. Both point predictors target the same future sequence:

$$\widehat{Y}_g = g_\omega(X_g), \qquad \widehat{Y}_f = f_\theta(X_f), \qquad \widehat{Y}_g, \widehat{Y}_f \in \mathbb{R}^{B\times H\times C}. \tag{1}$$

The parameter sets ω and θ are trained independently with a point-forecasting objective. Different initialization and context lengths introduce complementary views.

For the two-predictor configuration, the residual center is

$$m = \alpha_m \widehat{Y}_g + (1-\alpha_m)\widehat{Y}_f, \qquad 0 \le \alpha_m \le 1. \tag{2}$$

The center coefficient α_m is distinct from the final ensemble weights in Equation 12. The fixed-candidate optimization in Appendix A treats m and the resulting Flow candidate as independent of those final weights.

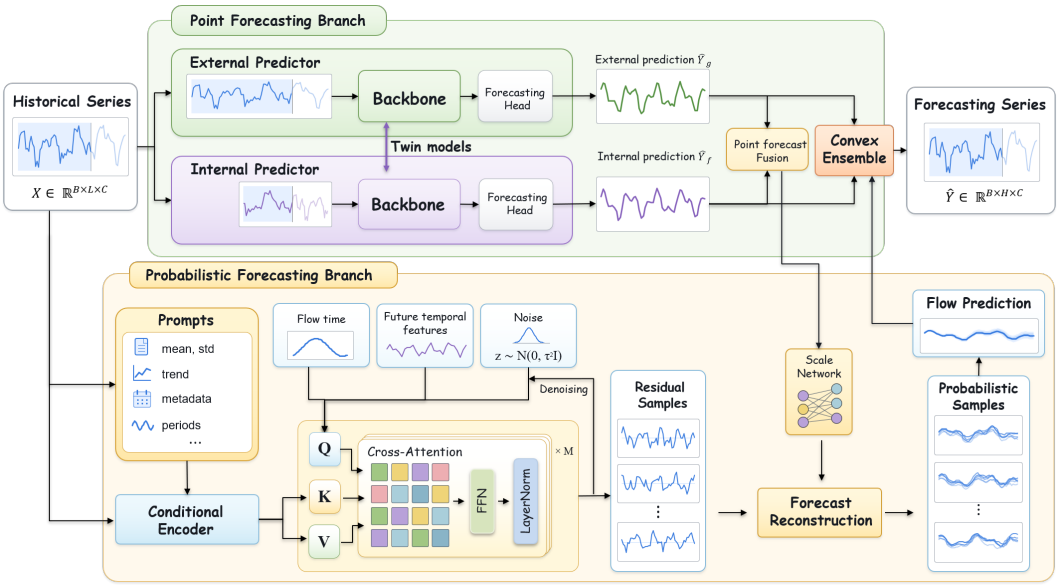


Figure 2: Overall architecture of DiverseFlow. The upper branch contains the internal and external point predictors. The lower branch contains the conditional encoder, scale network, and residual flow-matching branch. Forecast reconstruction combines the center m, scale s, and sampled normalized residuals. The Flow sample mean is combined with the two point predictions through a convex combination.

A scale network with parameters η predicts $s = s_\eta(X_c, m) \in (0,\infty)^{B\times H\times C}$. The normalized residual is

$$r = (Y - m) \oslash s, \qquad m, r \in \mathbb{R}^{B\times H\times C}. \tag{3}$$

Here, $\oslash$ denotes element-wise division. Unless stated otherwise, residual states and generated forecasts retain the batch dimension. Forecasts are reconstructed by the inverse transformation in Equation 10.

3.2 Network Architecture

3.2.1 Point Forecasting Branch

Diversity is introduced through random initialization and different window lengths. Each predictor divides a continuous historical window into short patches, treats each patch as a token, and uses a Transformer to predict the future sequence in a single step. We describe one predictor and one batch sample below, omitting their indices and writing $\ell \in \{L_f, L_g\}$ for its context length.

For channel $i \in \{1,\dots,C\}$, let $x^{(i)}_{1:\ell}$ be its historical signal. We compute the window mean μ_i and standard deviation σ_i, and normalize $\tilde{x}^{(i)}_t = (x^{(i)}_t - \mu_i)/(\sigma_i + \varepsilon)$ for $t = 1,\dots,\ell$, where $\varepsilon > 0$ ensures numerical stability. Channels share parameters within each predictor but are processed separately, without cross-channel attention. With patch length P and stride S, we extract

$$\mathbf{p}^{(i)}_n = \left[\tilde{x}^{(i)}_{(n-1)S+1}, \dots, \tilde{x}^{(i)}_{(n-1)S+P}\right]^\top \in \mathbb{R}^P, \quad n = 1, ..., N_{patch}. \tag{4}$$

Each patch is projected and combined with positional encoding as $\mathbf{e}^{(i)}_n = W_p\mathbf{p}^{(i)}_n + \mathbf{e}_{\text{pos},n}$, where $W_p \in \mathbb{R}^{D\times P}$, $\mathbf{e}_{\text{pos},n} \in \mathbb{R}^D$, and D is the hidden dimension. Stacking the tokens gives $E^{(i)} \in \mathbb{R}^{N_{\text{patch}}\times D}$. A stack of N_{enc} Transformer encoder layers, each containing multi-head attention and a feed-forward network, produces $Z^{(i)} \in \mathbb{R}^{N_{\text{patch}}\times D}$.

We flatten the encoded patches as $\mathbf{z}^{(i)} = \text{vec}(Z^{(i)}) \in \mathbb{R}^{N_{\text{patch}}D}$, where vec stacks all matrix entries into a vector. A linear head predicts the next H normalized values:

$$\hat{\tilde{\mathbf{x}}}^{(i)}_{\ell+1:\ell+H} = W_f\mathbf{z}^{(i)} + \mathbf{b}_f. \tag{5}$$

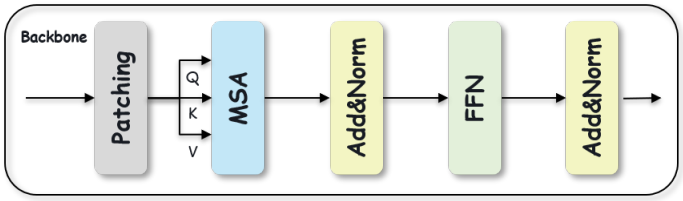


Figure 3: The description of backbone.

