Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Partial recovery of meter-scale surface weather

Sherrie Wang

sherwang@mit.edu

Massachusetts Institute of Technology https://orcid.org/0000-0002-4618-5675

Jonathan Giezendanner

Massachusetts Institute of Technology

Qidong Yang

Massachusetts Institute of Technology

Ruizhe Huang

Massachusetts Institute of Technology

Eric Schmitt

Shell Information Technology International Inc. https://orcid.org/0009-0009-1642-7934

Anirban Chandra

Shell Information Technology International Inc. https://orcid.org/0000-0002-9120-3051

Yawen Zhang

Allen Institute for AI

Jeremy Vila

Shell Information Technology International Inc.

Detlef Hohl

Shell Global Solutions Inc.

Campbell Watson

IBM Research

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8931286/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Partial recovery of meter-scale surface weather

Jonathan Giezendanner[1†], Qidong Yang[1†], Ruizhe Huang[1†],
Eric Schmitt[2], Anirban Chandra[2], Yawen Zhang[3], Jeremy Vila[2],
Detlef Hohl[2], Campbell Watson[4], Sherrie Wang[1*]

[1]Massachusetts Institute of Technology

[2]Shell Information Technology

[3]Allen Institute for AI

[4]IBM Research

[*]Corresponding author. Email: sherwang@mit.edu

[†]These authors contributed equally to this work.

Near-surface weather varies over tens to hundreds of meters, yet remains unresolved in analyses and forecasts. We test whether this variation can be inferred without resolving atmospheric dynamics. Combining sparse weather stations, high-resolution Earth observation, and coarse atmospheric dynamics, we infer temperature, dewpoint, and wind at 30-m resolution across the contiguous United States. Against measurements held out in space and time, estimates reduce error by 11–28% relative to the strongest baseline. Within held-out 0.25° grid cells, we recover more spatial variance than baselines, explaining nearly half of temperature variability in the median cell. The method captures time-varying differences between locations and produces coherent patterns associated with topography and land cover. Beyond weather, our findings illustrate how sparse observations of a dynamical system can be combined with dense observations of persistent environmental structure to recover otherwise unresolved spatial variability.

1 Introduction

Weather at the Earth's surface can change across meter scales: think of stepping from a shaded park into a sunlit street or breaking above the treeline into exposed wind. These small-scale variations — which we here refer to as "micro-weather" — are created by terrain, vegetation, buildings, and other surface features, and in turn impact human heat exposure (*1, 2*), renewable energy performance (*3*), wildfire propagation (*4, 5*), and pollutant dispersion (*6*). Despite their importance, micro-weather variations are not captured by numerical weather predictions (NWP) and reanalysis products at regional or global scales (*7, 8, 9, 10, 11*), since NWP, along with their recent machine-learning emulators (*12, 13, 14, 15, 16, 17, 18, 19*), are run at grid spacings of several kilometers (Fig. 1a). This coarse spatial resolution is likely to persist for some time as models continue to contend with computational limits and the breakdown of parameterizations at high resolution (*20, 21, 22*). At the time of writing, for example, the emerging digital twin initiative at ECMWF aims to simulate the atmosphere at ∼1 km resolution (*23, 22*).

Yet resolving fine-scale atmospheric dynamics may not be necessary to recover all fine-scale weather variability. Near-surface conditions are shaped not only by turbulent atmospheric processes but also by systematic interactions between the atmospheric state and persistent surface characteristics such as topography, vegetation, and buildings. Recent machine-learning systems already exploit such relationships at coarser scales: they forecast weather at individual station locations from gridded and observational inputs (*24, 25*), and densify sparse station targets into kilometer-scale forecast fields (*26*). These systems, however, predict at instrumented sites or on kilometer-scale grids, leaving open whether variation between stations at finer scales can be recovered. This raises the question: how much weather variability below the resolution of current models can be inferred from existing observations without explicitly resolving fine-scale atmospheric dynamics? We refer to this component as the "statistically recoverable component of micro-weather".

Our ability to address this question is enabled by the fact that Earth observation now resolves the land surface at meter to tens-of-meter resolution globally. Although these images do not directly encode the dynamic atmospheric state, multispectral satellite imagery, land-cover classifications, and digital elevation models capture properties that shape micro-scale atmospheric behavior (*27, 28, 29*). For example, wind must maneuver around terrain, vegetation has a cooling effect, and surface materials differ in moisture retention.

