UtoMe: Observation-Uncertainty-Guided Token Merging for Weather Foundation Models

Yunbei Zhang[1,2*†] **Janet Wang**[1,2*] **Xi Xiao**[2] **Jihun Hamm**[1] **Xiao Wang**[2†]
[1]Tulane University, [2]Oak Ridge National Laboratory

Abstract

Weather transformers process dense reanalysis fields whose reliability varies across variables, locations, and weather regimes. Token merging can reduce computation, but feature similarity alone does not account for this heterogeneous input uncertainty. We introduce UtoMe, an observation-uncertainty-guided token merging framework that combines dynamics saliency with ensemble data assimilation spread as a relative input-reliability proxy. The resulting importance score protects selected tokens as anchors and calibrates bipartite matching between the remaining tokens. Three variants use external uncertainty at both training and inference, during training only, or neither stage. In ERA5 72-hour forecasting at two spatial resolutions, UtoMe improves Overall error over train-time ToMe at matched token budgets. At $1.40625°$ and 50% token reduction, UtoMe-E lowers Overall error from 67.32 to 53.00 at the same $0.55\times$ encoder FLOPs. Small reductions remain close to the full-token baseline; larger reductions incur measurable accuracy losses. Component ablations and spatial uncertainty controls support combining dynamics with aligned uncertainty when allocating a weather transformer's token budget. Code and project page: **https://ornl-utome.github.io/**.

1 Introduction

Data-driven weather models learn forecast operators from reanalysis archives and provide efficient global forecasts [25, 1, 17, 23]. Training remains costly for transformer forecasters, which repeatedly process spatial tokens through attention and MLP layers. Uniformly increasing patch size reduces this cost but coarsens every location equally. Adaptive token reduction offers a more selective alternative: it can compress redundant regions while retaining representations of structures relevant to the forecast.

Token pruning and Token Merging (ToMe) reduce sequence length using attention, feature similarity, or learned importance [26, 4, 3]. Weather forecasting introduces an additional consideration: reanalysis inputs estimate a latent atmospheric state using heterogeneous observations and data assimilation [11, 13]. Their uncertainty is spatially structured. Similar hidden features can therefore reflect either reliable atmospheric redundancy or uncertainty in the input estimate, which a similarity-only rule does not distinguish.

We study input reliability, assessed through uncertainty quantification (UQ), as a compute-allocation signal. A latent-input predictive model makes explicit an observation-induced variance component alongside aleatoric variability and epistemic model uncertainty [14, 12, 21]. This motivates considering input uncertainty when allocating tokens, while leaving the allocation rule to empirical evaluation. High uncertainty can coincide with important weather dynamics; uncertainty must therefore be considered jointly with a saliency signal.

*Equal contribution. †Corresponding authors. Work done during Yunbei Zhang's internship at ORNL. Emails: yzhang111@tulane.edu, wangx2@ornl.gov.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

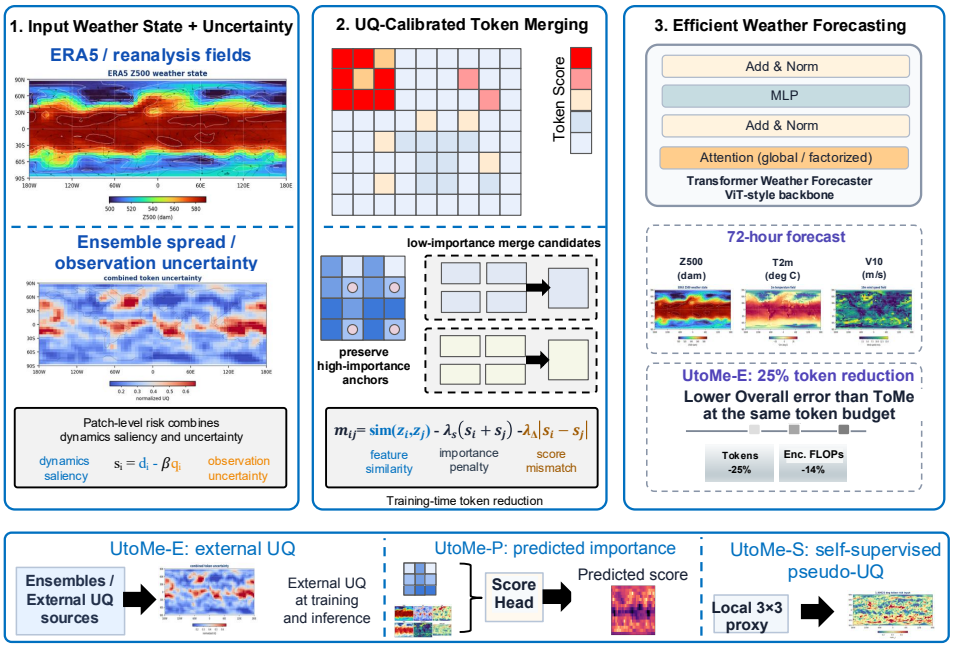


Figure 1: **UtoMe overview.** UtoMe uses observation uncertainty to guide training-time token merging for transformer-based weather forecasting. It first turns weather-state saliency and low observation uncertainty into a patch-level token score, then uses that score to protect forecast-sensitive anchors and merge lower-importance redundant tokens. The three variants use external uncertainty at both stages (E), during training only (P), or neither stage (S, using a 3×3 local spatial-inconsistency proxy). The illustrated 25% token reduction and 14% encoder FLOPs saving refer to UtoMe-E relative to no merging.

UtoMe combines patch-level dynamics saliency with normalized ensemble spread to form an importance score (Fig. 1). It protects high-scoring anchors and augments ToMe-style bipartite matching with importance penalties. Merging is active during both training and evaluation. The contribution is a weather-specific allocation policy that retains ToMe's merge operator and accounts for heterogeneous input reliability.

The three variants address distinct uncertainty-availability regimes. UtoMe-E uses external uncertainty at both stages; UtoMe-P learns an externally supervised importance predictor for deployment without uncertainty fields; and UtoMe-S learns from a local spatial-inconsistency proxy without external uncertainty. We make three contributions:

1. We formulate observation-uncertainty-guided token allocation for weather transformers, using ensemble spread as a relative input-reliability proxy.
2. We combine dynamics–uncertainty scoring, anchor protection, and importance-calibrated matching, and isolate their contributions with controlled ablations.
3. We evaluate three uncertainty-availability regimes on ERA5 72-hour forecasting at $5.625°$ and $1.40625°$, including spatial uncertainty controls, difficult-regime diagnostics, pruning comparisons, and a second transformer backbone.

2 Related Work

Weather foundation models. Data-driven weather models learn forecast operators from gridded atmospheric archives such as ERA5 [11, 22]. FourCastNet uses adaptive Fourier neural operators for high-resolution global forecasting [25]; Pangu-Weather uses a 3D Earth-specific architecture for medium-range forecasting [1]; GraphCast uses graph neural networks on the sphere [17]; and ClimaX studies a transformer foundation model that can be pretrained across heterogeneous weather

and climate datasets [23]. Aurora extends foundation modeling to diverse Earth-system forecasting tasks [2]. Recent efforts continue to scale these models and related Earth-system foundation models [32, 33]. These works improve forecast skill and model generality at a fixed tokenization. UtoMe is complementary: it asks how to reduce the token sequence during training while preserving the forecast information that matters.

Token pruning and merging. Computational efficiency is a central concern in transformer-based learning and adaptation [29, 18]. Token pruning, routing, pooling, and merging address this concern by reducing sequence length. DynamicViT learns to remove low-importance tokens progressively [26]. ToMe merges similar tokens with a lightweight bipartite matching algorithm and can be applied to pretrained ViTs, with extensions to diffusion models and other modalities [4, 3]. Recent work also explores token routing and sparse computation for diffusion models [35, 15, 27, 24]. Co-Me uses a confidence predictor to guide token merging for visual geometric transformers [6]. Sparse-Reslim brings sparse token routing to weather forecasting: it processes a sampled token subset in intermediate transformer blocks and scatters residual updates back to the full sequence, leaving unselected representations intact [31]. UtoMe uses dynamics saliency and observation uncertainty from the data source to guide anchor protection and pairwise merging during weather-model training.

Uncertainty quantification and observation error. Deep uncertainty quantification is often organized around aleatoric and epistemic uncertainty [14, 12], with practical estimators based on MC dropout, deep ensembles, evidential models, and deterministic distance-aware networks [9, 16, 28, 19]. This two-way split usually assumes the input is observed without error. Errors-in-variables models, input-noise propagation, and unified uncertainty formulations make this assumption explicit and add an input or observation uncertainty term [8, 5, 21, 10, 20, 30, 34]. In geoscience, observation error and representativeness error are core parts of data assimilation [13, 11]. UtoMe brings these ideas into the architecture: observation uncertainty is used not only as a diagnostic quantity, but as a token-level signal for allocating computation.

3 Method

3.1 Forecasting Setup

Let $x_t \in \mathbb{R}^{C\times H\times W}$ denote a multivariate atmospheric state at time t, where C includes pressure-level, surface, and static variables. Given lead time Δ, the model predicts $y_{t+\Delta}$. For a ViT-style weather transformer, the grid is split into non-overlapping patches $\{\mathcal{P}_i\}_{i=1}^{N}$ and embedded into tokens $z^0 \in \mathbb{R}^{N\times d}$. A standard transformer processes all N tokens in every layer. Token merging reduces the sequence by combining selected source tokens into destination tokens. The merge policy only assumes this patch-token interface, so it is not tied to a particular weather-transformer implementation.

The main design question is the merge policy. A similarity-only method merges tokens whose hidden features are close. This is attractive because it is cheap and batchable, but it ignores whether the apparent similarity comes from reliable atmospheric redundancy or from noisy observations. UtoMe augments the merge policy with an uncertainty-guided importance score computed before merging.