Here, $W_f \in \mathbb{R}^{H\times(N_{\text{patch}}D)}$, $\mathbf{b}_f \in \mathbb{R}^H$, and the output is in $\mathbb{R}^H$. These head parameters belong to the predictor being described. Using the same statistics and stabilizer as in normalization, we recover

$$\widehat{\mathbf{x}}^{(i)}_{\ell+1:\ell+H} = (\sigma_i + \varepsilon)\widehat{\widehat{\mathbf{x}}}^{(i)}_{\ell+1:\ell+H} + \mu_i \mathbf{1}_H, \tag{6}$$

where $\mathbf{1}_H \in \mathbb{R}^H$ is the all-ones vector. Thus, both predictors return forecasts on the same scale as Y before residual construction or ensembling.

Restoring the batch dimension, each predictor is trained separately using

$$\mathcal{L}^{(j)}_{\text{point}} = \frac{1}{BHC}\|\widehat{Y}_j - Y\|_F^2 = \frac{1}{BHC}\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{i=1}^{C}\left(\widehat{Y}_{j,b,h,i} - Y_{b,h,i}\right)^2, \quad j \in \{f, g\}. \tag{7}$$

The squared Frobenius norm $\|\cdot\|_F^2$ sums squared entries over all displayed tensor dimensions; b, h, and i index batch samples, forecast steps, and variables, respectively.

3.2.2 Probabilistic forecasting branch

Conditional encoding. A conditional encoder with parameters ψ constructs

$$c = E_\psi(X_c, T_{\text{past}}, P_{\text{dyn}}, P_{\text{meta}}), \tag{8}$$

where c collects the condition tokens read by the vector-field decoder through cross-attention. T_{past} denotes temporal features aligned with X_c; P_{dyn} contains prompts extracted from that observed window, including its mean, standard deviation, and overall trend; and P_{meta} contains dataset-level and variable-level metadata, including the total number of observations, sampling interval, and primary and secondary periods. The future temporal features T_{future} are aligned with the H forecast steps and are known at the forecast origin; they do not contain future observations.

Residual flow matching. The decoder takes the current residual state, continuous flow time, and future temporal features as inputs, and reads the conditional representation c through cross-attention. The initial noise is sampled from a Gaussian source distribution $z \sim \mathcal{N}(0, \tau^2 I)$, and the flow time is sampled independently as $u \sim \mathcal{U}(0, 1)$. Along a linear conditional path, the interpolated state and target velocity are defined as $r_u = (1-u)z + ur, v^* = r - z$. The vector field is trained by regressing the conditional velocity along this path:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{X,Y,z,u}\left[\frac{1}{BHC}\|v_\phi(r_u, u, c, T_{\text{future}}) - (r - z)\|_F^2\right]. \tag{9}$$

At inference, let $\widetilde{r}^{(k)}_u$ denote the generated ODE trajectory, distinguished from the training interpolation r_u. For each independent source draw $z^{(k)}$, $k = 1, \ldots, K$, we solve

$$\begin{aligned} \frac{\mathrm{d}\widetilde{r}^{(k)}_u}{\mathrm{d}u} &= v_\phi(\widetilde{r}^{(k)}_u, u, c, T_{\text{future}}), \\ \widetilde{r}^{(k)}_0 &= z^{(k)}, \qquad \widetilde{Y}^{(k)} = m + s \odot \widetilde{r}^{(k)}_1. \end{aligned} \tag{10}$$

Here $\odot$ denotes element-wise multiplication. DiverseFlow uses fixed-step Euler integration with N_{step} steps of size $1/N_{\text{step}}$. Equation 9 uses the observed training target through r, whereas Equation 10 requires only historical conditions, known future temporal features, the point center, the predicted scale, and random noise. New source draws produce new conditional forecast trajectories.

Conditional scale normalization. Residual magnitudes can vary substantially across datasets and even across different time periods. Directly learning flow matching on highly volatile residuals can increase the optimization difficulty. To address this issue, we introduce conditional scale normalization. A conditional scale model $s(X, m)$ predicts a positive element-wise scale. During training, the residual is normalized by dividing it by s, and during sampling, its magnitude is restored by multiplying it by s. This design decouples how large the residual is from how the residual is distributed, allowing the vector field to focus on learning the distributional shape and temporal structure of normalized residuals, thereby stabilizing distribution learning under heterogeneous time-series dynamics.

3.2.3 ENSEMBLING

Flow Matching generates K future prediction trajectories based on the conditional information and averages them to form the third point-prediction member. Let the k-th reconstructed prediction trajectory be $\widetilde{Y}^{(k)}$. The Flow prediction output is

$$\widehat{Y}_F = \frac{1}{K}\sum_{k=1}^{K}\widetilde{Y}^{(k)} = m + s \odot \left(\frac{1}{K}\sum_{k=1}^{K}\widetilde{r}_1^{(k)}\right). \tag{11}$$

Next, the outputs of the external point predictor, internal point predictor, and Flow predictor are directly combined through a three-model convex combination:

$$\widehat{Y} = w_g\widehat{Y}_g + w_f\widehat{Y}_f + w_F\widehat{Y}_F, \qquad w_g, w_f, w_F \geq 0, \quad w_g + w_f + w_F = 1. \tag{12}$$

The ensemble weights of the three models are jointly optimized according to the mean squared error on the validation set and obtained by solving a convex quadratic optimization problem. In this way, the final point prediction can simultaneously exploit the predictive differences introduced by different historical window lengths and the residual correction information generated by Flow Matching.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. The main point-forecasting experiments cover eight public datasets: ETTh1, ETTh2, ETTm1, ETTm2, ECL, Traffic, Weather, and Exchange, encompassing electricity, traffic, weather, and exchange-rate forecasting. The main probabilistic forecasting experiments cover six datasets: ETTm1, ETTm2, Weather, Solar-Energy, Electricity, and Traffic. Both ablation studies cover the eight point-forecasting datasets listed above. The data are divided chronologically into training, validation, and test sets, with normalization statistics estimated from the training set.

