Generative AI for Weather Data Assimilation

Ruizhe Huang[1], Qidong Yang[3], Jonathan Giezendanner[2], Sherrie Wang[1,2]

Email: {rzhuang, qidong, jgiezend, sherwang}@mit.edu

[1]Department of Mechanical Engineering, MIT
[2]Institute for Data, Systems, and Society, MIT
[3]Electrical Engineering and Computer Science, MIT

Summary

Problem: Standard ERA5 weather data is too "smooth" and often fails to capture real, local weather variations seen at ground stations.

Method: We compare multiple GenAI approaches for weather data assimilation.

Result: Guidance++ matches the best performance while requiring far less compute. Guidance++ successfully corrects known ERA5 biases (like wind near lakes and temperature in mountains) and achieves a mean improvement of 20.7% at random U.S. locations compared to the ERA5 baseline. Guidance++ effectively refines the ERA5 data locally and captures station variance by assimilating observations from thousands of local weather stations.

1. Motivation

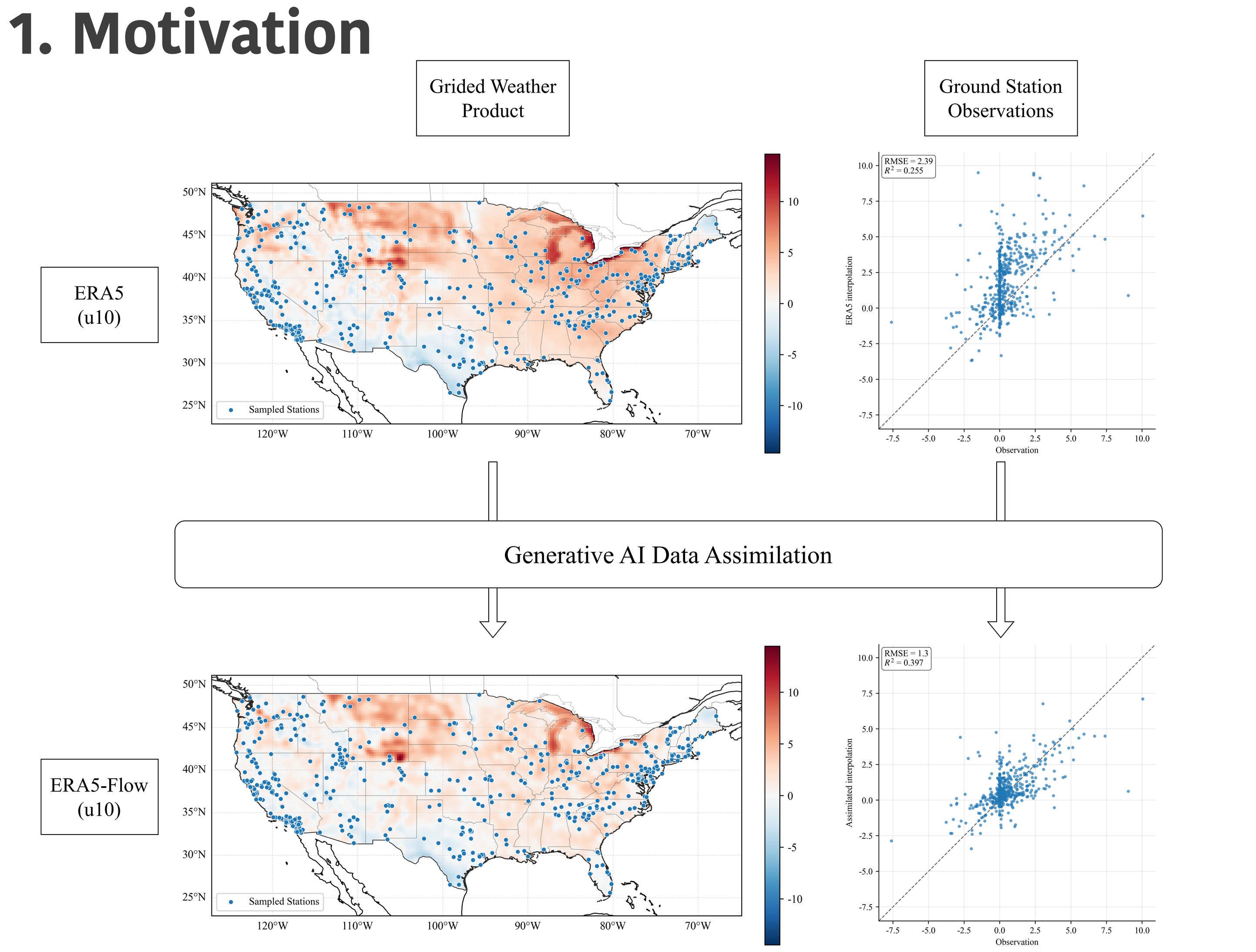

Figure 1. ERA5 data and weather station observations exhibit foundational distribution misalignments. ERA5 is known for its spatial smoothness and often fails to reflect true local values, while weather stations exhibit greater variance. Can we, using station observations, adjust ERA5 to better reflect these station-level variables? Furthermore, can we downscale ERA5 to capture this station variance within its native grid cells?