3.2 Observation Uncertainty as Analytical Motivation

We model the input as a noisy observation of a latent true atmospheric state,

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}^* + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim p_\varepsilon(\cdot \mid \boldsymbol{x}^*) \tag{1}$$

For training data $\mathcal{D}$, we assume the latent-input and parameter posteriors factorize as $p(\boldsymbol{x}^*, \boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \mathcal{D}) = p(\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}})p(\boldsymbol{\theta} \mid \mathcal{D})$. With finite conditional second moments, the predictive distribution is

$$p(y \mid \tilde{\boldsymbol{x}}, \mathcal{D}) = \int\int p(y \mid \boldsymbol{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{x}^*\, d\boldsymbol{\theta}. \tag{2}$$

The latent-input formulation motivates accounting for observation reliability when allocating a token budget. If one sets $\boldsymbol{x}^* = \tilde{\boldsymbol{x}}$, Eq. (2) reduces to the standard setting and predictive variance splits into aleatoric and epistemic terms [14, 7]. Keeping the latent state explicit and applying the law of total variance twice gives

$$\mathrm{Var}[y \mid \tilde{\boldsymbol{x}}, \mathcal{D}] = \mathbb{E}_{\boldsymbol{x}^*}\mathbb{E}_{\boldsymbol{\theta}}[\mathrm{Var}(y \mid \boldsymbol{x}^*, \boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{x}^*}\mathrm{Var}_{\boldsymbol{\theta}}(\mathbb{E}[y \mid \boldsymbol{x}^*, \boldsymbol{\theta}]) + \mathrm{Var}_{\boldsymbol{x}^*}(\mathbb{E}[y \mid \boldsymbol{x}^*])\,. \tag{3}$$

The first two terms describe aleatoric variability and epistemic model uncertainty, averaged over possible latent states. The third term is observation uncertainty: it measures how much the forecast mean changes because several latent states could have produced the same observed input. Appendix A gives the derivation and the first-order approximation that connects this term to local input sensitivity and latent-input posterior covariance.

The decomposition motivates including input reliability in token allocation; it does not imply an optimal compression policy. We use ensemble spread as a relative proxy for uncertainty in the input estimate and combine it with dynamics saliency. The subtraction sign, anchor rule, and matching penalties are evaluated through ablations rather than derived from the variance identity.

3.3 Patch-Level Importance from Dynamics and Observation Spread

For variables with uncertainty information, we use an aligned field $u_t \in \mathbb{R}^{C\times H\times W}$ from ECMWF's Ensemble of Data Assimilations (EDA). Its spread is a relative estimate of reanalysis uncertainty, including observation and assimilation-model effects; it is not an isolated measurement of observation-error variance.[1] We refer to this relative input-reliability signal as observation uncertainty. Low spread strengthens protection when a patch is dynamically salient; a patch with high spread can also remain important when its dynamics signal is large. In our reported runs, the UQ tensor follows the input-channel layout, with static variables assigned zero uncertainty. We normalize each variable and patch-average the field:

$$q_i = \frac{1}{|\mathcal{P}_i|\,C}\sum_{c=1}^{C}\sum_{(h,w)\in\mathcal{P}_i}\hat{u}_{t,c,h,w}, \tag{4}$$

where $\hat{u}$ denotes the normalized spread and $\mathcal{P}_i$ is the spatial support of token i. We then standardize the patch scores per sample and keep the notation q_i for the normalized patch-level observation-uncertainty score.

Dynamics saliency combines spatial variation and anomaly amplitude in the normalized input-time fields of the five forecast variables. For their channel set $\mathcal{C}_{\mathrm{dyn}}$, define

$$D_{h,w} = \frac{1}{|\mathcal{C}_{\mathrm{dyn}}|}\sum_{c\in\mathcal{C}_{\mathrm{dyn}}}\left(|\delta_h x_{t,c,h,w}| + |\delta_w x_{t,c,h,w}| + \tfrac{1}{2}|x_{t,c,h,w} - \bar{x}_{t,c}|\right), \tag{5}$$

where δ_h and δ_w are adjacent-grid differences and $\bar{x}_{t,c}$ is the spatial mean of input channel c. Patch averaging followed by per-sample standardization gives d_i. This construction emphasizes sharp gradients and anomalies using information available at the input time. The final importance score is

$$s_i = d_i - \beta q_i, \tag{6}$$

with weight β. We use $\beta = 1$, selected on validation data: dynamics saliency increases merge protection, while observation spread subtracts from the score. Thus, high-dynamics low-spread patches receive the strongest protection, and high-spread patches become better merge candidates unless the dynamics signal is strong enough to dominate. Fig. 3 illustrates the spatial structure retained by patch aggregation.

3.4 Anchored, Uncertainty-Calibrated Token Merging

UtoMe keeps the efficiency and batchability of ToMe-style bipartite matching, but changes the merge decision so that it respects the importance score in Eq. (6). Let $\mathcal{A}$ contain the top ρ fraction of tokens according to s_i, with $\rho = 0.25$ by default. Tokens in $\mathcal{A}$ are treated as anchors: they retain separate representations and are excluded from both merge-source and merge-destination pools. The anchor fraction is applied when merging begins, and the protection mask is carried through subsequent layers.

After anchors are removed from both candidate pools, UtoMe applies an uncertainty-calibrated matching score to the remaining tokens:

$$m_{ij} = \mathrm{sim}(z_i, z_j) - \lambda_s(s_i + s_j) - \lambda_\Delta|s_i - s_j|. \tag{7}$$

[1] ECMWF/Copernicus ERA5 uncertainty estimation documentation: https://confluence.ecmwf.int/display/CKB/ERA5:+uncertainty+estimation.

We use $\lambda_s = \lambda_\Delta = 0.05$, with the selection protocol and sensitivity analysis in Appendix C. Here z_i and z_j are hidden tokens, sim is cosine similarity in our implementation, and $\lambda_s, \lambda_\Delta \geq 0$ control the strength of the importance penalties. The similarity term keeps the original ToMe criterion: if two tokens encode the same local atmospheric information, they remain good merge candidates. The sum penalty discourages high-importance pairs, while the difference penalty discourages pairs with dissimilar joint importance. For each source token i, the destination is selected as

$$j^*(i) = \arg\max_{j \in \mathcal{D}_{\mathrm{dst}}} m_{ij}, \tag{8}$$

where $\mathcal{D}_{\mathrm{dst}}$ is the destination set in the bipartite partition. The top-scoring edges are merged until the layer budget is reached, and merged token features are averaged with size tracking as in ToMe.

3.5 Variants for Different UQ Availability Regimes

UtoMe-E: external UQ at training and inference. UtoMe-E uses the EDA-derived uncertainty field in Eq. (4) at both stages. This is the cleanest setting for reanalysis workflows where ensemble spread is available alongside the input fields.

UtoMe-P: external UQ during training only. UtoMe-P handles the case where uncertainty fields are available during training but not at deployment. We add a lightweight predictor g_ϕ that maps early hidden tokens to scores $\hat{s}_i$. During training, g_ϕ is supervised by the EDA-derived score s_i with a pairwise-ranking loss, since the merge policy depends primarily on the ordering used for anchor selection and edge penalties. Ranking supervision targets anchor selection; the matching penalties additionally depend on score magnitudes. At inference, $\hat{s}_i$ replaces s_i in anchor selection and matching.

UtoMe-S: no external UQ. UtoMe-S removes the dependency on ensemble products. We build a pseudo-uncertainty target from local spatial inconsistency,

$$\tilde{q}_{h,w} = \frac{1}{|\mathcal{C}_{\mathrm{dyn}}|} \sum_{c \in \mathcal{C}_{\mathrm{dyn}}} \left| x_{t,c,h,w} - \mathrm{AvgPool}_{3\times 3}(x_{t,c})_{h,w} \right| \tag{9}$$

using the same input-time variable set as dynamics saliency. We patch-average and standardize this proxy, substitute it for q_i in Eq. (6), and train the same predictor used by UtoMe-P on the resulting importance score. The proxy measures local inconsistency and is not a calibrated observation-error estimate.

4 Experiments

4.1 Experimental Setup

Data and task. We evaluate global ERA5 deterministic forecasting at two resolutions. The 5.625° setting uses 2010 to 2014 for training, 2015 for validation, and 2016 for testing, with a 32×64 grid and 2×2 patches, giving 512 initial spatial tokens. The 1.40625° setting uses 1979 to 2018 for training, 2019 for validation, and 2020 for testing, with a 128×256 grid and 4×4 patches, giving 2048 initial spatial tokens. These angles specify the latitude–longitude grid spacing. Each sample uses the atmospheric state x_t to predict selected targets at $t + 72$ hours. The input features use the ERA5 variable set from ClimaX and contain 41 channels covering static fields, surface variables, and pressure-level variables; the output features are Z500 (500 hPa geopotential, $\mathrm{m^2\,s^{-2}}$), T850 (850 hPa temperature, K), T2m (2 m temperature, K), and U10/V10 (zonal/meridional 10 m winds, $\mathrm{m\,s^{-1}}$). These targets cover mid-tropospheric circulation, lower-tropospheric temperature, near-surface temperature, and surface winds. EDA spread is aligned with the input time, grid, and variables where available. It is not concatenated to the forecast input; it supervises or calibrates token merging.

Model and baselines. The main comparisons use the same ViT-style transformer forecaster: eight transformer layers, hidden dimension 1024, 16 attention heads, MLP ratio 4, dropout 0.1, drop-path 0.1, and a two-layer decoder. The model embeds each selected variable with a one-channel patch embed, adds variable embeddings, aggregates variables into a spatial token sequence, and applies token merging on that spatial sequence. We compare no merging, train-time ToMe with the same layerwise token budget, and UtoMe-E/P/S. For all merging rows, merging is active during both

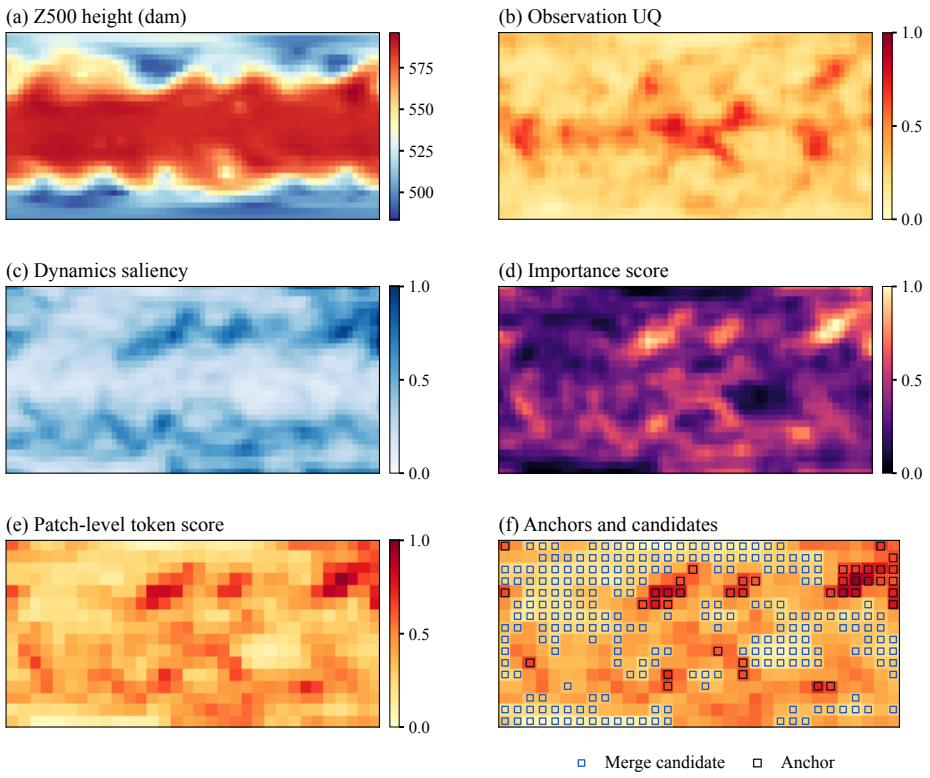


Figure 2: $5.625°$ **case summary.** The figure shows the complete merge-decision pipeline: weather state, observation uncertainty, dynamics saliency, importance score, patch-level token score, and the final anchor or merge-candidate overlay.

training and evaluation. E and ToMe share a schedule; P/S delay merging until after the predictor probe (Appendix B). The main tables report final token reductions of 25% and 50%.