Evaluation metrics. Point forecasting is evaluated using mean squared error (MSE) and mean absolute error (MAE). Probabilistic forecasting is evaluated using the continuous ranked probability score (CRPS) and the cross-variable aggregate score $\mathrm{CRPS}_{\mathrm{sum}}$, which measure the predictive distributions of individual variables and the distribution of the sum across variables, respectively. Lower values are better for all four metrics.

Compared methods. The point-forecasting baselines include T3Time, TimeCMA, TimeLLM, UniTime, TimesNet, DLinear, iTransformer, and PatchTST (Chowdhury et al., 2026; Liu et al., 2025; Jin et al., 2024; Liu et al., 2024a; Wu et al., 2023; Zeng et al., 2023; Liu et al., 2024b; Nie et al., 2023). The probabilistic-forecasting baselines include TimeGrad, CSDI, TimeDiff, TMDM, and DşU (Rasul et al., 2021a; Tashiro et al., 2021; Shen & Kwok, 2023; Li et al., 2024; 2025).

Implementation details. DiverseFlow uses a staged training procedure, and model selection and ensemble weights are determined on the validation set. Point forecasting is evaluated at four forecast horizons: 96, 192, 336, and 720. Table 1 summarizes the average performance across the four horizons, while the results for each individual horizon are provided in Appendix C. The external and internal predictors use historical contexts of 720 and 96 steps, respectively. The forecast horizon for

Table 1: Point forecasting averaged over horizons 96, 192, 336, and 720. MSE and MAE occupy separate columns; lower is better. Bold marks the lowest displayed value for each dataset and metric. ECL denotes Electricity.

Dataset	DiverseFlow (Ours)		T3Time (2026)		TimeCMA (2025)		Time-LLM (2024)		UniTime (2024a)		TimesNet (2023)		DLinear (2023)		iTransformer (2024b)		PatchTST (2023)	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.341**	**0.377**	0.372	0.393	0.380	0.392	0.410	0.409	0.385	0.399	0.400	0.406	0.403	0.407	0.407	0.410	0.392	0.402
ETTm2	**0.250**	**0.312**	0.279	0.322	0.275	0.323	0.296	0.340	0.293	0.334	0.291	0.333	0.350	0.401	0.288	0.332	0.285	0.328
ETTh1	0.419	0.432	**0.418**	**0.430**	0.423	0.431	0.448	0.443	0.442	0.448	0.458	0.450	0.456	0.452	0.454	0.447	0.463	0.449
ETTh2	0.351	0.392	**0.348**	**0.390**	0.372	0.397	0.381	0.404	0.378	0.403	0.414	0.427	0.559	0.515	0.383	0.407	0.395	0.414
ECL	**0.164**	**0.265**	0.170	0.266	0.174	0.269	0.195	0.288	0.216	0.306	0.192	0.295	0.212	0.300	0.178	0.270	0.207	0.289
Weather	**0.216**	**0.261**	0.244	0.275	0.250	0.276	0.275	0.291	0.253	0.276	0.259	0.287	0.265	0.317	0.258	0.278	0.257	0.280
Traffic	**0.406**	0.283	0.466	0.287	0.428	**0.282**	0.626	0.378	0.481	0.304	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383
Exchange	0.357	0.438	**0.353**	**0.401**	0.395	0.429	0.372	0.416	0.364	0.404	0.416	0.443	0.354	0.414	0.360	0.403	0.390	0.429

Table 2: Probabilistic forecasting at horizon 96. Bold marks the lowest CRPS or $\mathrm{CRPS}_{\mathrm{sum}}$ for each dataset. These scores evaluate the probabilistic configuration, separately from the point ensemble in Table 1.

Dataset	DiverseFlow (Ours)		D^3U (2025)		TMDM (2024)		TimeDiff (2023)		CSDI (2021)		TimeGrad (2021a)	
	CRPS	$\mathrm{CRPS}_{\mathrm{sum}}$	CRPS	$\mathrm{CRPS}_{\mathrm{sum}}$	CRPS	$\mathrm{CRPS}_{\mathrm{sum}}$	CRPS	$\mathrm{CRPS}_{\mathrm{sum}}$	CRPS	$\mathrm{CRPS}_{\mathrm{sum}}$	CRPS	$\mathrm{CRPS}_{\mathrm{sum}}$
ETTm1	**0.277**	**0.533**	0.285	0.574	0.429	0.633	0.454	0.846	0.773	0.852	0.665	0.996
ETTm2	**0.204**	**0.108**	0.243	0.141	0.380	0.226	0.316	0.180	0.625	0.782	0.785	1.051
Weather	**0.163**	**0.243**	0.207	0.283	0.226	0.292	0.293	0.400	0.508	0.465	0.482	0.503
Solar-Energy	**0.177**	**0.235**	0.186	0.266	0.375	0.267	0.900	1.164	0.649	0.681	0.783	1.167
Electricity	**0.197**	**0.125**	0.202	0.160	0.446	0.137	0.475	0.594	0.465	0.823	0.503	1.452
Traffic	0.252	0.231	**0.232**	0.186	0.552	**0.179**	0.671	0.823	0.612	1.275	0.657	1.683

probabilistic forecasting and both ablation studies is 96. Probabilistic forecasting uses an independent fixed-center configuration with a 96-step context and generates 64 test samples using four-step Euler integration. The ablation results are aggregated over three random seeds.