2. Method and Datasets

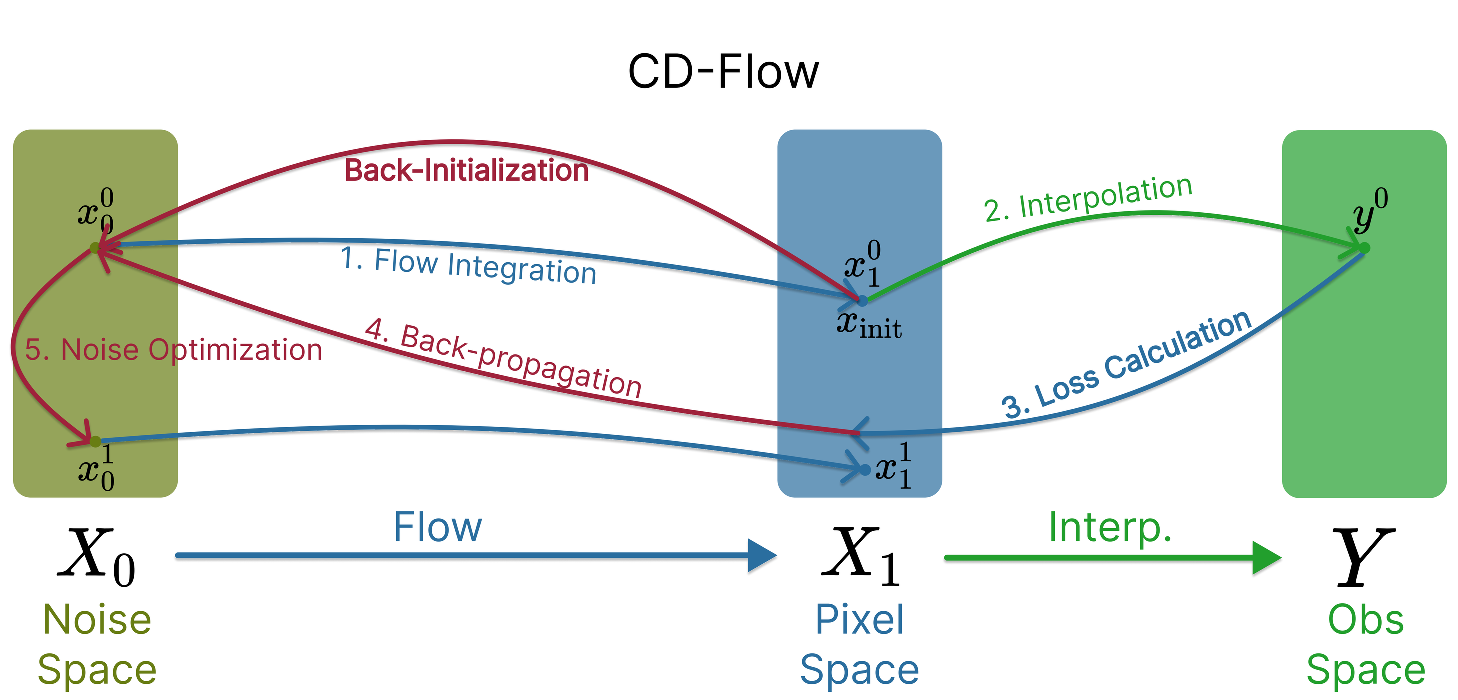

Figure 2. Overview of CD-Flow, an improved version of D-Flow. CD-Flow optimizes the initial noise x0 through end-to-end backpropagation.