Training and metrics. Training uses DDP on an HPC cluster of 8 nodes, each exposing 8 AMD MI250X GPU devices, for 64 GPU devices in total. We optimize with AdamW using a peak learning rate of 5×10^{-4}, weight decay 10^{-5}, 1000 warmup steps, gradient clipping at 1.0, and a 12000-step learning-rate schedule. The per-process batch size is 16 at $5.625°$ and 1 at $1.40625°$. Predictor-based variants use an auxiliary pairwise-ranking loss with weight 0.2. We report mean and standard deviation over three random seeds for the main forecast comparisons; the additional controls report point estimates. Metrics are latitude-weighted root mean square error (wRMSE) for each target variable, Overall error aggregated over the targets, and anomaly correlation coefficient (ACC) after climatology removal, following ClimaX [23]. Compute is the encoder FLOPs ratio relative to no merging at the same resolution. We use normalized FLOPs as the main efficiency metric because it is controlled by the token schedule and is comparable across resolutions. More details are in Appendix B.

Overall is the arithmetic mean of the five target-specific errors:

$$\text{Overall} = \frac{1}{5} \sum_{k \in \{\text{Z500,T850,T2m,U10,V10}\}} \text{wRMSE}_k. \tag{10}$$

Overall mixes physical units and is numerically dominated by Z500; it is not a scale-invariant measure of multivariate skill. We therefore interpret it alongside per-variable wRMSE and ACC.

4.2 Forecast Skill at Matched Token Budgets

Joint scoring improves the accuracy–compute trade-off. Tables 1 and 2 compare allocation policies at matched final token budgets. At $5.625°$ and 25% reduction, UtoMe-P obtains Overall error

Table 1: Main results on 5.625° ERA5 72-hour forecasting. The table compares no merging, train-time ToMe, and the three UtoMe variants under matched final token-reduction ratios. Values are mean and standard deviation over three random seeds. Lower is better for Overall and per-variable wRMSE, higher is better for ACC, and Compute is the encoder FLOPs ratio relative to no merging at the same resolution; the main resolution tables follow this convention. Green shades identify E/P/S uncertainty availability; blue bands identify token budgets. Bold/underlined values are best/second-best within each budget.

Method	Overall ↓	ACC ↑	Z500 ↓	T850 ↓	T2m ↓	U10 ↓	V10 ↓	Compute ↓
No merging	$143.07_{\pm 1.81}$	$0.651_{\pm 0.008}$	$701.64_{\pm 8.92}$	$3.108_{\pm 0.023}$	$2.502_{\pm 0.008}$	$3.944_{\pm 0.024}$	$4.151_{\pm 0.051}$	1.00×
25% token reduction								
ToMe	$160.82_{\pm 2.07}$	$0.562_{\pm 0.011}$	$789.00_{\pm 10.34}$	$3.544_{\pm 0.041}$	$3.378_{\pm 0.038}$	$\mathbf{4.041}_{\pm 0.047}$	$\mathbf{4.142}_{\pm 0.044}$	0.86×
UtoMe-E	$\underline{144.74}_{\pm 1.42}$	$\underline{0.640}_{\pm 0.007}$	$\underline{709.50}_{\pm 7.21}$	$\underline{3.220}_{\pm 0.026}$	$\underline{2.730}_{\pm 0.021}$	$4.060_{\pm 0.032}$	$4.190_{\pm 0.039}$	0.86×
UtoMe-P	$\mathbf{144.10}_{\pm 1.16}$	$\mathbf{0.650}_{\pm 0.006}$	$\mathbf{706.34}_{\pm 6.43}$	$\mathbf{3.210}_{\pm 0.024}$	$\mathbf{2.717}_{\pm 0.019}$	$\underline{4.054}_{\pm 0.030}$	$\underline{4.182}_{\pm 0.034}$	0.87×
UtoMe-S	$145.36_{\pm 1.69}$	$0.635_{\pm 0.009}$	$712.50_{\pm 8.18}$	$3.260_{\pm 0.031}$	$2.760_{\pm 0.024}$	$4.075_{\pm 0.036}$	$4.205_{\pm 0.041}$	0.87×
50% token reduction								
ToMe	$220.05_{\pm 3.84}$	$0.423_{\pm 0.014}$	$1080.19_{\pm 18.76}$	$5.189_{\pm 0.092}$	$5.822_{\pm 0.117}$	$4.586_{\pm 0.066}$	$4.464_{\pm 0.058}$	0.55×
UtoMe-E	$\mathbf{151.95}_{\pm 1.58}$	$\mathbf{0.605}_{\pm 0.010}$	$\mathbf{744.41}_{\pm 7.82}$	$\underline{3.466}_{\pm 0.034}$	$\underline{3.485}_{\pm 0.036}$	$\underline{4.125}_{\pm 0.039}$	$\underline{4.255}_{\pm 0.042}$	0.55×
UtoMe-P	$\underline{152.42}_{\pm 1.47}$	$\underline{0.580}_{\pm 0.012}$	$\underline{747.19}_{\pm 7.31}$	$\mathbf{3.412}_{\pm 0.029}$	$\mathbf{3.481}_{\pm 0.033}$	$\mathbf{3.963}_{\pm 0.028}$	$\mathbf{4.045}_{\pm 0.035}$	0.57×
UtoMe-S	$167.63_{\pm 2.26}$	$0.502_{\pm 0.015}$	$821.37_{\pm 11.94}$	$3.732_{\pm 0.047}$	$4.507_{\pm 0.071}$	$4.196_{\pm 0.044}$	$4.335_{\pm 0.049}$	0.57×

Table 2: Main results on 1.40625° ERA5 72-hour forecasting. Values are mean and standard deviation over three seeds; conventions follow Table 1.

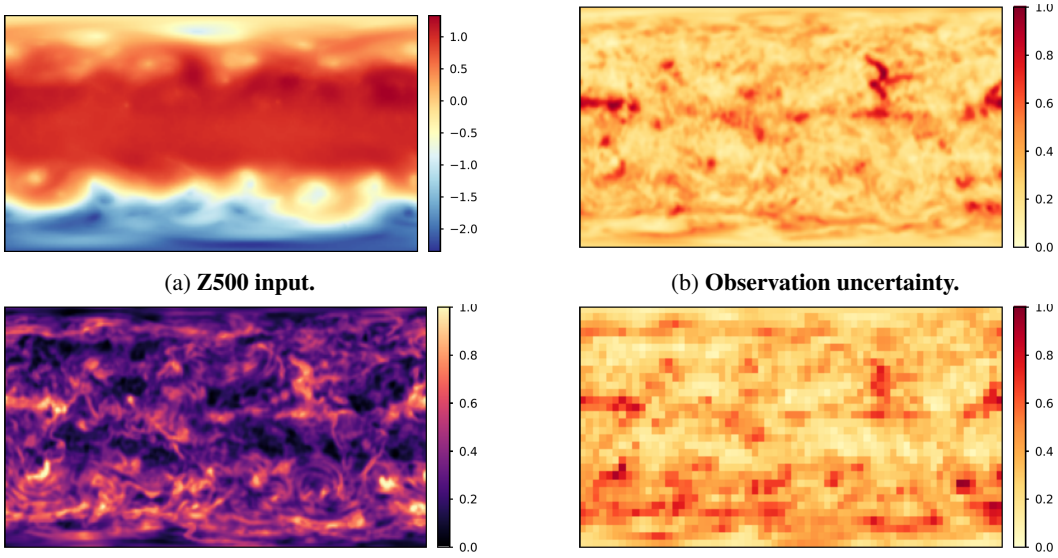
Method	Overall ↓	ACC ↑	Z500 ↓	T850 ↓	T2m ↓	U10 ↓	V10 ↓	Compute ↓
No merging	$32.41_{\pm 0.28}$	$0.964_{\pm 0.004}$	$156.76_{\pm 1.34}$	$1.299_{\pm 0.007}$	$1.269_{\pm 0.006}$	$1.339_{\pm 0.008}$	$1.360_{\pm 0.009}$	1.00×
25% token reduction								
ToMe	$42.45_{\pm 0.24}$	$0.949_{\pm 0.003}$	$206.02_{\pm 1.17}$	$1.452_{\pm 0.006}$	$1.483_{\pm 0.007}$	$1.572_{\pm 0.008}$	$1.724_{\pm 0.009}$	0.86×
UtoMe-E	$\mathbf{39.59}_{\pm 0.21}$	$\mathbf{0.952}_{\pm 0.003}$	$\mathbf{191.93}_{\pm 1.05}$	$\mathbf{1.413}_{\pm 0.005}$	$\mathbf{1.454}_{\pm 0.006}$	$\mathbf{1.527}_{\pm 0.007}$	$\mathbf{1.645}_{\pm 0.008}$	0.86×
UtoMe-P	$\underline{39.61}_{\pm 0.01}$	$\mathbf{0.952}_{\pm 0.000}$	$\underline{191.98}_{\pm 0.03}$	$1.425_{\pm 0.001}$	$1.459_{\pm 0.000}$	$\underline{1.530}_{\pm 0.000}$	$\underline{1.648}_{\pm 0.001}$	0.87×
UtoMe-S	$39.84_{\pm 0.27}$	$\underline{0.951}_{\pm 0.004}$	$193.14_{\pm 1.31}$	$\underline{1.416}_{\pm 0.007}$	$\underline{1.458}_{\pm 0.008}$	$1.533_{\pm 0.009}$	$1.656_{\pm 0.010}$	0.87×
50% token reduction								
ToMe	$67.32_{\pm 0.43}$	$0.905_{\pm 0.005}$	$328.21_{\pm 2.11}$	$1.846_{\pm 0.011}$	$2.056_{\pm 0.015}$	$2.136_{\pm 0.014}$	$\underline{2.347}_{\pm 0.018}$	0.55×
UtoMe-E	$\mathbf{53.00}_{\pm 0.36}$	$\mathbf{0.913}_{\pm 0.004}$	$\mathbf{256.72}_{\pm 1.76}$	$\underline{1.759}_{\pm 0.009}$	$\underline{2.051}_{\pm 0.012}$	$\underline{2.038}_{\pm 0.013}$	$2.433_{\pm 0.015}$	0.55×
UtoMe-P	$\underline{53.14}_{\pm 0.01}$	$0.907_{\pm 0.000}$	$\underline{257.37}_{\pm 0.05}$	$1.762_{\pm 0.001}$	$2.056_{\pm 0.001}$	$2.045_{\pm 0.001}$	$2.444_{\pm 0.001}$	0.57×
UtoMe-S	$57.01_{\pm 0.04}$	$\underline{0.908}_{\pm 0.000}$	$276.99_{\pm 0.18}$	$\mathbf{1.752}_{\pm 0.001}$	$\mathbf{1.968}_{\pm 0.004}$	$\mathbf{2.009}_{\pm 0.002}$	$\mathbf{2.335}_{\pm 0.002}$	0.57×

