Towards Disentangling Predictive Uncertainty in End-to-End AI Weather Forecasts

Rodrigo Almeida [1] Noelia Otero [1] Jackie Ma [1]

[1]Applied Machine Learning Group, Fraunhofer Heinrich-Hertz Institute, Berlin, Germany, rodrigo.almeida@hhi.fraunhofer.de

Background & Motivation

End-to-End AI Weather Forecasting

- Data-driven AI weather prediction (AIWP) models now match or exceed numerical weather prediction (NWP) models deterministic and probabilistic skill at medium range [1].
- Aardvark Weather [2] goes further: producing skillful global and weather station forecasts directly from observations, bypassing numerical solvers entirely at inference time.

Uncertainty Gap

- Operational usage of weather forecasts relies on ensemble systems like IFS ENS to quantify forecast uncertainty.
- Probabilistic AIWP has been developed using input perturbations, multi-model and multi-checkpoints ensembles.
- No existing work explores how to model uncertainty and where uncertainty originates in end-to-end observation-driven architectures.
- Aleatoric (input) uncertainty: arises from noise and sparsity in the observations themselves, cannot be reduced by training on more data.
- Epistemic (model) uncertainty: reflects gaps in model knowledge, can be captured by using Monte Carlo Dropout at inference time [3].
- Probabilistic forecasts from end-to-end systems are essential for downstream decision-making in high-stakes applications (energy, disaster response).

Research Questions

- How can we model and quantify uncertainty in end-to-end AI weather models?
- What is the contribution of aleatoric and epistemic uncertainty on total forecast uncertainty?

Methodology

- *Data:* Reanalysis — **ERA5**, Satellite — **ASCAT** (surface wind), **AMSU-A/B**, **MHS**, **HIRS** (temperature and humidity profiles), **IASI** (hyperspectral), **GRIDSAT** (geostationary), In-situ — **HadISD** (land weather stations), **ICOADS** (ships), and **IGRA** (radiosondes). Training data is from 2007–2017, validation from 2020 and test dataset is 2019. Forecasts — **IFS ENS**.
- *Model:* Aardvark architecture [2] with modified Encoder (NoiseEncoder) inspired by AIFS CRPS [4].
- *Pre-training*: Train NoiseEncoder, Processor, individually, supervised by ERA5.
- *NoiseEncoder*: processes multi-source satellite and in-situ observations into n perturbed initial state estimates, trained via LW-CRPS to capture the ensemble spread of the atmospheric initial condition.
- *Fine-tuning*: auto-regressively finetune the pre-trained processor on the output of the NoiseEncoder, for each lead time (1 to 10 days).
- *Ensemble generation*:
  1. $n = 5$ NoiseEncoder, $k = 1$ MC Dropout passes (which we expect to capture aleatoric uncertainty);
  2. $n = 1$ NoiseEncoder, $k = 5$ MC Dropout passes (which we expect to capture epistemic uncertainty);
  3. $n = 5$ NoiseEncoder, $k = 5$ MC Dropout passes (which we expect to capture total uncertainty).
- *Metrics:* RMSE, CRPS, Ensemble Std, SSR.