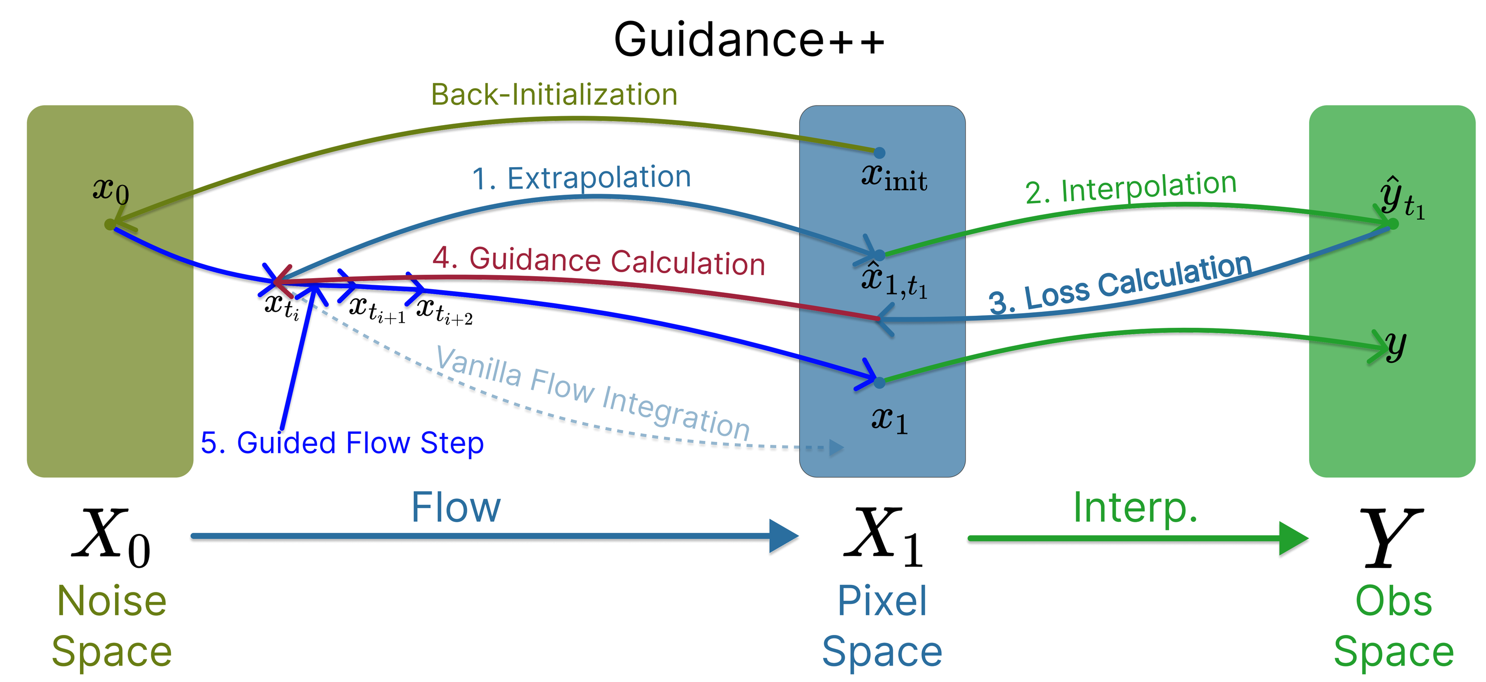

Figure 3. Overview of Guidance++, an improved version of Guidance. Guidance++ corrects velocity during ODE integration.