144.10, compared with 143.07 without merging and 160.82 for ToMe. At 50%, UtoMe-E obtains 151.95, compared with ToMe's 220.05. At 1.40625°, UtoMe-E lowers ToMe's Overall error by 2.86 and 14.32 at the two budgets, respectively, at identical encoder FLOPs. These gains do not establish parity with the full-token model: at high resolution and 25% reduction, Z500 wRMSE increases from 156.76 to 191.93 (22.4%). Nor are the gains uniform across targets: ToMe has lower U10/V10 errors at low-resolution 25% reduction and lower V10 error than E/P at high-resolution 50% reduction.

Takeaway 1: Joint scoring improves budgeted forecasting

At matched token budgets, UtoMe lowers Overall error relative to ToMe at both resolutions. The per-variable results identify the benefits and costs of reallocating the token budget.

Token-reduction ratio ablation. Table 3 quantifies the high-resolution accuracy–compute frontier. At 5% reduction, Overall error rises from 32.41 to 32.80 (1.2%), ACC remains 0.964, and encoder FLOPs decrease by 3%. At 10%, Overall error rises by 6.4% for a 7% FLOPs reduction. At 25%, the 14% FLOPs saving comes with a 22.2% increase in Overall error. The 50% setting saves 45% of encoder FLOPs but increases Overall error by 63.5%. Thus, the low-loss operating point is the small-reduction regime; the larger budgets test robustness to substantial compression.



(a) **Z500 input.** (b) **Observation uncertainty.**

(c) **Importance score.** (d) **Patch-level token score.**

Figure 3: $1.40625°$ **high-resolution case study.** The observation-uncertainty field contains coherent bands and high-spread structures. After saliency combination and patch aggregation, these structures produce a heterogeneous token score for uncertainty-guided merging.

Table 3: Token-reduction ratio ablation at $1.40625°$. UtoMe corresponds to UtoMe-E.

Method	Token Red.	Overall ↓	ACC ↑	Z500 ↓	T850 ↓	T2m ↓	U10 ↓	V10 ↓	Compute ↓
No merging	0%	$32.41_{\pm 0.28}$	$0.964_{\pm 0.004}$	$156.76_{\pm 1.34}$	$1.299_{\pm 0.007}$	$1.269_{\pm 0.006}$	$1.339_{\pm 0.008}$	$1.360_{\pm 0.009}$	1.00×
UtoMe	5%	$32.80_{\pm 0.18}$	$0.964_{\pm 0.002}$	$158.70_{\pm 0.94}$	$1.305_{\pm 0.005}$	$1.278_{\pm 0.005}$	$1.349_{\pm 0.006}$	$1.372_{\pm 0.007}$	0.97×
UtoMe	10%	$34.49_{\pm 0.25}$	$0.961_{\pm 0.003}$	$167.00_{\pm 1.16}$	$1.330_{\pm 0.006}$	$1.310_{\pm 0.006}$	$1.380_{\pm 0.007}$	$1.410_{\pm 0.008}$	0.93×
UtoMe	25%	$39.59_{\pm 0.21}$	$0.952_{\pm 0.003}$	$191.93_{\pm 1.05}$	$1.413_{\pm 0.005}$	$1.454_{\pm 0.006}$	$1.527_{\pm 0.007}$	$1.645_{\pm 0.008}$	0.86×
UtoMe	50%	$53.00_{\pm 0.36}$	$0.913_{\pm 0.004}$	$256.72_{\pm 1.76}$	$1.759_{\pm 0.009}$	$2.051_{\pm 0.012}$	$2.038_{\pm 0.013}$	$2.433_{\pm 0.015}$	0.55×

Takeaway 2: Compression has a measurable accuracy cost

Small reductions preserve most full-token skill. At 25% and 50% reduction, UtoMe improves the trade-off relative to ToMe while incurring material losses relative to no merging.

Complementary uncertainty regimes. At high resolution and 50% reduction, UtoMe-S has lower errors than E/P on T850, T2m, U10, and V10, while UtoMe-E is stronger on Z500, Overall, and ACC. The mixed result is consistent with complementary allocation signals and does not establish that external spread is preferable for every target. Z500 dominates the numerical scale of Overall.

Additional baselines and backbone. With the joint score held fixed, UtoMe-E merging improves on score-based pruning at both budgets. On a plain 12-layer ViT, it improves over ToMe by 2.7%/4.9% at matched FLOPs. Appendix F reports the complete comparisons.

Uncertainty predictor ablation. UtoMe-P and UtoMe-S obtain forecasting gains without external uncertainty at inference (Tables 1 and 2). On three held-out $5.625°$ cases, P's predicted scores have Pearson correlations 0.769–0.825 with the EDA-derived teacher, while S reaches 0.858–0.900 against its distinct proxy target (Appendix G). These correlations describe score agreement, not uncertainty calibration, and are not directly comparable across targets. Ranking supports anchor selection, whereas matching also uses score magnitudes. The transformation controls below test empirical robustness to selected magnitude changes.

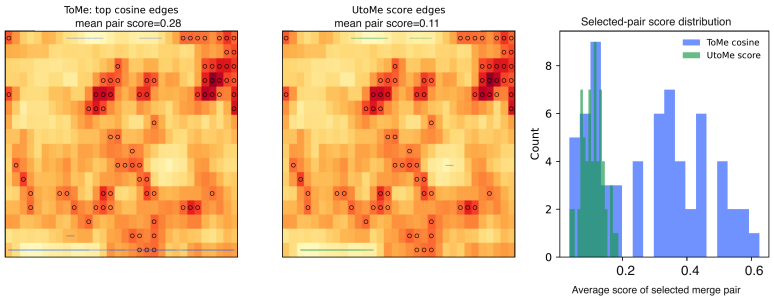


Figure 4: **Cosine-only matching can select merge edges in high-importance regions because feature similarity alone does not encode observation uncertainty.** UtoMe protects anchors and penalizes importance-incompatible pairs, shifting the merge budget toward lower-importance regions.

Takeaway 3: External spread is optional at deployment

P distills an externally supervised score; S learns a local-inconsistency proxy. Both support merging without external uncertainty at inference, with target-dependent differences in forecast skill.

Qualitative Analysis. Fig. 2 shows the low-resolution allocation pipeline; Fig. 3 illustrates its high-resolution signals. Fig. 4 compares selected edges under cosine-only and joint-importance matching. At 25% reduction, the merge-edge Jaccard overlap is 0.139, and merged-token membership in the top importance quartile falls from 46.8% to 0%. These diagnostics establish that the policy changes selected tokens; the forecasting comparisons establish its utility. Appendix G.1 distinguishes quantitative statistics from illustrative figure markers.

4.3 Component Contributions and UQ Robustness

Table 4: **UQ, anchors, and matching contribute at both budgets.** Overall error (↓) at $1.40625°$ with matched token schedules.

Configuration	25% reduction	50% reduction
ToMe	42.45	67.32
Dynamics only ($s = d$)	40.95	58.60
No anchors	40.31	61.40
No matching penalties	40.68	55.90
Full UtoMe-E	**39.59**	**53.00**

Table 5: **Spatial alignment matters more than the tested rescalings.** Score rank correlation and anchor overlap are relative to the EDA-derived score; errors are Overall error (↓) at $1.40625°$. The two-member row uses simulated sampling noise.

UQ field	Rank corr.	Anchor overlap	25%	50%
Original EDA	1.000	100%	39.59	53.00
$\sqrt{\text{UQ}}$	0.995	94%	39.62	53.05
UQ^2	0.977	88%	39.68	53.19
Rank-mapped UQ	0.969	84%	39.72	53.28
Simulated 2-member	0.960	79%	39.94	54.10
Constant UQ	–	–	40.95	58.60
Spatially shuffled UQ	≈ 0	26%	41.88	62.30

Contribution of the score, anchors, and matching. In Table 4, removing UQ increases Overall error by 1.36/5.60 at $25\%/50\%$ reduction. Removing anchors costs 0.72/8.40, and removing the matching penalties costs 1.09/2.90. The full component matrix in Appendix C also shows that UQ-only scoring underperforms the combined score.

Spatial information and rescaling. In Table 5, shuffling UQ degrades wRMSE to 41.88/62.30. Square-root, square, and rank transformations change wRMSE by at most 0.13/0.28, indicating that preserving spatial information matters more than these magnitude transformations. Appendix D reports the full controls and the finite-ensemble sampling-noise simulation.