Training strategy. The model is trained in two stages. In the first stage, the point-prediction branch is trained using mean squared error. In the second stage, the point-prediction branch, including its parameters and BatchNorm running statistics, is frozen, and only the conditional encoder and vector-field decoder are trained so that the point output remains unchanged during probabilistic training. The conditional scale network is jointly trained during this stage.

4.2 Point and probabilistic forecasting results

Point forecasting. In Table 1, DiverseFlow achieves the lowest average MSE on ETTm1, ETTm2, ECL, Weather, and Traffic, and the lowest average MAE on the first four. On ETTm1, its MSE and MAE are 0.341 and 0.377, compared with the best baseline values of 0.372 and 0.392. On Weather, the corresponding values are 0.216 and 0.261. The complete results show that both metrics remain favorable across all four horizons on ETTm1 and Weather, so their average gains are not confined to a single horizon.

Rankings nevertheless depend on the dataset and metric. Traffic MSE is 0.406, below TimeCMA's 0.428, but MAE is 0.288, above TimeCMA's 0.282. T3Time obtains lower average errors on ETTh1 and ETTh2. On Exchange, DiverseFlow's average MSE is close to T3Time and DLinear, but its MAE is less competitive; the favorable 96-step result does not persist at longer horizons. These comparisons support benefits on multiple datasets rather than uniform superiority.

Probabilistic forecasting. Table 2 presents the probabilistic forecasting results at a forecast horizon of 96 across six datasets. DiverseFlow achieves the lowest CRPS and $\mathrm{CRPS}_{\mathrm{sum}}$ on ETTm1, ETTm2, Weather, Solar-Energy, and Electricity, obtaining the best results on 10 of the 12 dataset–metric pairs. On ETTm2, the two scores are 0.204 and 0.108, respectively, lower than the corresponding D^3U scores of 0.243 and 0.141. On Weather, the scores are 0.163 and 0.243, also lower

Table 3: Ablation of conditional scale normalization at horizon 96. Results are CRPS (mean ± std) over three seeds.

Dataset	Unit scale	Conditional scale	Decrease
ETTh1	0.343 ± 0.0137	$\mathbf{0.305 \pm 0.0029}$	11.09%
ETTh2	0.378 ± 0.0182	$\mathbf{0.267 \pm 0.0016}$	29.37%
ETTm1	0.296 ± 0.0081	$\mathbf{0.277 \pm 0.0039}$	6.10%
ETTm2	0.229 ± 0.0096	$\mathbf{0.204 \pm 0.0017}$	10.77%
Electricity	0.298 ± 0.0004	$\mathbf{0.197 \pm 0.0010}$	33.78%
Traffic	0.361 ± 0.0030	$\mathbf{0.253 \pm 0.0012}$	29.94%
Weather	0.209 ± 0.0022	$\mathbf{0.164 \pm 0.0011}$	21.45%
Exchange	0.217 ± 0.0043	$\mathbf{0.158 \pm 0.0048}$	27.34%

Table 4: Point ensemble ablation at horizon 96: mean errors over three seeds. External and internal denote individual PatchTST models; "+ Flow" adds the flow sample mean to the two-member candidate set. Weights may be zero. Bold marks the lowest value for each dataset and metric.

Dataset	External		Internal		Two members		+ Flow	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.388	0.417	0.383	0.398	0.373	0.395	**0.366**	**0.393**
ETTh2	0.299	0.359	0.298	0.344	0.281	0.342	**0.274**	**0.339**
ETTm1	0.296	0.353	0.326	0.364	0.289	0.349	**0.282**	**0.340**
ETTm2	0.165	0.259	0.174	0.258	0.163	0.255	**0.158**	**0.250**
Electricity	0.141	0.246	0.185	0.274	0.141	0.246	**0.137**	**0.242**
Traffic	0.381	0.277	0.502	0.324	0.380	0.277	**0.380**	**0.272**
Weather	0.160	0.210	0.178	0.219	0.158	0.207	**0.150**	**0.203**
Exchange	0.099	0.227	0.085	0.202	0.085	0.203	**0.084**	**0.202**

than D^3U's 0.207 and 0.283. On Electricity, DiverseFlow achieves a CRPS of 0.197, close to D^3U's 0.202, while its aggregate score of 0.125 is lower than TMDM's 0.137. These results indicate that the model performs well in both per-variable forecasting and forecasting after aggregating variables.

On Traffic, DiverseFlow obtains a CRPS of 0.252, which is higher than D^3U's 0.232. Its aggregate score is 0.231, also higher than TMDM's 0.179 and D^3U's 0.186. Together with the point-forecasting results, these findings show that lower point-forecasting MSE does not automatically correspond to better probabilistic scores. The two tasks focus on different properties of the predictions and therefore need to be evaluated separately. The aggregate score provides additional information about forecasting at the cross-variable aggregation level, but it cannot by itself characterize the quality of fitting the complete joint distribution.

4.3 Component ablations

Separating scale from distributional shape. We compare direct residual modeling with $s = 1$ against flow matching with conditional scale normalization. Both settings share the fixed point center and flow training configuration. Table 3 reports mean CRPS and standard deviation over three seeds.

Conditional scaling lowers mean CRPS on all eight datasets. Electricity improves from 0.297 to 0.197, and Traffic from 0.360 to 0.252. The difference is smaller on ETTm1 but has the same direction. Standard deviations decrease on ETTh1, ETTh2, ETTm1, ETTm2, Traffic, and Weather, and increase slightly on Electricity and Exchange. Improved average predictive quality therefore does not imply uniformly reduced variation across training seeds.

Adding the flow output to the ensemble. Table 4 compares the external and internal PatchTST models, their two-member ensemble, and the ensemble that additionally includes the flow mean. Both ensembles reuse the same point models and fit weights on validation data. The external and internal contexts are 720 and 96 steps, respectively, with a 96-step horizon.