The Flow Model was trained on continental U.S. ERA5 (2019–2023), assimilating observations from 8,294 training weather stations, and evaluated on 1,778 separate test stations.

3. Benchmark

Method	Batch Size	u10	v10	t2m	d2m	Average
ERA5 (Baseline)	—	Reference	Reference	Reference	Reference	Reference
Vanilla Guidance	4	23.16 ± 10.03	23.82 ± 10.01	−138.64 ± 80.17	−113.53 ± 85.79	−51.30 ± 38.04
Guidance++	4	**33.15 ± 5.46**	**35.05 ± 5.93**	26.02 ± 4.57	30.72 ± 6.04	31.24 ± 3.10
Vanilla D-Flow	4	26.24 ± 6.87	28.22 ± 7.41	20.40 ± 5.99	24.77 ± 7.02	24.91 ± 4.27
CD-Flow	4	32.34 ± 5.70	34.49 ± 6.24	**27.16 ± 4.69**	**31.28 ± 6.12**	**31.32 ± 3.25**

Table 1. RMSE change (%) relative to ERA5 (mean ± std across runs) on test stations. Positive = lower RMSE than ERA5 (improvement); negative = higher RMSE (degradation). "Average" is the unweighted mean across variables.

Method	Batch Size	Training		Assimilation		
		Mem (GB)	Time (h)	Mem (GB)	Time (s)	Total Time (s)
Vanilla Guidance	4	37.96	47.50	0.785 ± 0.060	127.04 ± 57.12	11687.7
Guidance++	4	37.96	47.50	**0.106 ± 0.007**	**2.72 ± 0.16**	**250.2**
Vanilla D-Flow	4	37.96	47.50	37.756 ± 3.708	308.06 ± 13.37	28341.5
CD-Flow	4	37.96	47.50	41.425 ± 3.787	322.09 ± 14.05	29632.3

Table 2. Compute cost for pretraining and inference-time assimilation. Values are Mean ± Std; lower is better; best per column in bold. Guidance++ is able to assimilate 365 snapshots within 5 mins.