4.4 Forecast Quality in Difficult Regimes

Table 6: **The gain concentrates in dynamically active regions, including high-UQ regions.** Z500 wRMSE (↓) at $1.40625°$ and 25% token reduction. L/H denote below/at-or-above the per-sample median of UQ (q) and dynamics (d). The two low-dynamics regimes expose the allocation trade-off.

Method	qL/dH	qH/dH	qL/dL	qH/dL	Global
No merging	172.0	200.0	116.0	124.0	156.76
ToMe	219.8	268.4	158.2	155.7	206.02
UtoMe-E	185.2	220.0	176.0	181.0	191.93

High uncertainty with strong dynamics. Using per-sample median splits of UQ and dynamics, Table 6 shows that UtoMe-E lowers high-UQ/high-dynamics Z500 wRMSE from 268.4 to 220.0, an 18.0% improvement over ToMe at 25% reduction. It also improves the low-UQ/high-dynamics regime, while errors increase in the two low-dynamics regimes. The result supports joint allocation in this regime, while exposing a cost in less dynamically active regions.

Extremes and spatial structure. At 50% reduction, UtoMe-E lowers upper-tail Z500, lower-tail Z500, and upper-tail T2m errors by 22.4%, 22.7%, and 16.9% relative to ToMe. Spatial standard-deviation, gradient, kinetic-energy, and vorticity ratios also remain closer to the no-merging reference at both budgets. These amplitude diagnostics do not establish conservation or full physical consistency. Appendix E provides the full results and definitions.

5 Discussion

The component and spatial controls support joint allocation by atmospheric structure and input reliability. Dynamics identifies gradients and anomalies, while uncertainty modulates confidence in the input estimate. In the joint score, strong dynamics can counterbalance elevated spread; anchor protection then retains high-scoring structures. The gains in both high-dynamics regimes, together with losses in low-dynamics regions (Table 6), are consistent with concentrating representation capacity on dynamically active regions. This trade-off makes uncertainty most useful when interpreted alongside the atmospheric state.

The choice among E/P/S depends on access to aligned uncertainty products. E uses spread directly, P transfers an externally supervised scoring rule to deployment, and S learns from an input-derived proxy. Their comparison should account for the delayed merge schedule in P/S as well as the final token budget. Within each regime, the reduction ratio should reflect acceptable errors in the target variables: the small-reduction frontier supports accuracy-sensitive use, while larger reductions offer greater encoder savings at a measurable forecast cost.

6 Conclusion

We introduced UtoMe, a weather-specific token-allocation policy combining dynamics saliency, relative input uncertainty, anchors, and importance-calibrated matching. Across two ERA5 resolutions, it improves Overall error over ToMe at matched token budgets; controls support contributions from aligned uncertainty and both allocation mechanisms. The three variants accommodate different availability of external uncertainty. Small token reductions preserve most full-token skill, while aggressive compression incurs substantial error. Evidence is limited to two transformer backbones and direct 72-hour ERA5 forecasts; EDA spread may miss shared biases or fine-scale extremes, and encoder FLOPs do not measure end-to-end training time. Appendix I details these limitations.

Acknowledgments and Disclosure of Funding

This manuscript has been authored by UTBattelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). This research was supported by the ORNL's AI Initiative sponsored by the Director's Research and Development Program at ORNL. This work's computing time was supported through the Oak Ridge Leadership Computing Facility.

References

[1] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[2] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641: 1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

[3] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4599–4603, 2023.

[4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[5] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. Measurement error in nonlinear models: A modern perspective. *Chapman and Hall/CRC*, 2006.

[6] Yutian Chen, Yuheng Qiu, Ruogu Li, Ali Agha, Shayegan Omidshafiei, Jay Patrikar, and Sebastian Scherer. Co-me: Confidence-guided token merging for visual geometric transformers, 2025. URL https://arxiv.org/abs/2511.14751.

[7] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[8] Wayne A. Fuller. *Measurement Error Models*. John Wiley & Sons, 1987.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[10] Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3369–3378, 2018.

[11] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

[12] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3.

[13] Tijana Janjić, Niels Bormann, Marc Bocquet, James A. Carton, Stephen E. Cohn, Sarah L. Dance, Svetlana N. Losa, Nancy K. Nichols, Roland Potthast, Joanne A. Waller, and Peter Weston. On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144(713):1257–1278, 2018. doi: 10.1002/qj.3130.

[14] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[15] Felix Krause, Timy Phan, Ming Gui, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Tread: Token routing for efficient architecture-agnostic diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15703–15713, 2025.

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[17] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[18] Yuqi Li, Xi Xiao, Yunbei Zhang, Lin Zhao, Yu Li, Aiden Zhao, Tianyang Wang, Hao Xu, and Yingli Tian. Rethinking layer-wise information allocation for vision foundation model adaptation. *arXiv preprint arXiv:2607.21973*, 2026.

[19] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[20] Jörg Martin and Clemens Elster. Aleatoric uncertainty for errors-in-variables models in deep regression. *Neural Processing Letters*, 2022. doi: 10.1007/s11063-022-11066-3.

[21] Andrew McHutchon and Carl Edward Rasmussen. Gaussian process training with input noise. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[22] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383, 2021.

[23] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[24] Dogyun Park, Moayed Haji-Ali, Yanyu Li, Willi Menapace, Sergey Tulyakov, Hyunwoo J Kim, Aliaksandr Siarohin, and Anil Kag. Sprint: Sparse-dense residual fusion for efficient diffusion transformers. *arXiv preprint arXiv:2510.21986*, 2025.

[25] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[26] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[27] Vikash Sehwag, Xianghao Kong, Jingtao Li, Michael Spranger, and Lingjuan Lyu. Stretching each dollar: Diffusion training from scratch on a micro-budget. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28596–28608, 2025.

[28] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[29] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[30] Matias Valdenegro-Toro, Ivo Pascal de Jong, and Marco Zullich. Unified uncertainties: Combining input, data and model uncertainty into a single formulation. In *ICML 2024 LatinX in AI (LXAI) Research Workshop*, 2024.

[31] Janet Wang, Yunbei Zhang, Lin Zhao, Xi Xiao, Jihun Hamm, and Xiao Wang. Less tokens, better forecasts: Sparse residual routing for efficient weather prediction. *arXiv preprint arXiv:2607.02829*, 2026.

[32] Xiao Wang, Jong-Youl Choi, Takuya Kurihaya, Isaac Lyngaas, Hong-Jun Yoon, Xi Xiao, David Pugmire, Ming Fan, Nasik Muhammad Nafi, Aristeidis Tsaris, et al. Orbit-2: Scaling exascale vision foundation models for weather and climate downscaling. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 86–98, 2025.

[33] Xiao Wang, Zezhong Zhang, Isaac Lyngaas, Hong-Jun Yoon, Jong-Youl Choi, Siming Liang, Janet Wang, Hristo G Chipilski, Ashwin M Aji, Feng Bao, et al. Global attention with linear complexity for exascale generative data assimilation in earth system prediction. *arXiv preprint arXiv:2604.16590*, 2026.

[34] Yingfan Xu, Tieming Liu, Ye Liang, and Taiping Liu. RiVaT-Fuse: Reliability-calibrated variational tensor fusion for multimodal prediction under modality uncertainty. *arXiv preprint arXiv:2609.10798*, 2026. URL https://arxiv.org/abs/2609.10798.

[35] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

Appendix

A Proof of the Three-Way Variance Decomposition

We give the short derivation used in Sec. 3.2. All expectations and variances are conditioned on $\tilde{\boldsymbol{x}}$ and $\mathcal{D}$ unless stated otherwise, with the factorized posterior assumed in Sec. 3.2. The conditional second moments are finite. Applying the law of total variance over $\boldsymbol{x}^* \sim p(\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}})$ gives

$$\mathrm{Var}[y] = \mathbb{E}_{\boldsymbol{x}^*}[\mathrm{Var}(y \mid \boldsymbol{x}^*)] + \mathrm{Var}_{\boldsymbol{x}^*}(\mathbb{E}[y \mid \boldsymbol{x}^*]) \,. \tag{11}$$

The second term is the observation uncertainty term σ^2_{obs}. For a fixed $\boldsymbol{x}^*$, applying the law of total variance again over $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D})$ gives

$$\mathrm{Var}(y \mid \boldsymbol{x}^*) = \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{Var}(y \mid \boldsymbol{x}^*, \boldsymbol{\theta})] + \mathrm{Var}_{\boldsymbol{\theta}}(\mathbb{E}[y \mid \boldsymbol{x}^*, \boldsymbol{\theta}]) \,. \tag{12}$$

Substitution yields Eq. (3). If the observation noise is zero, $p(\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}})$ collapses to a point mass at $\tilde{\boldsymbol{x}}$, the observation term vanishes, and the usual aleatoric and epistemic decomposition is recovered.

The delta-method estimate follows by expanding $\mu(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}}[f_{\boldsymbol{\theta}}(\boldsymbol{x})]$ around $\tilde{\boldsymbol{x}}$:

$$\mu(\boldsymbol{x}^*) \approx \mu(\tilde{\boldsymbol{x}}) + \boldsymbol{J}_\mu(\tilde{\boldsymbol{x}})^\top (\boldsymbol{x}^* - \tilde{\boldsymbol{x}}). \tag{13}$$

Taking variance over $\boldsymbol{x}^*$ gives

$$\sigma^2_{\mathrm{obs}}(\tilde{\boldsymbol{x}}) = \mathrm{Var}_{\boldsymbol{x}^*}[\mu(\boldsymbol{x}^*)] \approx \boldsymbol{J}_\mu(\tilde{\boldsymbol{x}})^\top \boldsymbol{\Sigma}_{\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}}} \boldsymbol{J}_\mu(\tilde{\boldsymbol{x}}). \tag{14}$$

Here $\boldsymbol{\Sigma}_{\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}}} = \mathrm{Cov}(\boldsymbol{x}^* \mid \tilde{\boldsymbol{x}})$ is the latent-input posterior covariance. Identifying it with an additive observation-noise covariance requires further assumptions, such as a locally flat latent-state prior and input-independent Gaussian noise. With a diagonal posterior covariance, the first-order term sums squared sensitivities weighted by latent-input variances. This motivates considering both sensitivity and reliability, but does not derive $s_i = d_i - \beta q_i$, an optimal merging rule, or a forecast-error bound. We do not estimate the decomposition's components or use it as a training loss.