The three-member ensemble achieves the lowest MSE on all eight datasets and the lowest MAE on six. On Electricity, the two-member ensemble has the same errors as the external model, while adding the flow mean reduces both metrics. On Exchange, the internal model outperforms the external one, yet the three-member ensemble improves slightly on both the internal model and the two-member ensemble. In contrast, the two-member ensemble has the lowest MAE on ETTh1 and Weather. The very small differences on datasets such as ETTh2 should be interpreted as observations under the current evaluation, not as established statistically significant gains.

4.4 Discussion

Complementarity of the Flow output and metric differences. Across the eight datasets, adding the Flow sample mean reduces ensemble MSE. On Electricity, the two-member ensemble reduces to the external predictor, whereas adding Flow reduces both errors. The probabilistic branch thus supplies a useful residual-based correction for point forecasting, rather than only producing predictive samples. We note that random variation among samples does not by itself yield complementarity: ensemble error depends jointly on individual errors and member disagreement, and a full account of the mechanism would require analyzing error correlations and varying the number of samples. Finally, MSE improves while MAE increases on ETTh1 and Weather; since the weights are fitted on validation MSE, the combination rule should be chosen according to the target metric.

Conditional scale and probabilistic distribution modeling. The conditional-scale ablation reduces mean CRPS on all eight datasets, supporting the design of first estimating residual magnitude and then learning the standardized distributional shape: rescaling the generated residuals preserves the appropriate magnitude for each historical window, while allowing Flow to learn the conditional distribution in standardized space. We note that the main probabilistic experiment and this ablation address different questionsthe former evaluates the complete pipeline against competing methods, while the latter isolates the role of conditional scaling: on Traffic, scaling greatly improves the unscaled Flow model, yet the complete method still does not attain the best probabilistic score, showing that component-level gains do not imply outperforming every baseline. Finally, CRPS and $CRPS_{sum}$ reflect per-variable and aggregated forecasting quality, respectively, and reporting both gives a more comprehensive evaluation.

Scope and future research. Across four forecast horizons, DiverseFlow shows persistent advantages on ETTm1 and Weather, whereas rankings on ETTh1, ETTh2, and Exchange vary with the forecast horizon and evaluation metric, so its effectiveness should be assessed relative to the specific dataset and forecast range. The point ensemble and the fixed-center probabilistic head also use different evaluation configurations, providing separate evidence for ensemble complementarity and conditional distribution modeling; future work could jointly evaluate point accuracy, probabilistic scores, and inference cost under a unified configuration, and examine the effects of sample count, integration steps, and forecast horizon. Finally, the three-seed summaries provide a reference for training variability, but more repeated experiments and paired analyses are needed to assess the stability of small metric differences.

5 Conclusion

We presented DiverseFlow, a unified framework for modeling forecast diversity and predictive uncertainty. It combines deterministic predictors with different historical contexts and conditional residual flow matching to produce point forecasts and probabilistic samples. Conditional scale normalization decouples residual magnitude from distributional shape, while validation-weighted ensembling incorporates the Flow sample mean as an additional forecasting candidate.

Experiments across multiple benchmarks demonstrate improvements in point and probabilistic forecasting under their respective evaluation settings. Ablation studies further support the benefits of conditional scale normalization for probabilistic forecasting and the contribution of the Flow sample mean to point-forecasting accuracy. These gains vary across datasets, horizons, and metrics. Future work will extend probabilistic evaluation to longer horizons and jointly assess point accuracy, probabilistic quality, and inference cost under a unified configuration.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, assisted with research ideation and method discussion, experimental code development and execution, interpretation of experimental results, and manuscript drafting, translation, editing, and typesetting. The numerical tables in this English manuscript were transcribed from the author-provided manuscript. Scientific judgments, verification of experimental claims, and responsibility for the final text, code, and reported results remain with the authors.

REPRODUCIBILITY STATEMENT

Section 3 specifies the residual parameterization, flow objective, sampling procedure, and ensemble optimization. Section 4.1 distinguishes the point and probabilistic evaluation configurations. Appendix A derives the validation optimization, Appendix B reports supplementary two-member weights, and Appendix D provides the complete four-horizon point results. The ablation tables explicitly identify the number of random seeds and the components held fixed.

REFERENCES

Marin Bilo, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with stochastic process diffusion, 2023. URL https://arxiv.org/abs/2211.02590.

Abdul Monaf Chowdhury, Rabeya Akter, and Safaeid Hossain Arib. T3Time: Tri-modal time series forecasting via adaptive multi-head alignment and residual fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20597–20605, 2026. doi: 10.1609/aaai.v40i25.39196. URL https://ojs.aaai.org/index.php/AAAI/article/view/39196.

Emmanuel de Bézenac, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohlke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, and Tim Januschowski. Normalizing kalman filters for multivariate time series analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2995–3007. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f47cef5e38c952f94c5d61726027439-Paper.pdf.

Graham Elliott and Allan Timmermann. Economic forecasting. *Journal of Economic Literature*, 46 (1):356, March 2008. doi: 10.1257/jel.46.1.3. URL https://www.aeaweb.org/articles?id=10.1257/jel.46.1.3.

Fei Hao, Xiaofeng Zhang, Yepeng Liu, Yujuan Sun, Hua Wang, Lin Yang, and Ren Wang. Ftdasc: A frequency-time domain approach with stationarity correction for multivariate time series forecasting. *Expert Systems with Applications*, 308:131164, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2026.131164. URL https://www.sciencedirect.com/science/article/pii/S0957417426000783.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01728.

Chih-Yu Lai, Yu-Chien Ning, and Duane S. Boning. Rdit: Residual-based diffusion implicit models for probabilistic time series forecasting. *arXiv preprint arXiv:2509.02341*, 2025. URL https://arxiv.org/abs/2509.02341.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 95104, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210006. URL https://doi.org/10.1145/3209978.3210006.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a877dcfbf01491ef9e6dbc31cea4bba2-Abstract-Conference.html.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Baolin Sun, and Mingyuan Zhou. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/516a9317af9d89e9f2251bd7fde49b8f-Abstract-Conference.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025. doi: 10.1609/aaai.v39i18.34067. URL https://ojs.aaai.org/index.php/AAAI/article/view/34067.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. UniTime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM Web Conference 2024*, 2024a. URL https://github.com/liuxu77/UniTime.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2310.06625.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.14730.