4. Guidance++ Captures Station Variance

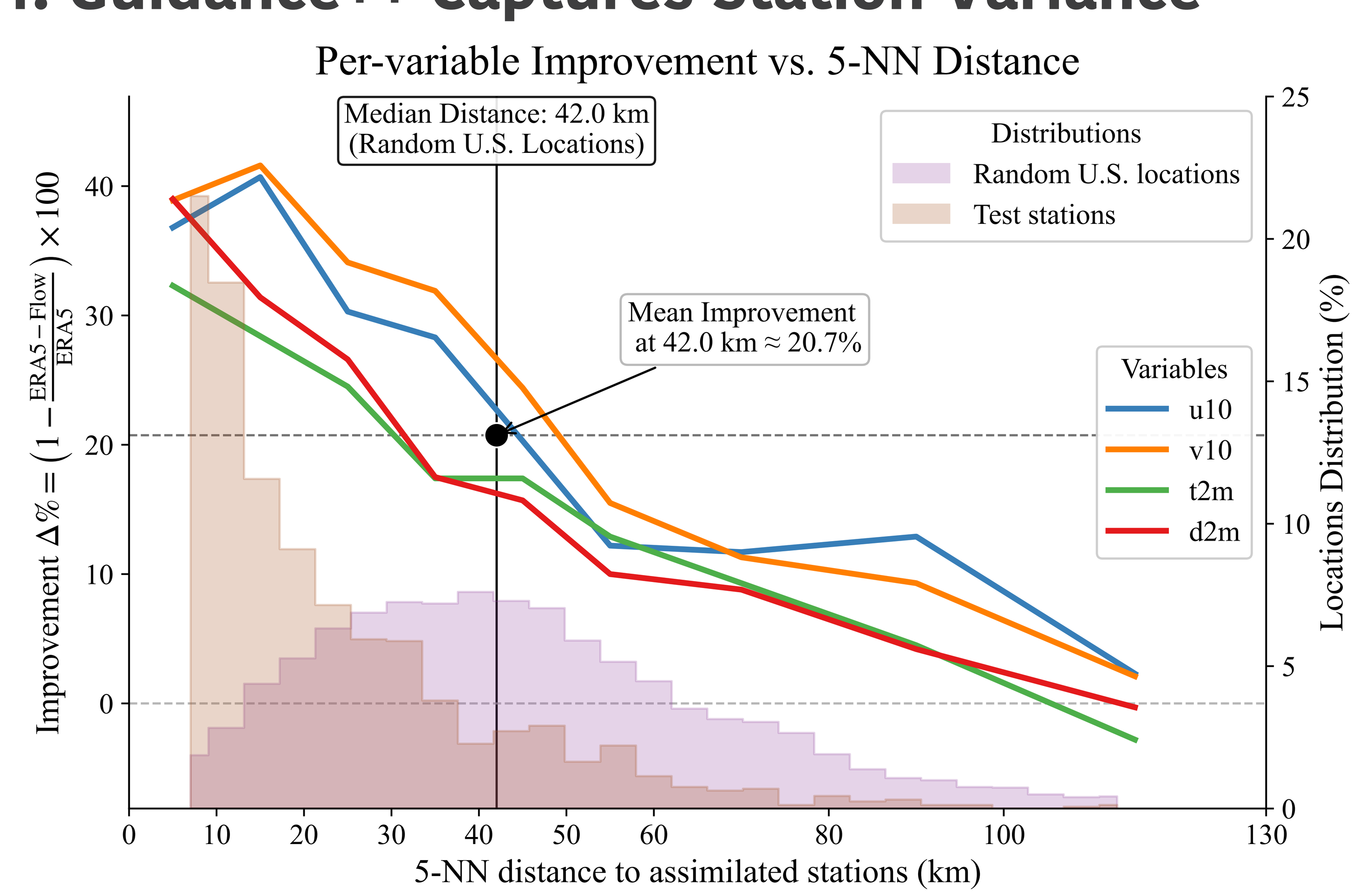

Figure 4. Per-variable improvement vs. 5-NN distance to assimilated stations. Lines show u10, v10, t2m, and d2m relative RMSE improvement $\Delta\% = 100\,(1 - \text{ERA5-Flow}/\text{ERA5})$ as a function of mean 5-NN distance. Improvements are highest near stations and decay with distance; winds decline more gradually than temperatures. The annotation marks the median distance for random U.S. locations (42.0 km) and the corresponding average improvement (~20.7%). This confirms that Guidance++ captures station variance, demonstrating its promise for downscaling.