B Implementation and Evaluation Details

Data splits and tokenization. The $5.625°$ experiments use the split-sharded ERA5 setting with 2010 to 2014 for training, 2015 for validation, and 2016 for testing. The files are split into 320 train shards, 64 validation shards, and 64 test shards so that file-level DDP sharding remains balanced over the 64-GPU setting. Each variable has spatial shape 32×64, and the patch size is 2, producing a 16×32 token grid. The $1.40625°$ experiments use 1979 to 2018 for training, 2019 for validation, and 2020 for testing. Each variable has spatial shape 128×256, and the patch size is 4, producing a 32×64 token grid. In both settings, the forecast pair is constructed as input x_t and target y_{t+72}.

Variables. The 41 input channels are the static fields land-sea mask, orography, and latitude; the surface variables T2m, U10, and V10; and geopotential, zonal wind, meridional wind, temperature, and specific humidity at pressure levels 50, 250, 500, 600, 700, 850, and 925 hPa. The five forecast targets are Z500, T850, T2m, U10, and V10. Inputs and outputs are normalized using the training-set mean and standard deviation. RMSE and ACC are computed after denormalization.

Uncertainty fields. External UQ denotes ECMWF EDA spread fields aligned with the same time and grid as the ERA5 samples. In the reported runs, UQ is constructed for the full input-channel set used by the forecaster, with the same year split and shard layout as the weather fields. UQ fields are not used as extra forecast input channels. They are normalized per variable, patch-averaged, and used as teacher or calibration signals for token merging. Static variables receive zero observation uncertainty when constructing the full input-channel UQ tensor.

Importance score. Dynamics saliency uses the input-time fields of the selected forecast variables and Eq. (5). Adjacent-grid differences are backward differences, with a zero difference at the first row/column. The anomaly term is centered on the spatial mean of each input channel and has weight $1/2$. The dynamics map and the UQ map are patch-averaged with the model patch size and normalized per sample, yielding d_i and q_i. The reported configurations use $s_i = d_i - \beta q_i$ with $\beta = 1.0$. This makes high-dynamics, low-spread patches strong anchor candidates and treats high-spread patches as

less reliable merge candidates unless the dynamics signal is large. The anchor fraction is $\rho = 0.25$. The matching weights in Eq. (7) are $\lambda_s = 0.05$ and $\lambda_\Delta = 0.05$ unless otherwise stated.

Variants and predictor. UtoMe-E uses EDA-derived teacher importance directly for merge bias and anchor protection and requires external UQ at evaluation. UtoMe-P uses EDA-derived teacher importance during training, trains a convolutional importance head from the hidden tokens at probe layer 3 (zero-based indexing), and uses predicted importance for merging after a warmup period and during testing. UtoMe-S uses no external UQ: it builds a proxy UQ target from local 3×3 spatial inconsistency, trains the same predictor, and uses predicted proxy importance at inference. Predictor-based variants use a pairwise-ranking importance loss with weight 0.2.

Merge schedules. The main tables use final token-reduction ratios of 25% and 50%. At $5.625°$, these remove 128 and 256 tokens from an initial sequence of 512. At $1.40625°$, they remove 512 and 1024 tokens from an initial sequence of 2048. For ToMe and UtoMe-E, ratio-based merging is spread over all eight transformer layers. For UtoMe-P and UtoMe-S, the predictor is produced at probe layer 3, so the current implementation delays ratio-based merging until after the probe layer; the first four layers therefore run with the full token count.

Optimization. Training uses DDP on an HPC cluster of 8 nodes, each exposing 8 AMD MI250X GPU devices, for 64 GPU devices in total. We use AdamW with $\beta = (0.9, 0.99)$, weight decay 10^{-5}, peak learning rate 5×10^{-4}, 1000 warmup steps, minimum learning rate 10^{-8}, gradient clipping at 1.0, periodic validation, and a learning-rate schedule calibrated to 12000 steps. The per-process batch size is 16 at $5.625°$ and 1 at $1.40625°$, corresponding to global batch sizes 1024 and 64 under the 64-GPU setting.

B.1 Normalization, selection, and evaluation

For a patch vector v, per-sample standardization is

$$\mathcal{Z}(v)_i = \frac{v_i - \bar{v}}{\max(\mathrm{std}(v), 10^{-6})}. \tag{15}$$

Thus $d_i = \mathcal{Z}(\mathrm{PatchAvg}(D))_i$ and q_i is the corresponding standardized spread score. A spatially constant spread gives $q_i = 0$, recovering the dynamics-only score. Anchor selection protects the highest-scoring fraction when merging begins; protected tokens are excluded from both endpoints of subsequent merges.

We select β, the anchor fraction, and the matching weights using the held-out validation split, then fix the configuration before evaluating the test set. The sensitivity tables report test performance of the stated settings; the default is $\beta = 1$, $\rho = 0.25$, and $\lambda_s = \lambda_\Delta = 0.05$. The main resolution tables report three-seed means and standard deviations. The additional control tables report point estimates.

For a target field, latitude weighting uses $w_h = \cos(\varphi_h)$ normalized over the evaluated grid. Per-variable wRMSE is the square root of the weighted squared prediction error; Overall averages the five variable-specific errors as in Eq. (10). ACC is the weighted correlation of predicted and reference anomalies after removal of climatology. We report per-variable errors because their units and magnitudes differ. Supplemental masked diagnostics use per-sample weighted errors and average the valid sample values.

C Component Ablations and Parameter Sensitivity

The component study compares the full score and allocation mechanism with controlled removals at the same token budgets. Table 7 expands Table 4. UQ alone is weaker than the combined dynamics–UQ score, and removing either anchors or matching penalties degrades the full method. The anchor removal has the largest effect at 50% reduction.

Table 8 varies each major parameter around the validation-selected configuration. Increasing β to 4 raises the 50% error to 69.40, and increasing the matching weights to 0.50 raises it to 63.70. Excessive reliance on UQ or matching penalties can therefore overconstrain allocation. Increasing the anchor fraction above 0.25 also raises error. The moderate settings provide the best results in these reported sweeps.

Table 7: **Both score inputs and both allocation mechanisms contribute.** Overall error (↓) at $1.40625°$. The no-anchor and no-matching rows retain the combined score. Disabling both allocation mechanisms recovers ToMe.

Configuration	25%	50%
ToMe	42.45	67.32
Dynamics only ($s = d$)	40.95	58.60
UQ only ($s = -q$)	41.40	61.70
No anchors	40.31	61.40
No matching penalties	40.68	55.90
Full UtoMe-E	**39.59**	**53.00**

Table 8: **Moderate parameter values balance protection and compression.** One-parameter sweeps at $1.40625°$, with other settings held at the default. Columns report Overall error (↓). The validation-selected default values are $\beta = 1$, $\rho = 0.25$, and $\lambda_s = \lambda_\Delta = 0.05$.

Parameter	Value	25%	50%
β	0	40.95	58.60
	0.25	40.12	55.10
	0.5	39.74	53.62
	1	39.59	53.00
	2	40.36	54.85
	4	43.80	69.40
ρ	0	40.31	61.40
	0.05	40.05	58.20
	0.10	39.78	55.30
	0.25	39.59	53.00
	0.40	39.86	53.94
	0.50	40.62	56.80
$\lambda_s = \lambda_\Delta$	0	40.68	55.90
	0.01	40.02	54.20
	0.05	39.59	53.00
	0.10	39.71	53.35
	0.20	40.94	56.10
	0.50	44.60	63.70

D Robustness to Uncertainty Estimates

D.1 Spatial controls and monotone transformations

The UQ controls preserve the weather input and token budget while modifying the uncertainty field. Spatial shuffling retains the spread values but changes their locations. The random-noise control matches the per-channel scale statistics of the reference spread. Constant UQ removes the standardized uncertainty contribution. Square-root and square transformations change magnitude spacing, and rank mapping replaces spread magnitudes with normalized spatial ranks. The sign control changes only the score direction to $s = d + \beta q$.

Table 9 reports both forecast performance and agreement with the original importance signal. Rank correlation compares patch-score orderings; anchor overlap measures the shared fraction of protected tokens. The combined score includes dynamics as well as UQ, so its ordering can change under a monotone transform of UQ. The measured stability is an empirical property of these transformations.

D.2 Simulated ensemble sampling noise

To study finite-ensemble estimation noise, we perturb the reference EDA spread at several nominal ensemble-size settings. ERA5 EDA itself uses ten members; the counts in this simulation label estimator-noise settings, rather than observed subsets or additional EDA members. Table 11 re-

Table 9: **Tested rescalings preserve useful spatial order.** Score rank correlation and anchor overlap are relative to the original EDA field. The last two columns are Overall error (↓) at $1.40625°$. Dashes mark agreement metrics not reported for the constant field.

UQ field	Rank corr.	Anchor overlap	25%	50%
Original EDA	1.000	100%	39.59	53.00
$\sqrt{\text{UQ}}$	0.995	94%	39.62	53.05
UQ^2	0.977	88%	39.68	53.19
Rank-mapped UQ	0.969	84%	39.72	53.28
Simulated 2-member	0.960	79%	39.94	54.10
Constant UQ	–	–	40.95	58.60
Shuffled UQ	≈ 0	26%	41.88	62.30

Table 10: **Removing spatial information or reversing the UQ direction degrades performance.** Overall error (↓) at $1.40625°$.

Control	25%	50%
Original EDA spread	39.59	53.00
Constant UQ / dynamics only	40.95	58.60
Spatially shuffled UQ	41.88	62.30
Matched random-noise UQ	42.07	63.15
Inverted sign ($s = d + \beta q$)	42.64	64.90

ports score correlations and forecast errors. These results use simulated spread perturbations, not subsampling of archived raw ensemble members.

Table 11: **Finite-ensemble sampling noise causes gradual degradation.** This experiment simulates noisier spread estimates, rather than subsampling archived raw ensemble members. Rank correlation is measured against the reference importance score; errors are Overall wRMSE at $1.40625°$.