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

Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs Bergmann, and Roland Vollgraf. Multivariate probabilistic time series forecasting via conditioned normalizing flows, 2021b. URL https://arxiv.org/abs/2002.06103.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31016–31029, 2023. URL https://proceedings.mlr.press/v202/shen23d.html.

Tianen Shen, Zhengyu Li, Yutong Li, Xiangfei Qiu, Xingjian Wu, Bin Yang, and Jilin Hu. Enhancing irregular time series forecasting with continuous-time modeling framework. *arXiv preprint arXiv:2607.28035*, 2026. URL https://arxiv.org/abs/2607.28035.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2107.03502.

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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02186.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting, 2026. URL https://arxiv.org/abs/2509.22295.

Huibo Xu, Runlong Yu, Likang Wu, Xianquan Wang, and Qi Liu. Bridging the last mile of prediction: Enhancing time series forecasting with conditional guided flow matching. *arXiv preprint arXiv:2507.07192*, 2025. URL https://arxiv.org/abs/2507.07192.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. 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/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2205.13504.

Tianlong Zhao, Guangle Song, Xuemei Li, Lizhen Cui, and Caiming Zhang. Diff-mgr: Dynamic causal graph attention and pattern reproduction guided diffusion model for multivariate time series probabilistic forecasting. *Information Sciences*, 675:120742, 2024. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2024.120742. URL https://www.sciencedirect.com/science/article/pii/S002002552400656X.

A VALIDATION OPTIMIZATION AND ERROR DECOMPOSITION

Three-member ensemble used by DiverseFlow. The formal point-prediction output of DiverseFlow is the three-member convex combination in Equation 12. The candidates are the external predictor g, the internal predictor f, and the Flow sample-mean predictor F. Let validation predictions be flattened into N scalar entries, and let $e_j \in \mathbb{R}^N$ denote the error vector of candidate $j \in \{g, f, F\}$. Define the weight vector $w = (w_g, w_f, w_F)^\top$ and the error Gram matrix G by

$$G_{ij} = \frac{1}{N} e_i^\top e_j. \tag{13}$$

Because the weights sum to one, the ensemble residual is $e(w) = w_g e_g + w_f e_f + w_F e_F$. Fitting the weights on the validation set is therefore the three-variable constrained quadratic program

$$\min_{w \in \mathbb{R}^3} w^\top G w \quad \text{subject to} \quad w \geq 0, \quad \mathbf{1}^\top w = 1. \tag{14}$$

Equivalently, if $A_j = \|e_j\|_2^2/N$ and $C_{ij} = e_i^\top e_j/N$, the objective is

$$J(w) = w_g^2 A_g + w_f^2 A_f + w_F^2 A_F + 2w_g w_f C_{gf} + 2w_g w_F C_{gF} + 2w_f w_F C_{fF}. \tag{15}$$

The matrix G is positive semidefinite, so this is a convex problem. For an invertible G, the equality-constrained interior stationary candidate is

$$w = \frac{G^{-1}\mathbf{1}}{\mathbf{1}^\top G^{-1}\mathbf{1}}, \tag{16}$$

provided all three components are nonnegative. If an interior component is negative, the Karush–Kuhn–Tucker solution lies on the boundary of the simplex. Each of the three edges sets one member's weight to zero and reduces to a two-member problem, while each vertex corresponds to a single predictor. We compare the feasible interior candidate with all edge and vertex candidates and select the one with the smallest validation objective. Singular cases can have multiple minimizers; the defining constrained objective, rather than the inverse formula, determines the solution. In particular, the three-member feasible set contains the two-member set with $w_F = 0$, so adding the Flow candidate cannot worsen the optimized validation MSE, although this validation-set property does not guarantee an improvement on the test set.

The three-member error decomposition makes the role of pairwise complementarity explicit. Define the pairwise diversity term

$$D_{ij} = A_i + A_j - 2C_{ij} = \frac{1}{N}\|e_i - e_j\|_2^2, \qquad i, j \in \{g, f, F\}. \tag{17}$$

Substituting this definition into Equation 15 and using $w_g + w_f + w_F = 1$ gives

$$J(w) = w_g A_g + w_f A_f + w_F A_F - w_g w_f D_{gf} - w_g w_F D_{gF} - w_f w_F D_{fF}. \tag{18}$$

Thus, the three-member ensemble is determined jointly by the individual errors and all three pairwise error relationships. The Flow output is useful only when its accuracy and its relationships with the two point predictors make the resulting validation objective smaller; stochastic variation alone is not sufficient.

Two-member ablation and supplementary weights. For the two-member comparison used in the ablation, write $d = \widehat{Y}_g - \widehat{Y}_f$ and minimize

$$\frac{1}{N}\|\widehat{Y}_f + \alpha d - Y\|_2^2, \qquad 0 \leq \alpha \leq 1. \tag{19}$$

The derivative gives the unconstrained ratio for this two-member objective; projection onto $[0, 1]$ gives the constrained minimizer. When $d = 0$, every weight produces the same forecast, and we use 0.5.

For completeness, the two-member diversity identity follows from expanding the squared difference:

$$D_{g,f} = A_g + A_f - 2C_{gf}. \tag{20}$$

This is the two-member special case of Equation 18 and is included because the ablation and supplementary weight table compare the external and internal predictors without Flow. Both decompositions are algebraic identities, not assumptions that adding noise or a member improves a predictor on unseen data.