We show that a substantial portion of micro-weather is statistically recoverable by combin-

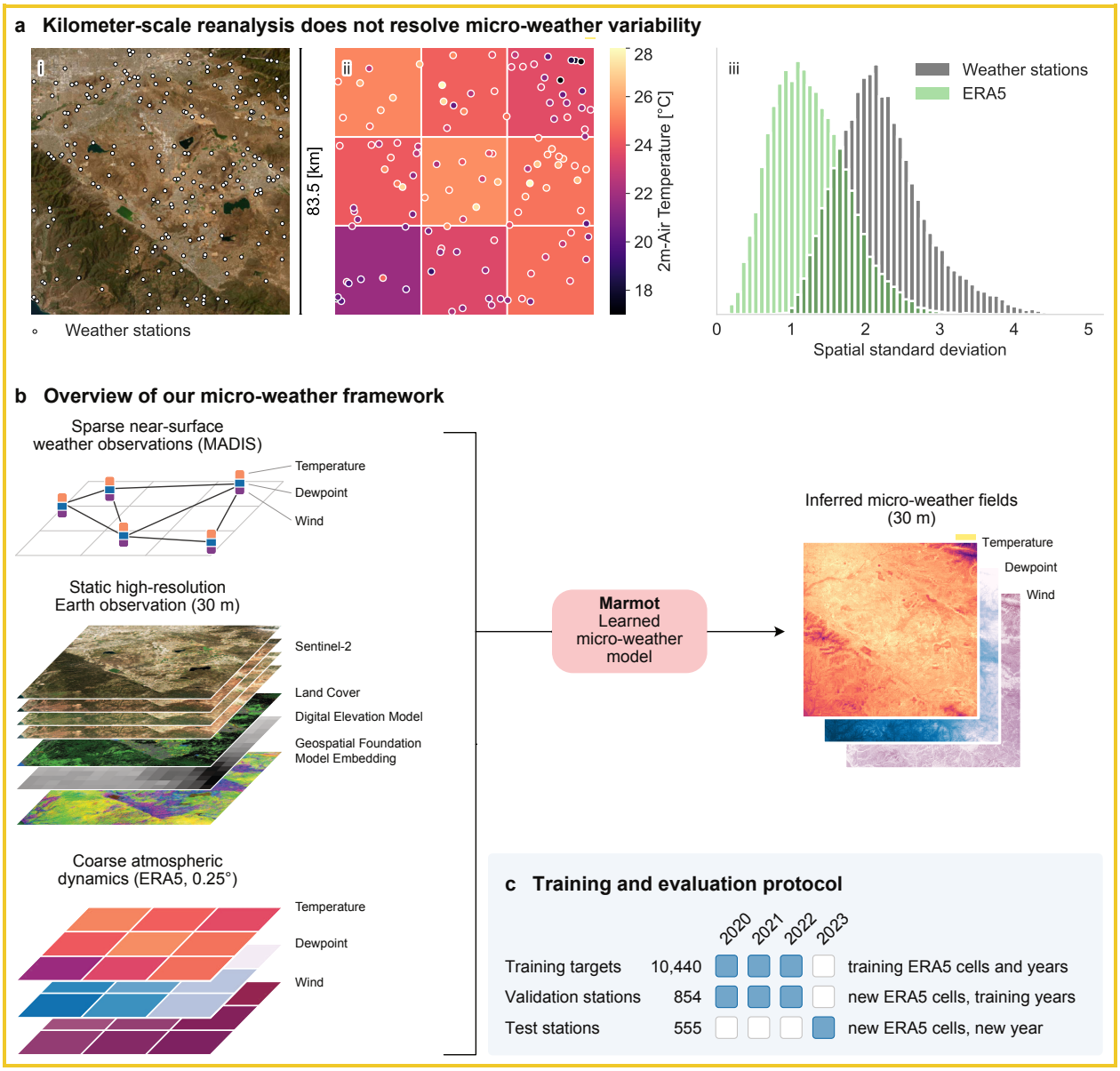


Figure 1: **Framework combining surface stations, Earth observation, and reanalysis to recover sub-kilometer weather variability. (a)** (i) Example over the Los Angeles metropolitan area illustrating that kilometer-scale reanalysis (ERA5) does not resolve meter-scale weather variability. (ii) 2 m air temperature from ERA5 (grid) and as measured by surface weather stations (points). (iii) Histogram of the spatial standard deviation of 2 m air temperature, 2020–2023, across the weather stations versus across the nine ERA5 grid cells. **(b)** Overview of the inference framework. The model combines observations from sparse surface stations, high-resolution Earth observation data, and coarse atmospheric dynamics from ERA5 to generate spatially continuous 30 m resolution fields of 2 m air temperature, dewpoint temperature, and 10 m wind. **(c)** Training and evaluation protocol. Stations are split into backbone, training, validation, and test sets by whole 0.25° ERA5 cells; backbone and training stations serve as training targets. Filled squares mark the years each set contains; test stations are new to the model in both space and time.