5. Guidance++ Refines ERA5-Flow Locally with Observations

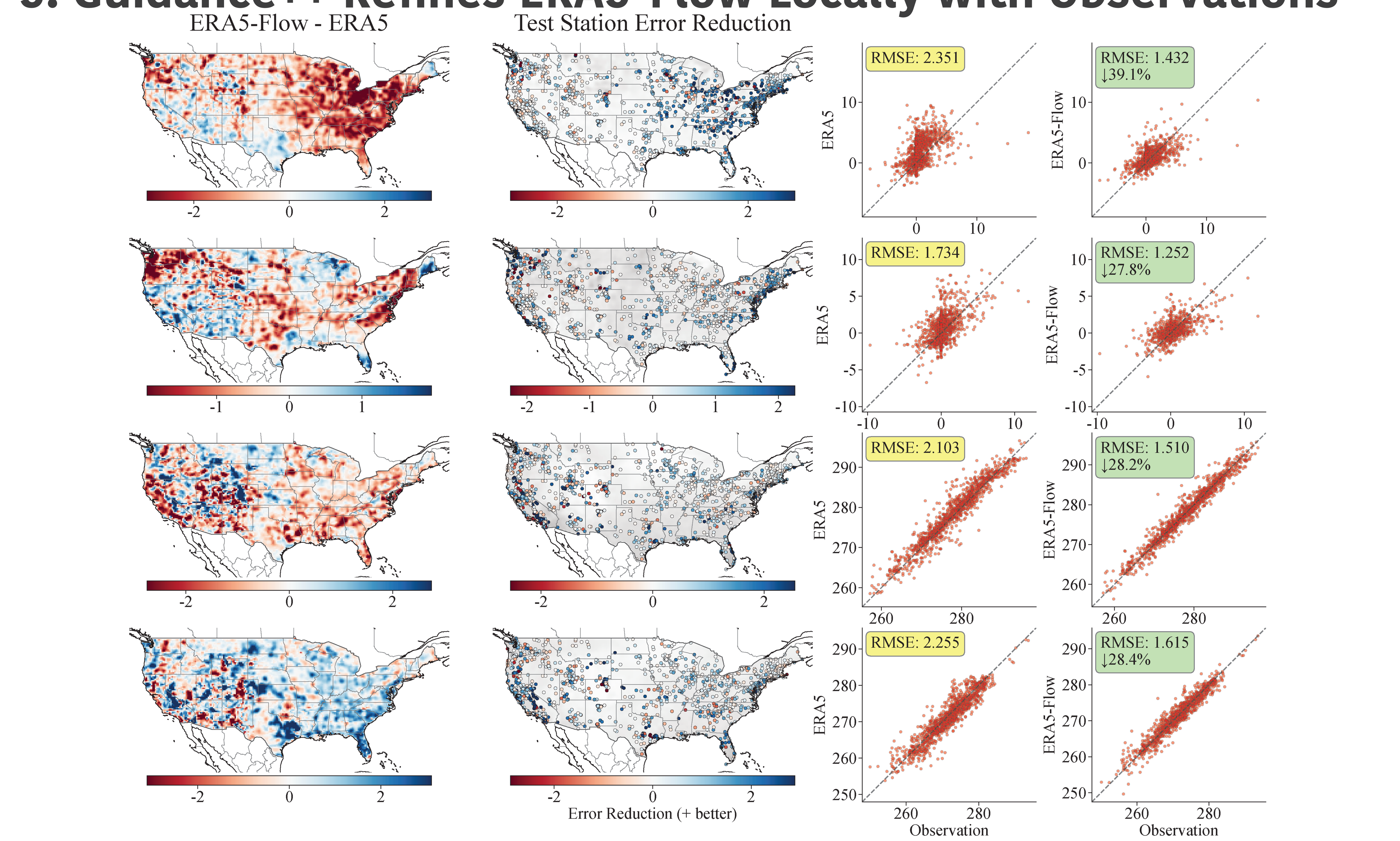

Figure 5. Guidance++ guides ERA5-Flow to more closely match observations from assimilated weather stations locally. This leads to improvements at test stations due to spatial correlation between assimilated and test stations.