B SUPPLEMENTARY TWO-MEMBER VALIDATION WEIGHTS

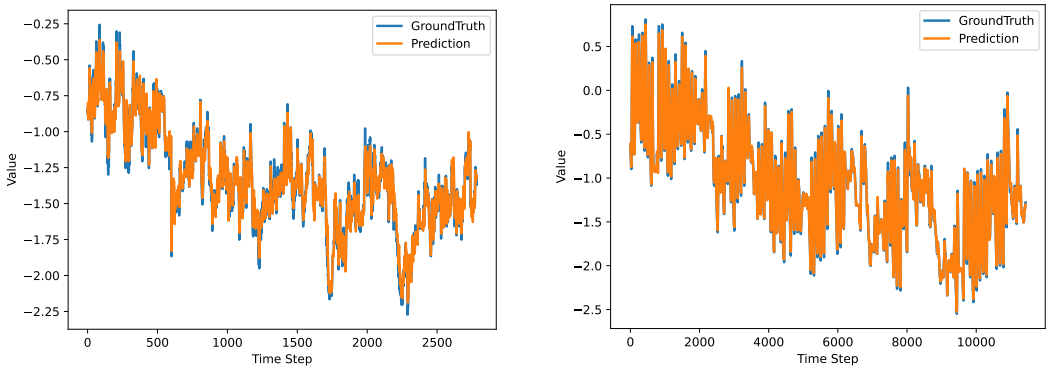
Table 5 records the external-model weight α for a supplementary two-member configuration. The internal weight is $1-\alpha$. These scalar weights describe deterministic two-member combinations and are not the three-member weights in Equation 12.

Table 5: External-model weights in a supplementary two-member ensemble, by random seed.

Dataset	Seed 2	Seed 2021	Seed 2023
Electricity	1.0000	1.0000	1.0000
ETTh1	0.3394	0.3801	0.3797
ETTh2	0.4738	0.4403	0.4595
ETTm1	0.5763	0.5935	0.5516
ETTm2	0.6040	0.6097	0.6459
Exchange	0.1532	0.1556	0.2690
Traffic	0.9849	0.9931	0.9848
Weather	0.7948	0.7887	0.8268

C VISUAL ANALYSIS OF MODEL PREDICTIONS

To further assess the point-forecasting behavior of DiverseFlow, Fig. 4 compares its predictions with the ground truth on the ETTh1 and ETTm2 datasets. Across these long evaluation sequences, the forecasts closely track the temporal evolution of the observations and remain responsive to local variations as the dynamics vary. The final prediction combines two independently trained deterministic predictors that use different historical contexts with the Flow sample mean through validation-fitted convex weights. The qualitative agreement in Fig. 4 is thus consistent with the intended roles of these complementary signals: the two deterministic branches offer predictions from different temporal perspectives, while the Flow output contributes an additional residual-based forecasting candidate.



(a) ETTh1 (b) ETTm2

Figure 4: Full prediction results on ETTh1 and ETTm2.

D COMPLETE FOUR-HORIZON POINT FORECASTING RESULTS

Table 6 reports the complete results for all eight datasets, four horizons, and their reported averages. Dataset and horizon form the row index, while each forecasting method occupies adjacent MSE and MAE columns. The displayed precision follows the reported results; ties are indicated by boldface. The main comparison uses the averages in Table 1, while the matched component comparison at horizon 96 is reported separately in Table 4.

Table 6: Complete point forecasting results on eight datasets and four forecast horizons. Each method has separate MSE and MAE columns; lower is better. The best and second-best results are highlighted in red bold and blue underline, respectively. Avg denotes the arithmetic mean over the four forecast horizons. 1st Count reports the number of best entries, including Avg rows and ties; higher is better for 1st Count.