ing sparse surface observations with land surface characteristics and ERA5 reanalysis. We train **Marmot**, a multimodal deep learning model, to predict observations at more than 10,000 weather stations across the contiguous United States. Then, at inference time, we generate spatially continuous near-surface micro-weather fields at 30 m resolution. We evaluate Marmot against millions of observations in entirely held-out ERA5 grid cells and find that it reduces error relative to statistical and dynamical baselines, including station interpolation, ERA5-MOS, and HRRR. To distinguish sub-grid recovery from improvements in aggregate accuracy, we develop an evaluation framework based on observed spatial variance within coarse grid cells and contrasts between held-out station pairs as a function of their separation. Under these tests, Marmot recovers substantially more sub-grid variation than the baseline methods, with evidence of temperature recovery in some sub-kilometer station pairs. These recovered contrasts change through time, including reversals in which location is warmer or more humid, rather than reflecting persistent differences between locations. Improvements also extend into the tails of the weather distribution, with particularly strong performance for extreme heat and humidity.

Despite being trained on sparse station measurements, the inferred fields also exhibit physically interpretable and coherent patterns at meter scales, including nighttime temperature inversions, humidity gradients due to crop irrigation, and wind channeling with terrain. Independent comparisons with land surface temperature observations from the ECOSTRESS sensor lend further support to the spatial structure of the inferred air temperature fields (*30*). Ablation experiments reveal that nearby weather stations provide a strong estimate of conditions at unobserved locations, while explicit topographic relationships and high-resolution Earth observation enable additional spatial variation to be recovered; coarse reanalysis contributes comparatively modest additional information. Together, these results show that near-surface micro-weather contains a component that can be inferred from existing observations without explicitly modeling fine-scale atmospheric dynamics. More broadly, our work demonstrates how sparse ground observations can be densified with fine-scale environmental information to recover previously unresolved variability.

2 The micro-weather model

We develop Marmot, a transformer-based model that infers near-surface weather at arbitrary target locations from three complementary sources of information (Fig. 1b). First, a backbone network of surface weather stations from the NOAA Meteorological Assimilation Data Ingest System (MADIS)

provides spatially sparse, point-based observations of 2 m temperature, 2 m dewpoint, and 10 m wind. Second, high-resolution Earth observation (EO) data characterizes fine-scale surface heterogeneity at station and target locations. Our primary model uses 30 m AlphaEarth foundation model embeddings, which summarize information from multiple EO modalities (*31*); experiments using conventional EO inputs are described in Sections 8.1 and 8.6. Third, hourly ERA5 reanalysis describes how the coarse weather field at $0.25° \times 0.25°$ resolution varies around the target location and how ERA5 differs from observations at nearby backbone stations (*32*).

Marmot first uses self-attention to integrate weather information across the backbone station network and then uses cross-attention with the target location to infer target weather conditions. The model explicitly represents spatial relationships between backbone stations and the target, including relative position, distance, and elevation difference, allowing it to learn how weather varies with geographic and topographic separation. At the target location, Marmot additionally receives ERA5 weather and high-resolution EO information describing the local environment. Additional architecture and training details are provided in Sections 8.2 and 8.3.

We evaluate Marmot on observations that are held out both spatially and temporally (Fig. 1c). To enforce spatial separation, ERA5 grid cells across the contiguous U.S. are partitioned into disjoint backbone, training, validation, and test sets (Fig. 7). Backbone stations provide contemporaneous weather context during both training and inference, while stations in training and backbone cells provide supervision during model training. Validation cells are used for model selection, and test cells are reserved exclusively for evaluation. Importantly, no weather observations from held-out stations are seen by the model at any time; predictions at test station locations rely solely on ERA5 and EO features at the target location and observations from the backbone network. To simultaneously enforce temporal separation, the model is trained using observations from 2020–2022 and evaluated only on 2023 observations from these spatially held-out test cells. Thus, every test observation is from both a location and a time period unseen during training.

At inference time, Marmot can be queried at arbitrary locations and timesteps by combining the backbone station observations with EO features at the target location and coarse atmospheric information from ERA5. This enables the generation of spatially continuous micro-weather fields at 30 m resolution over large geographic domains. These fields estimate the component of micro-weather that is statistically recoverable from sparse weather observations, fine-scale environmental information, and coarse atmospheric context.