6. Case Study - Where Guidance++ Corrects the Most

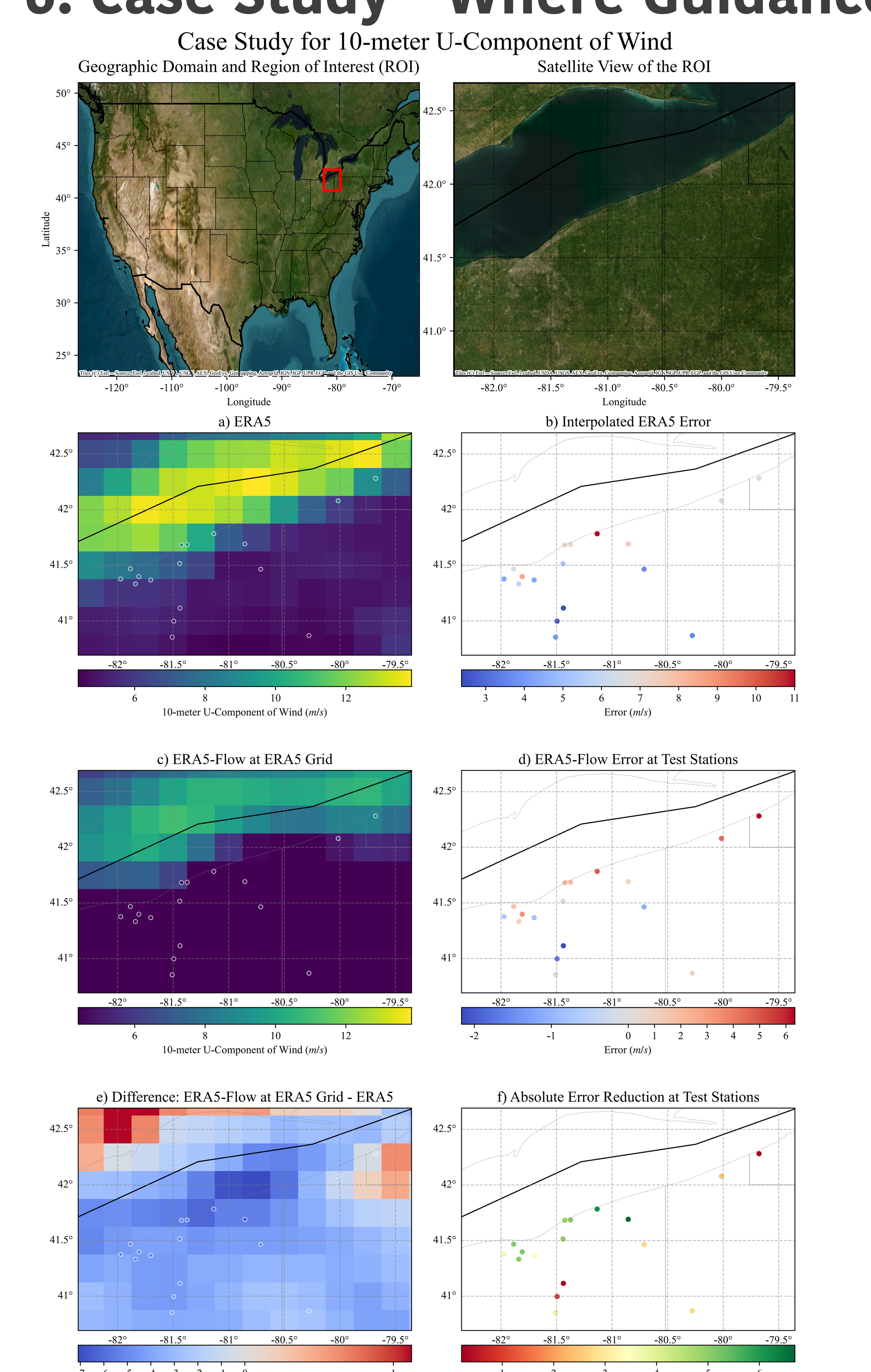

Figure 6. Case study for 10-meter horizontal wind (u10) on January 1, 2020, at 00:00 UTC. This figure highlights the area with the most correction for u10, located around the Great Lakes region. Large wind errors were identified along the shore, which are corrected by Guidance++.