Simulated members	Rank corr.	25%	50%
Full reference	1.000	39.59	53.00
32	0.999	39.60	53.02
16	0.997	39.62	53.08
8	0.994	39.66	53.21
4	0.986	39.76	53.55
2	0.960	39.94	54.10

E Weather-Regime, Extreme-Value, and Structural Diagnostics

E.1 Regime definitions and aggregation

Table 6 uses four masks defined by the per-sample medians of the patch-level UQ and dynamics fields. A patch is assigned to the high group when its value is at least the corresponding median. Patch masks are expanded to their spatial support. For a mask A and one forecast sample, we compute

$$\text{wRMSE}(A) = \sqrt{\frac{\sum_{h,w} w_h \, \mathbf{1}_A(h,w)(\hat{y}_{h,w} - y_{h,w})^2}{\sum_{h,w} w_h \, \mathbf{1}_A(h,w)}}. \tag{16}$$

Valid per-sample errors are averaged over the evaluation samples. Each quadrant is normalized by its own weight sum; the reported Global error is evaluated over the full field, not an unweighted average of the four quadrant errors. At 25% reduction, the high-dynamics improvements of 34.6 and 48.4 Z500 wRMSE coexist with low-dynamics increases of 17.8 and 25.3, respectively. This describes where the budget reallocation helps and where it incurs a cost.

E.2 Tail errors

For each sample and target variable, upper/lower-tail masks select values above/below the spatial 95th/5th percentiles of the reference anomaly field. We evaluate prediction error in the target's physical units within those masks using the weighted expression above. Table 12 shows the complete upper/lower Z500 and upper T2m results.

Table 12: **UtoMe-E better retains forecast extremes than ToMe at both budgets.** Conditional wRMSE (↓) on the upper/lower 5% target-anomaly masks at $1.40625°$. Z500 errors use $\mathrm{m}^2\,\mathrm{s}^{-2}$ and T2m errors use K.

Method	Reduction	Z500 upper	Z500 lower	T2m upper
No merging	0%	189.9	242.5	1.953
ToMe	25%	251.4	318.7	2.318
UtoMe-E	25%	226.8	285.3	2.147
ToMe	50%	402.6	512.4	3.145
UtoMe-E	50%	312.5	396.1	2.612

E.3 Spatial and wind-structure ratios

The standard-deviation ratio is the predicted spatial standard deviation divided by the reference standard deviation. The gradient ratio compares latitude-weighted mean gradient magnitudes, using physical grid spacing. For U10/V10, kinetic energy per unit mass is $K = (u^2 + v^2)/2$, and the relative-vorticity diagnostic uses the finite-difference expression $\zeta = \partial_x v - \partial_y u$. The KE ratio compares weighted mean K and the vorticity ratio compares weighted mean $|\zeta|$ between prediction and reference. Ratios are formed per sample and then averaged. Longitude differences are periodic for these diagnostics, and latitude differences use one-sided boundary handling. A ratio of one indicates matching amplitude statistics.

Table 13 shows improved preservation of spatial variability, gradients, and wind statistics relative to ToMe.

Table 13: **Spatial and wind statistics remain closer to the full-token reference.** Prediction/reference ratios at $1.40625°$; unity denotes agreement with the reference statistic. The spatial standard-deviation and gradient columns concern Z500; KE and vorticity use U10/V10.

Method	Reduction	σ ratio	∇ ratio	KE ratio	Vort. ratio
No merging	0%	0.998	1.013	0.975	0.870
ToMe	25%	0.964	1.092	0.931	0.812
UtoMe-E	25%	0.981	1.048	0.958	0.846
ToMe	50%	0.907	1.213	0.874	0.731
UtoMe-E	50%	0.955	1.104	0.936	0.802

F Additional Baselines and Backbone

F.1 Pruning controls

Table 14 compares random, dynamics-only, UQ-only, and joint-score selection with merging. The comparison holds the token budget and schedule fixed across policies. The joint score improves over the strongest alternative pruning score by 2.64 at 25% reduction and 2.56 at 50%. Holding that joint score fixed, UtoMe-E merging further improves error by 1.15 and 4.21, respectively. Both the allocation signal and the representation-combining operator matter in this setting.

Table 14: **The joint score benefits pruning, and merging preserves more forecast skill.** Overall error (↓) at $1.40625°$, using the matched backbone, split, schedule, and token budgets of the comparison. Encoder FLOPs are $0.86\times$/$0.55\times$ at 25%/50% reduction.

Method	25%	50%
Random pruning	61.86	85.90
Dynamics-saliency pruning	44.42	61.02
UQ-only pruning	43.38	59.77
UtoMe-score pruning	40.74	57.21
ToMe merging	42.45	67.32
UtoMe-E merging	**39.59**	**53.00**

F.2 A plain 12-layer ViT

The second backbone uses a single joint patch embedding, twelve transformer layers, and no per-variable embeddings or variable-aggregation attention. The same score, anchor protection, and matching mechanism apply to its spatial tokens. Table 15 reports matched-FLOPs comparisons within this backbone. At 25% reduction, UtoMe-E obtains 141.13 versus 145.05 for ToMe; at 50%, it obtains 149.44 versus 157.14. These results extend the evidence beyond ClimaX-style variable tokenization to another patch-transformer architecture.

Table 15: **The allocation rule transfers to a plain patch-transformer backbone.** Overall wRMSE (↓) on a 12-layer ViT with a joint patch embedding. FLOPs are relative to this backbone without merging.

Method	Reduction	wRMSE	Encoder FLOPs
No merging	0%	141.96	$1.000\times$
ToMe	25%	145.05	$0.829\times$
UtoMe-E	25%	141.13	$0.829\times$
ToMe	50%	157.14	$0.661\times$
UtoMe-E	50%	149.44	$0.661\times$

G Predictor Quality and Merge-Decision Analysis

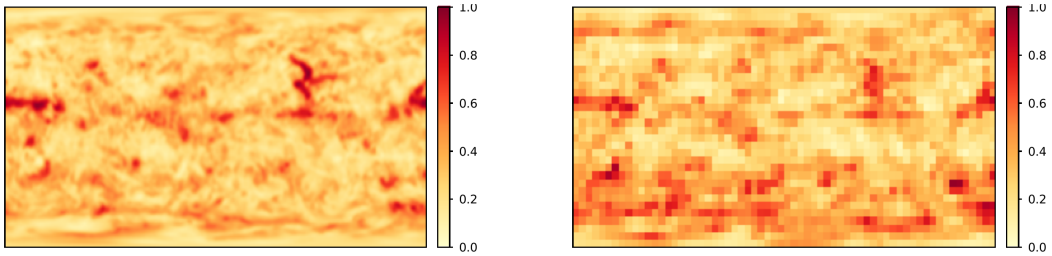
Table 16 reports the held-out patch-level predictor alignment summarized in Sec. 4.2. The EDA-derived teacher and the self-supervised proxy define different prediction targets; correlations should be read within each target. Forecast performance in the main tables provides the downstream measure of allocation quality.

Table 16: Predictor quality on held-out $5.625°$ cases. The case label denotes the spatial resolution, not a separate dataset. Correlation and z-normalized MAE are computed patch by patch between the predictor output and its supervision target.

Case	UtoMe-P corr. ↑	UtoMe-P zMAE ↓	UtoMe-S corr. ↑	UtoMe-S zMAE ↓
$5.625°$ case 01	0.820	0.466	0.858	0.403
$5.625°$ case 02	0.825	0.463	0.900	0.341
$5.625°$ case 03	0.769	0.525	0.870	0.396

G.1 Merge-set and importance statistics

Let E_T and E_U denote the selected merge-edge sets for ToMe and UtoMe. Their Jaccard overlap, $|E_T \cap E_U|/|E_T \cup E_U|$, is 0.139 at 25% reduction. Table 17 summarizes the accompanying importance diagnostics. The first two rows concern the merged-token statistics; the last row concerns the displayed pair endpoints in Fig. 4. These use different aggregation and score-display conventions and are



(a) **Combined observation uncertainty.** (b) **Patch-level token score.**

Figure 5: **Motivation for uncertainty-guided token merging at** $1.40625°$**.** The observation-uncertainty field is not uniform noise; it contains coherent spatial structure. After patch aggregation and combination with dynamics saliency, the token-score grid separates low-uncertainty high-saliency regions from high-spread merge candidates.

reported separately. The figure's top-15% circle markers are a visualization aid, whereas algorithmic protection uses the 25% anchor fraction and the allocation diagnostic uses the top importance quartile.

Table 17: **UtoMe changes the selected merge set.** Allocation diagnostics at 25% token reduction. Quartile membership and merged-token importance refer to the standardized importance score.

Diagnostic	ToMe	UtoMe
Merged tokens in top quartile	46.8%	0%
Mean merged-token importance	+0.732	-1.456
Displayed-pair mean (Fig. 4)	0.28	0.11

H Additional Qualitative Visualizations

Fig. 6 expands the score construction for a held-out $5.625°$ case. The dynamics map highlights fronts and high-gradient structures, while the observation-UQ map highlights regions with larger EDA spread. The teacher and predicted importance maps are not simple copies of either input map; they preserve broad dynamically important bands while suppressing less reliable high-spread regions. This supports the design choice of combining dynamics and uncertainty before deciding which tokens to protect.

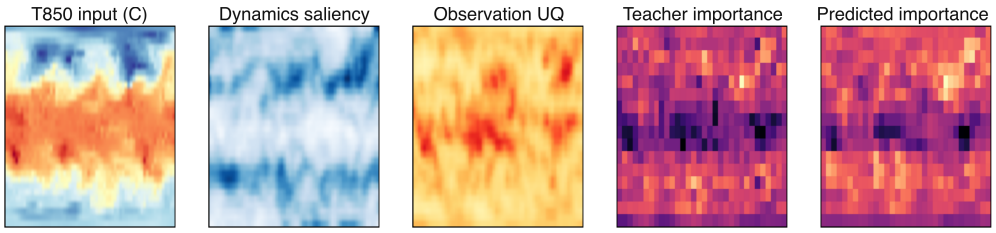


Figure 6: **Additional importance-score decomposition.** From left to right, the figure shows the input weather state, dynamics saliency, observation uncertainty, teacher importance, and predicted importance for a held-out $5.625°$ case.

Fig. 7 visualizes the predictor setting used by UtoMe-P and UtoMe-S. The top row compares the EDA-derived teacher score with the UtoMe-P predictor, while the bottom row compares the self-supervised pseudo target with the UtoMe-S predictor. The visual agreement is not perfect, but the broad high- and low-score regions are preserved. This agreement supports learning a useful score for allocation. Anchor selection uses score ordering, while pair penalties additionally depend on score magnitudes.