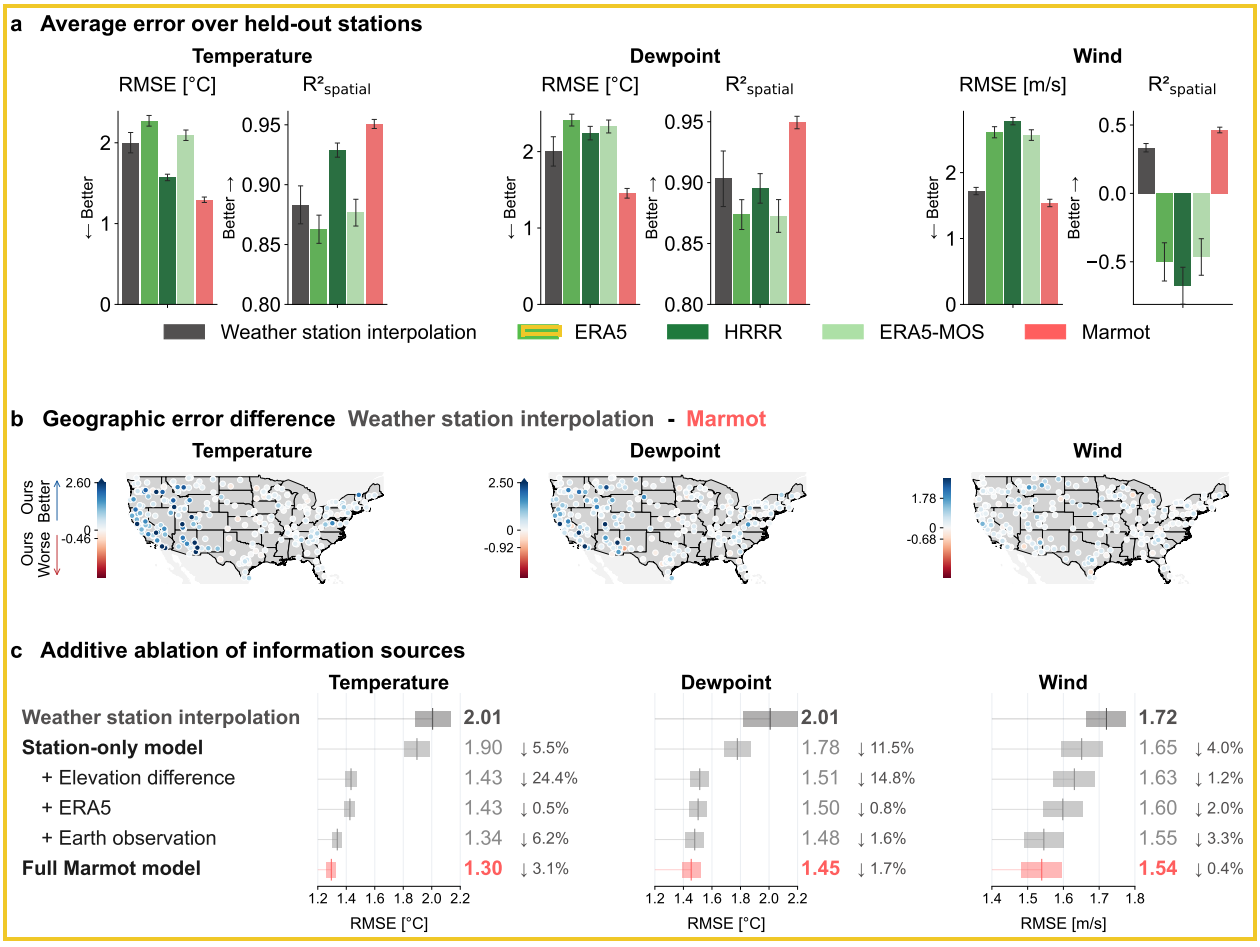


Figure 2: **U.S.-wide evaluation of inferred micro-weather accuracy relative to baseline methods.** **(a)** Root mean square error (RMSE) and spatial coefficient of determination (R^2_{spatial}, de-meaned per timestep to isolate spatial variability) evaluated on 2023 observations from stations in spatially held-out test grid cells, such that all evaluation observations are held out both in space and time. Marmot is compared with station interpolation, ERA5 reanalysis, HRRR analysis, and ERA5-MOS. Error bars denote one standard deviation across bootstrap replicates, each resampling both the held-out test grid cells and the five trained model seeds. **(b)** Map of station-level error reduction by Marmot relative to station interpolation, averaged over 2023. Positive values indicate improvement over station interpolation. **(c)** Additive ablation of information sources in Marmot. Model specifications progressively add information to station interpolation; values report RMSE and the percent reduction in RMSE relative to the preceding specification.