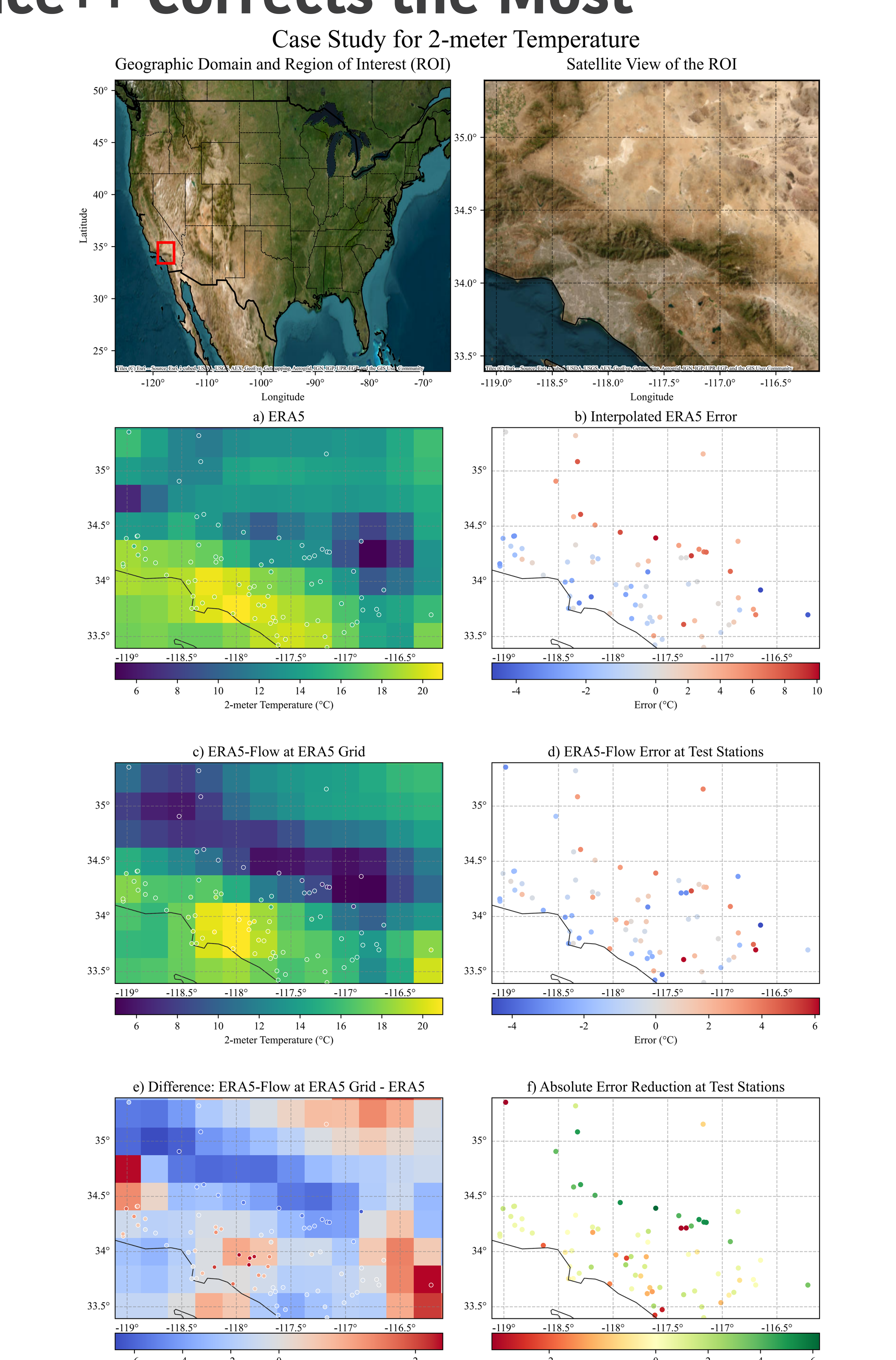

Figure 7. Case study for 2-meter Temperature (t2m) on January 1, 2020, at 00:00 UTC. This figure highlights the area with the most correction for t2m, located around Los Angeles. Because of the region's proximity to the sea and mountains, the temperature varies greatly with space. In the ERA5 data, mountains are not cold enough and cities are not hot enough; these biases are corrected by Guidance++.

7. Future Work

Downscaling ERA5 with foundation model embeddings and satellite images

Multi-sensor data assimilation