Fig. 8 shows the final token-level decision induced by the score. Anchor tokens concentrate in high-importance regions that should retain separate representations, while merge candidates occupy

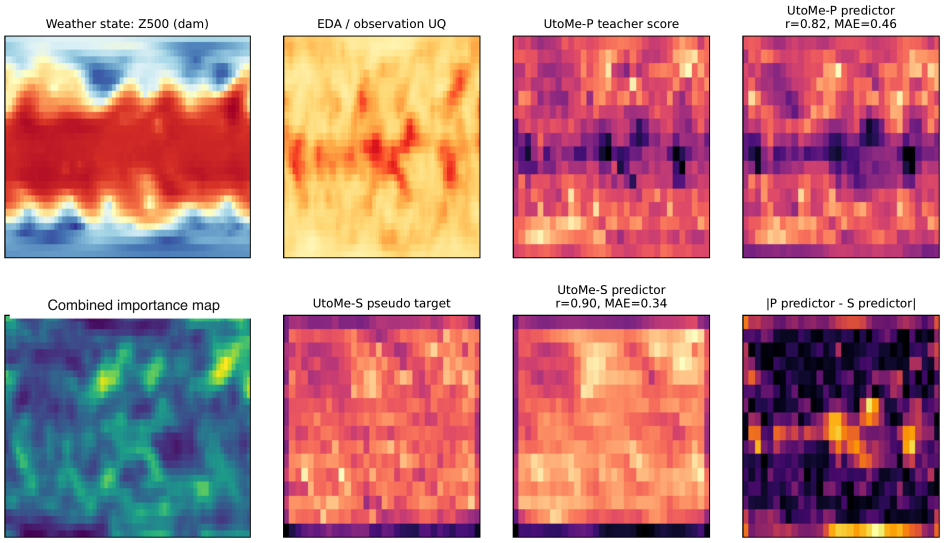


Figure 7: **Predictor visualization for UtoMe-P and UtoMe-S.** The learned predictors recover the broad spatial structure of their supervision targets, which explains why the predictor-based variants remain close to UtoMe-E in the main results.

lower-score regions where local redundancy is higher. The visualization complements Fig. 4: the former shows the absolute anchor and candidate layout, and the latter shows how this layout changes the selected merge edges relative to cosine-only matching.

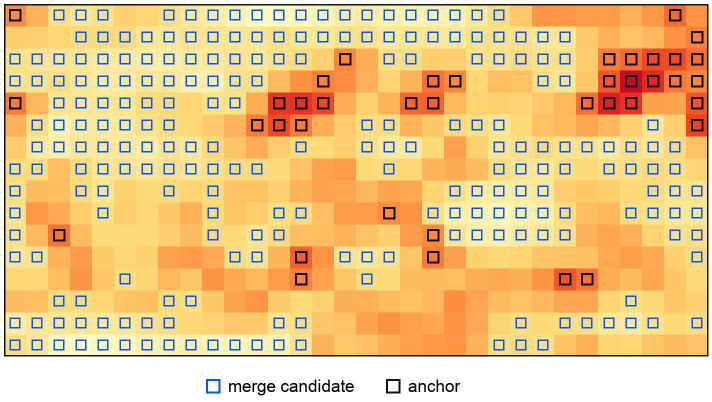
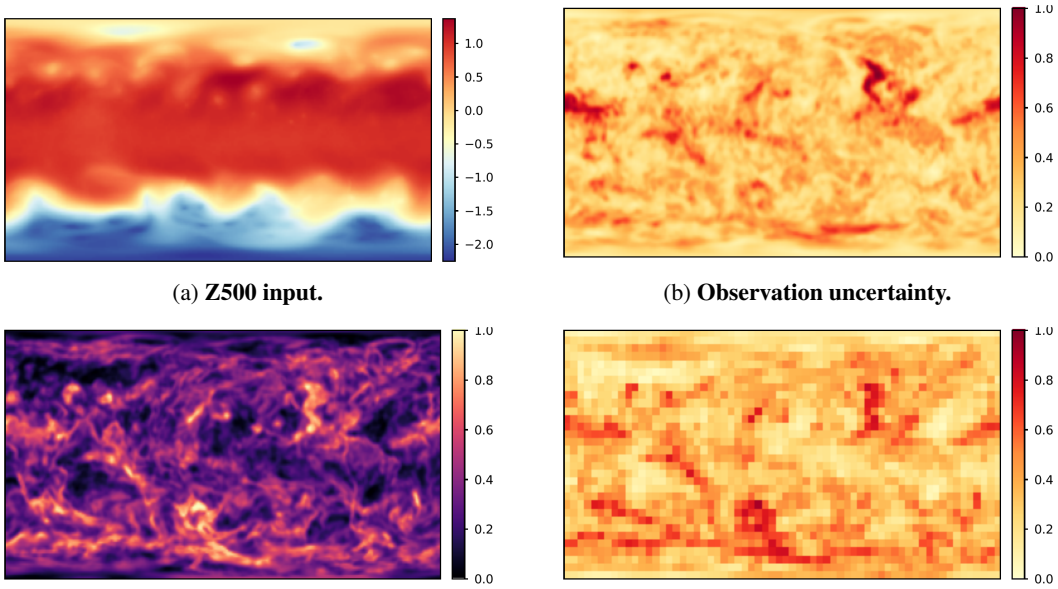


Figure 8: **Additional anchor and merge-candidate overlay at** $5.625°$**.** Blue squares denote merge candidates and black squares denote anchors selected from the patch-level token score.

Fig. 9 provides an additional high-resolution case study at $1.40625°$. The observation-uncertainty field contains coherent bands and localized high-spread structures that do not simply duplicate the weather state. After combining this field with dynamics saliency, the patch-level token score becomes smoother and more directly usable by the merge policy. This case supports the same qualitative conclusion as the main visualizations: uncertainty-guided scoring gives the model a structured basis for deciding which regions should be protected and which regions can be merged.

I Limitations

Uncertainty semantics and calibration. EDA spread reflects uncertainty in the assimilated reanalysis state, including observation and model effects, rather than isolating sensor noise. ECMWF describes it as a relative uncertainty estimate; common biases and under-resolved cyclone structure can produce



(a) **Z500 input.** (b) **Observation uncertainty.**

(c) **Importance score.** (d) **Patch-level token score.**

Figure 9: **Additional** $1.40625°$ **case study.** The panels show how a high-resolution weather state and its aligned observation-uncertainty field are converted into the importance and token-score maps used by UtoMe.

small spread despite substantial error. Spatial and temporal alignment, per-variable spread scaling, and changes in the observing system can affect the score. Our monotone-transformation and simulated sampling-noise controls do not establish robustness to arbitrary distribution shift. The simulated ensemble-size experiment does not subsample raw EDA members.

Allocation and difficult weather. Dynamics saliency is a heuristic based on gradients and anomalies, not a calibrated forecast sensitivity. High uncertainty need not imply irrelevance: fronts, data-sparse regions, and rapidly developing events can be both uncertain and important. The regime analysis supports joint scoring on the tested data while recording losses in low-dynamics regions. Conditional tail errors are not event-specific cyclone or precipitation evaluations. Gradient and wind-amplitude ratios are useful structural diagnostics but do not verify conservation laws or complete physical consistency.

Analytical and architectural scope. The variance identity is conditional on the stated posterior factorization and does not prove the direction or optimality of the allocation policy. EDA does not supply separate calibrated token-level aleatoric and epistemic estimates, so their effects on merging are not directly compared. We evaluate a ClimaX-style backbone and a plain ViT, not GraphCast, FourCastNet, Pangu-Weather, or other tokenizations. The evidence concerns direct 72-hour forecasts at two ERA5 resolutions; long autoregressive rollouts, operational forecasting, and other modalities remain untested.

Metrics, compute, and reproducibility. Overall averages errors with different physical units and is dominated by Z500; the per-variable metrics should guide interpretation. At high resolution, 25% and 50% token reductions incur material losses relative to no merging. Encoder FLOPs exclude data processing, uncertainty estimation, and other end-to-end costs, and do not establish proportional wall-clock savings. P/S have different merging schedules and slightly higher compute than E/ToMe. Additional controls are point estimates rather than three-seed comparisons. Code, configurations, and run instructions are available at https://github.com/ornl-utome/UtoMe.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state the contributions of UtoMe as observation-uncertainty-guided token merging with three UQ-availability variants, and these claims are supported by Sec. 3 and the experiments in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix I discusses the main limitations: experiments cover ERA5 forecasting at two resolutions on the primary backbone and an additional plain ViT comparison; the discussion also covers UQ shift, finite-ensemble simulation, difficult regimes, and the distinction between encoder FLOPs and end-to-end efficiency.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The three-way variance decomposition is stated in Sec. 3.2 with finite second moments and an explicit factorized latent-input/parameter posterior; Appendix A gives the full derivation and the delta-method estimate.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 4.1 and Appendix B document data splits, variable lists, model architecture, merge schedule, importance-score configuration, and optimizer settings, which together specify the procedure used to produce the reported numbers.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: ERA5 is publicly available. The project page linked in the abstract provides access to the public code repository, including the E/P/S implementations, training and checkpoint evaluation, data-preparation instructions, and configurations for the main forecast comparisons.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4.1 reports the data splits, model configuration, optimizer, learning-rate schedule, and merge ratios; Appendix B gives the full implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1 and 2 report mean and standard deviation over three random seeds. Additional control tables give point estimates; Appendix B distinguishes these from the repeated-run main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Sec. 4.1 and Appendix B report the GPU type and count, batch sizes, and training schedule. Per-run execution time, peak memory use, and total project compute are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses publicly available reanalysis data, does not involve human subjects, and the authors have reviewed and follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The work is a methodological contribution to efficient training of weather forecasting models on the publicly available ERA5 reanalysis, and we do not identify a direct path to negative societal impact distinct from the broader weather-modeling literature.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases an efficient training method for weather forecasters and does not release pretrained generative models, scraped datasets, or other artifacts with a high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: ERA5 reanalysis data [11] is used under the Copernicus Climate Data Store license, the ClimaX-style model and data conventions are credited to Nguyen et al. [23] (MIT license), and ToMe [4, 3] is credited as the train-time merging baseline.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released code includes a README, data-preparation instructions, training and evaluation scripts, configuration files, and licensing information. The paper does not release a new dataset or pretrained model.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing or any research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve human subjects, so no IRB approval is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method; any incidental use is limited to writing and formatting and therefore does not require declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.