Model	Horizon	DiverseFlow (Ours)		T3Time (2026)		TimeCMA (2025)		Time-LLM (2024)		UniTime (2024a)		TimesNet (2023)		DLinear (2023)		iTransformer (2024b)		PatchTST (2023)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.366**	0.393	0.371	0.397	0.373	**0.391**	0.398	0.410	0.397	0.418	0.384	0.402	0.386	0.400	0.386	0.405	0.404	0.413
	192	**0.410**	**0.420**	0.411	0.421	0.427	0.421	0.451	0.440	0.434	0.439	0.434	0.429	0.437	0.432	0.441	0.436	0.454	0.430
	336	**0.445**	0.448	0.447	**0.441**	0.458	0.448	0.473	0.451	0.468	0.457	0.491	0.469	0.481	0.459	0.487	0.458	0.497	0.462
	720	0.456	0.468	**0.441**	**0.460**	0.449	0.460	0.469	0.470	0.469	0.477	0.521	0.500	0.519	0.516	0.503	0.491	0.496	0.481
	Avg	0.419	0.432	**0.418**	**0.430**	0.427	0.430	0.448	0.443	0.442	0.448	0.458	0.450	0.456	0.452	0.454	0.448	0.463	0.447
ETTh2	96	**0.274**	0.339	0.278	0.338	0.286	**0.336**	0.295	0.345	0.296	0.345	0.340	0.374	0.333	0.387	0.297	0.349	0.312	0.358
	192	0.354	**0.386**	**0.351**	0.389	0.363	0.387	0.386	0.399	0.374	0.394	0.402	0.414	0.477	0.476	0.380	0.400	0.397	0.408
	336	0.366	0.401	**0.358**	**0.398**	0.406	0.421	0.419	0.429	0.415	0.427	0.452	0.452	0.594	0.541	0.428	0.432	0.435	0.440
	720	0.414	0.441	**0.404**	**0.433**	0.417	0.438	0.425	0.442	0.425	0.444	0.462	0.468	0.831	0.657	0.427	0.445	0.436	0.449
	Avg	0.351	0.392	**0.348**	**0.390**	0.368	0.396	0.381	0.404	0.378	0.403	0.414	0.427	0.559	0.515	0.383	0.407	0.395	0.414
ETTm1	96	**0.282**	**0.340**	0.308	0.354	0.312	0.351	0.359	0.381	0.322	0.363	0.338	0.375	0.345	0.372	0.334	0.368	0.344	0.373
	192	**0.325**	**0.366**	0.357	0.381	0.361	0.378	0.383	0.393	0.366	0.387	0.374	0.387	0.380	0.389	0.377	0.391	0.367	0.386
	336	**0.353**	**0.386**	0.382	0.400	0.392	0.401	0.416	0.414	0.398	0.407	0.410	0.411	0.413	0.413	0.426	0.420	0.392	0.407
	720	**0.407**	**0.418**	0.442	0.437	0.453	0.438	0.483	0.449	0.454	0.440	0.478	0.450	0.474	0.453	0.491	0.459	0.464	0.442
	Avg	**0.341**	**0.377**	0.372	0.393	0.380	0.392	0.410	0.409	0.385	0.399	0.400	0.406	0.403	0.407	0.407	0.410	0.392	0.402
ETTm2	96	**0.158**	**0.250**	0.172	0.254	0.173	0.258	0.193	0.280	0.183	0.266	0.187	0.267	0.193	0.292	0.180	0.264	0.177	0.260
	192	**0.216**	**0.291**	0.237	0.300	0.238	0.301	0.257	0.318	0.251	0.310	0.249	0.309	0.284	0.362	0.250	0.309	0.246	0.305
	336	**0.271**	**0.327**	0.306	0.337	0.297	0.338	0.317	0.353	0.312	0.364	0.319	0.351	0.369	0.427	0.311	0.348	0.305	0.343
	720	**0.358**	**0.383**	0.400	0.398	0.393	0.394	0.419	0.411	0.420	0.410	0.408	0.403	0.554	0.522	0.412	0.407	0.410	0.405
	Avg	**0.250**	**0.312**	0.279	0.322	0.275	0.323	0.297	0.341	0.292	0.338	0.291	0.333	0.350	0.401	0.288	0.332	0.285	0.328
ECL	96	**0.137**	0.242	0.138	**0.233**	0.143	0.238	0.172	0.265	0.196	0.287	0.168	0.272	0.197	0.282	0.148	0.240	0.186	0.269
	192	**0.154**	0.256	0.155	**0.250**	0.161	0.259	0.182	0.279	0.199	0.291	0.184	0.289	0.196	0.285	0.162	0.253	0.190	0.273
	336	0.169	0.271	**0.168**	0.265	0.169	**0.261**	0.195	0.288	0.214	0.305	0.198	0.300	0.209	0.301	0.178	0.269	0.206	0.290
	720	**0.195**	**0.293**	0.218	0.314	0.219	0.315	0.233	0.320	0.254	0.335	0.220	0.320	0.245	0.333	0.225	0.317	0.247	0.322
	Avg	**0.164**	**0.265**	0.170	**0.266**	0.173	0.268	0.196	0.288	0.216	0.305	0.193	0.295	0.212	0.300	0.178	0.270	0.207	0.289
Traffic	96	**0.380**	0.272	0.428	0.271	0.395	**0.268**	0.649	0.389	0.462	0.295	0.522	0.290	0.805	0.493	0.593	0.321	0.650	0.396
	192	**0.393**	0.278	0.448	0.282	0.417	**0.276**	0.601	0.366	0.466	0.296	0.530	0.293	0.756	0.474	0.617	0.336	0.598	0.370
	336	**0.408**	0.284	0.473	0.289	0.433	**0.283**	0.609	0.369	0.482	0.304	0.558	0.305	0.762	0.477	0.629	0.336	0.605	0.373
	720	**0.445**	0.303	0.516	0.307	0.467	**0.302**	0.647	0.387	0.514	0.322	0.589	0.328	0.719	0.449	0.640	0.350	0.645	0.394
	Avg	**0.406**	0.283	0.466	0.287	0.428	**0.282**	0.627	0.378	0.481	0.304	0.550	0.304	0.761	0.473	0.620	0.336	0.625	0.383
Weather	96	**0.150**	**0.203**	0.162	0.210	0.167	0.211	0.198	0.235	0.171	0.214	0.172	0.220	0.196	0.255	0.174	0.214	0.177	0.218
	192	**0.189**	**0.240**	0.211	0.253	0.212	0.253	0.240	0.269	0.217	0.254	0.219	0.261	0.237	0.296	0.221	0.254	0.222	0.259
	336	**0.236**	**0.279**	0.267	0.293	0.270	0.292	0.295	0.308	0.274	0.293	0.280	0.306	0.283	0.335	0.278	0.296	0.277	0.297
	720	**0.291**	**0.325**	0.335	0.346	0.350	0.348	0.368	0.353	0.351	0.343	0.365	0.359	0.345	0.381	0.358	0.349	0.352	0.347
	Avg	**0.216**	**0.261**	0.244	0.276	0.250	0.276	0.275	0.291	0.253	0.276	0.259	0.287	0.265	0.317	0.258	0.278	0.257	0.280
Exchange	96	**0.084**	**0.202**	0.085	0.205	0.099	0.224	0.087	0.208	0.086	0.209	0.107	0.234	0.088	0.218	0.086	0.206	0.109	0.236
	192	0.177	0.298	**0.172**	**0.296**	0.186	0.312	0.173	0.299	0.174	0.299	0.226	0.344	0.176	0.315	0.177	0.299	0.205	0.327
	336	0.326	0.412	0.318	**0.408**	0.364	0.444	0.375	0.454	0.319	0.408	0.367	0.448	**0.313**	0.427	0.331	0.417	0.356	0.436
	720	0.841	0.841	**0.836**	0.696	0.932	0.735	0.853	0.703	0.875	0.701	0.964	0.746	0.839	0.695	0.847	**0.691**	0.888	0.716
	Avg	0.357	0.438	**0.353**	**0.401**	0.395	0.429	0.372	0.416	0.364	0.404	0.416	0.443	0.354	0.414	0.360	0.403	0.390	0.429
1st Count		**29**	**20**	13	13	0	11	0	0	0	1	0	0	1	0	0	1	0	0