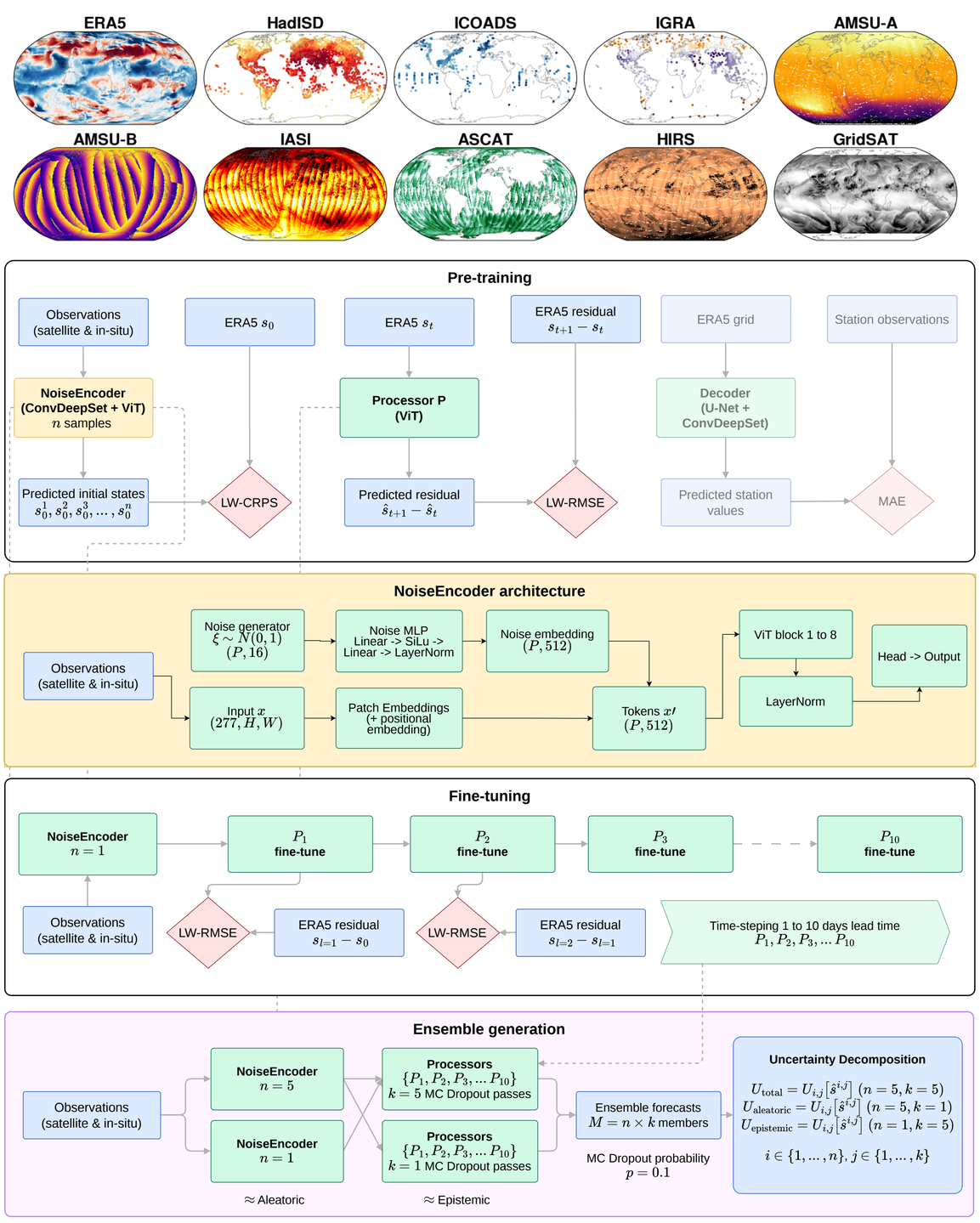

Figure 1. End-to-end pipeline for probabilistic weather forecasting, from multi-source data ingestion through ensemble generation and uncertainty decomposition. The introduced NoiseEncoder architecture is defined.

Preliminary Results

NoiseEncoder learns to produce meaningful spread.

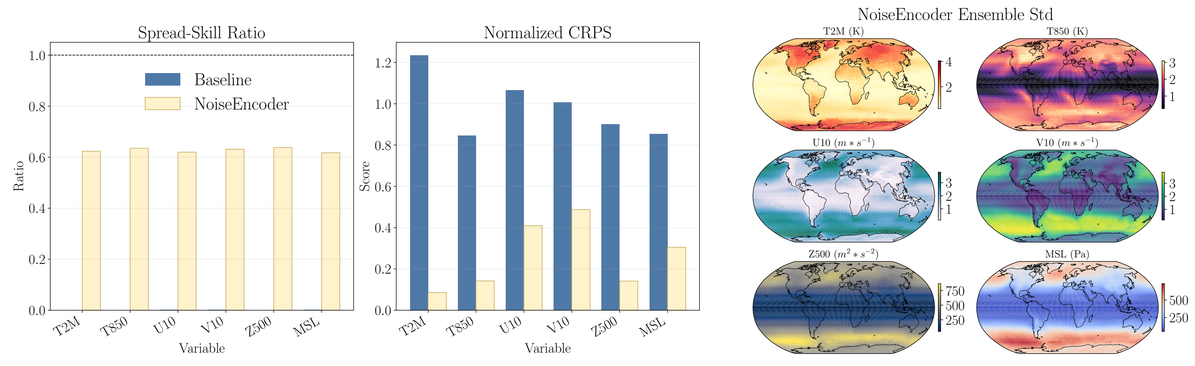

Figure 2. SSR, CRPS and Ensemble Std of NoiseEncoder. Baseline is direct noise injection in a deterministically trained model in the embedding space for test dataset.

Pre-trained Processor has high errors compared with NoiseEncoder Std, and original Aardvark paper.

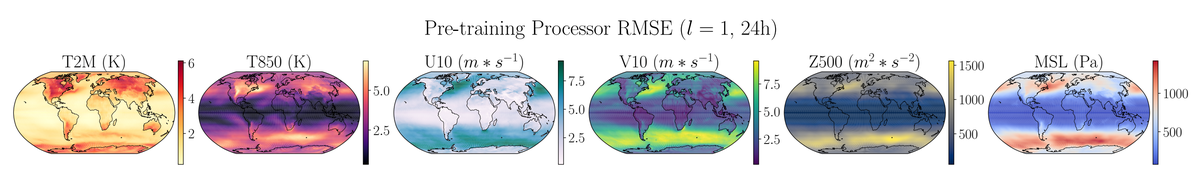

Figure 3. RMSE of the pre-trained processor for test dataset.

Aleatoric dominates spread (≈ 44% of total SSR at 5 days) while total uncertainty is smaller than sum of components (≈ 82% of sum) – this suggests uncertainty entanglement.

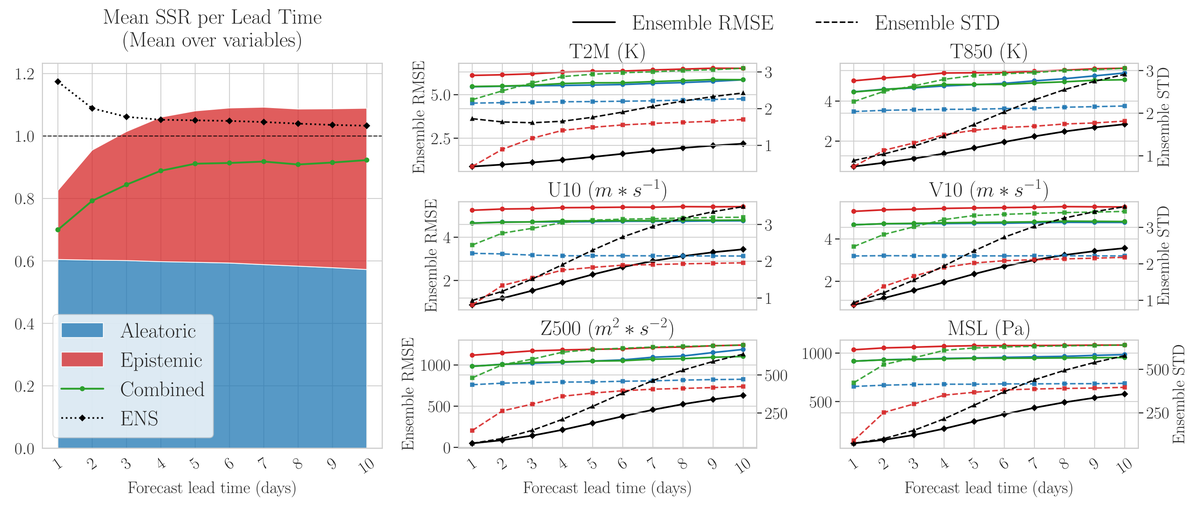

Figure 4. Latitude weighted mean SSR (across variables), Ensemble RMSE and Std of aleatoric, epistemic and combined ensembles for test dataset.

Aleatoric is stronger in spread in the poles. Epistemic increases overall with lead time.

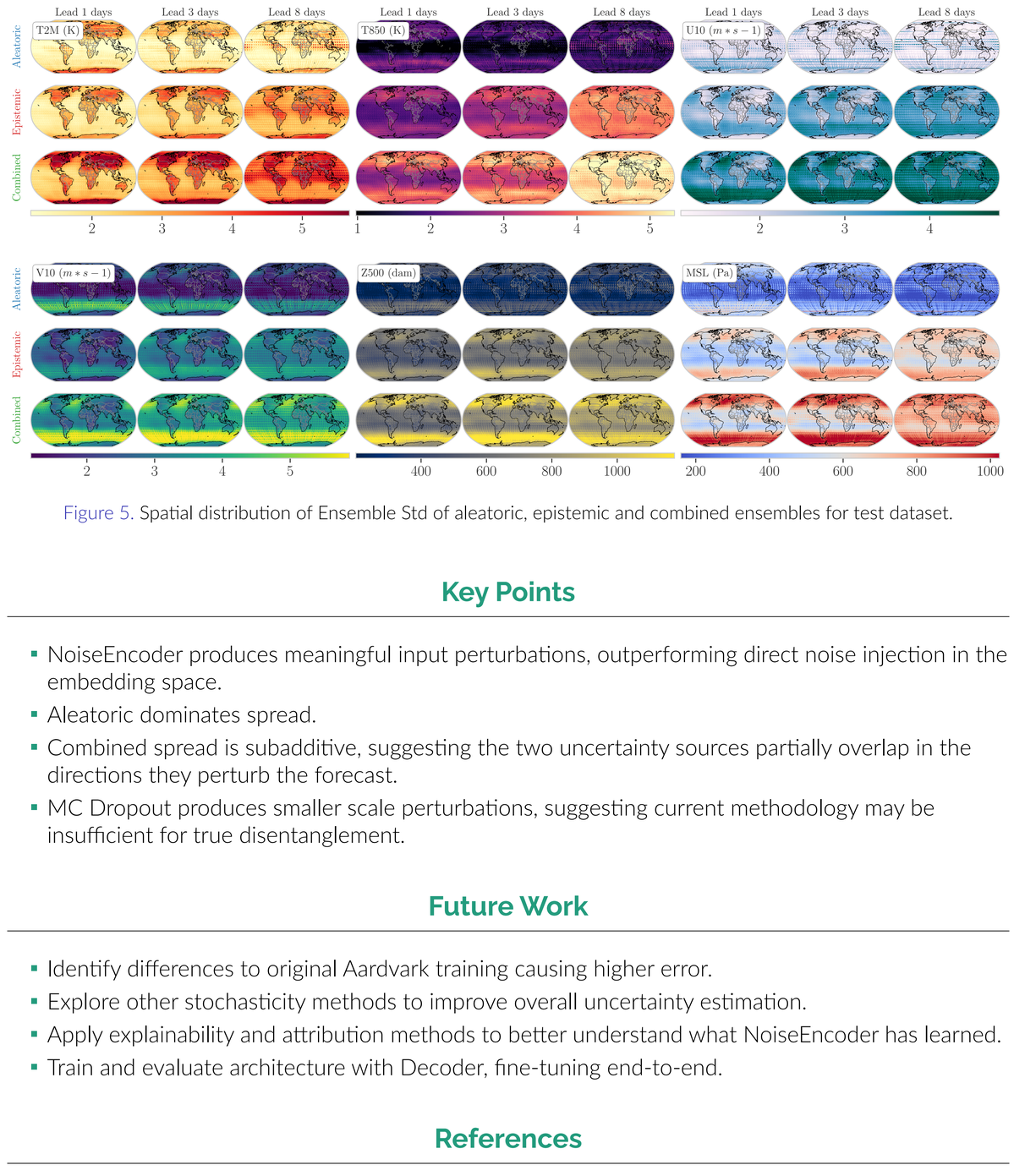

Figure 5. Spatial distribution of Ensemble Std of aleatoric, epistemic and combined ensembles for test dataset.

Key Points

- NoiseEncoder produces meaningful input perturbations, outperforming direct noise injection in the embedding space.
- Aleatoric dominates spread.
- Combined spread is subadditive, suggesting the two uncertainty sources partially overlap in the directions they perturb the forecast.
- MC Dropout produces smaller scale perturbations, suggesting current methodology may be insufficient for true disentanglement.

Future Work

- Identify differences to original Aardvark training causing higher error.
- Explore other stochasticity methods to improve overall uncertainty estimation.
- Apply explainability and attribution methods to better understand what NoiseEncoder has learned.
- Train and evaluate architecture with Decoder, fine-tuning end-to-end.

References

[1] Stephan Rasp et al. *WeatherBench 2: A Benchmark for the next Generation of Data-Driven Global Weather Models*. Jan. 2024.

[2] Anna Allen et al. "End-to-End Data-Driven Weather Prediction". In: *Nature* 641.8065 (May 2025), pp. 1172–1179.

[3] Katherine Haynes et al. "Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications". In: *Artificial Intelligence for the Earth Systems* 2.2 (Apr. 2023).

[4] Simon Lang et al. "AIFS-CRPS: Ensemble Forecasting Using a Model Trained with a Loss Function Based on the Continuous Ranked Probability Score". In: *npj Artificial Intelligence* 2.1 (Feb. 2026), p. 18.