3 Improved near-surface weather estimation

We compare Marmot against four baselines chosen to represent distinct approaches to estimating local weather conditions: interpolation of station observations; globally-available ERA5 reanalysis, which provides Marmot's coarse reanalysis input; ERA5-MOS (*33*), a statistical post-processing model trained on station observations and surface attributes; and HRRR (*34*), an operational 3-km convection-allowing regional analysis built on the WRF-ARW dynamical core with hourly cycled data assimilation (Section 8.4).

Across the contiguous U.S., Marmot consistently improves near-surface weather estimation relative to these baselines. Averaged over all held-out test stations and time steps, Marmot reduces 2 m air temperature RMSE by 18% relative to HRRR analysis, the strongest temperature baseline, and reduces 2 m dewpoint and 10 m wind vector RMSE by 28% and 11%, respectively, relative to station interpolation, the strongest baseline for those variables (Fig. 2a). Relative to ERA5, which is provided as an input to the model, temperature RMSE is reduced by 43%, dewpoint RMSE by 40%, and wind vector RMSE by 41%. Statistical post-processing improves ERA5 slightly, but ERA5-MOS remains less accurate than Marmot across all three variables. Marmot also achieves higher national R^2_{spatial} than baseline methods for all three variables (Fig. 2a), indicating that its improved pointwise accuracy arises not only from correction of systematic biases but also improved representation of spatial variability at fixed timesteps.

Performance gains are geographically widespread but heterogeneous (Fig. 2b). For temperature and dewpoint, improvements relative to station interpolation are largest across the western and eastern U.S. and smaller across much of the central U.S. For wind, error reductions are smaller and have less pronounced regional structure. The size of improvements also varies by land cover type: relative error reductions are largest for temperature and dewpoint in shrublands and for wind in forested and built-up areas, while the smallest reductions occur in cropland for temperature and dewpoint and in barren and shrubland areas for wind (Fig. S2). At the national scale, prediction error exhibits only a weakly positive relationship with distance to the nearest backbone stations (Fig. S1), suggesting that station proximity alone does not explain these patterns. The geographic and land-cover differences may instead reflect a combination of environmental heterogeneity, the information provided by the surrounding station network, and differences in the processes governing each weather variable.

We next examine which sources of information enable the model to estimate near-surface

weather between sparse observations (Fig. 2c; Table S3). Station interpolation itself provides a strong baseline across all three variables. A learned station-only model, which infers conditions at held-out locations from contemporaneous observations at neighboring backbone stations, improves upon interpolation (−7% RMSE averaged across the three variables). Explicitly providing the elevation difference between neighboring stations and the target produces a large reduction in error (−13% RMSE), particularly for temperature and dewpoint. Adding ERA5 provides a small additional improvement (−1% RMSE), whereas incorporating high-resolution Earth observation through AlphaEarth embeddings produces a more substantial reduction in error (−4% RMSE). The remaining components of the full Marmot model, including temporal encodings, provide modest additional gains (−2% RMSE).

These results indicate that station observations provide substantial information about conditions at unobserved locations, while fine-scale information about topography and surface characteristics enables recovery of additional spatial variation. The small marginal improvement from ERA5 is consistent with coarse atmospheric information being largely redundant with observations from the backbone station network. Fine-scale surface information remains informative even after explicit elevation differences are provided: removing either elevation difference or AlphaEarth embeddings from the full model increases error (Table S3). Conventional Earth observation inputs, including elevation, land cover, and Sentinel-2 imagery, also improve predictions but do not reproduce the full gains obtained with AlphaEarth embeddings (Table S4).

4 Recovery of sub-grid spatial weather variability

While reductions in pointwise error and national R^2_{spatial} indicate improved agreement with station observations, they do not establish that the model recovers weather at fine scales. To directly evaluate sub-grid spatial recovery, we measure Marmot's ability to capture observed spatial variance within entirely held-out ERA5 grid cells, in which no station observations are available to the model and the variance within each cell is unresolved by Marmot's coarse atmospheric input.

Using a de-meaned coefficient of determination R^2_{spatial} that isolates spatial variability at fixed time steps (Section 8.5), we find that Marmot explains a significantly larger fraction of observed sub-grid spatial variance than baseline methods for all three weather variables (Fig. 3a). Performance is especially strong for temperature; for the median held-out ERA5 grid cell, Marmot explains 47% of the spatial variance in temperature, compared to 16% for HRRR, which is the next-best

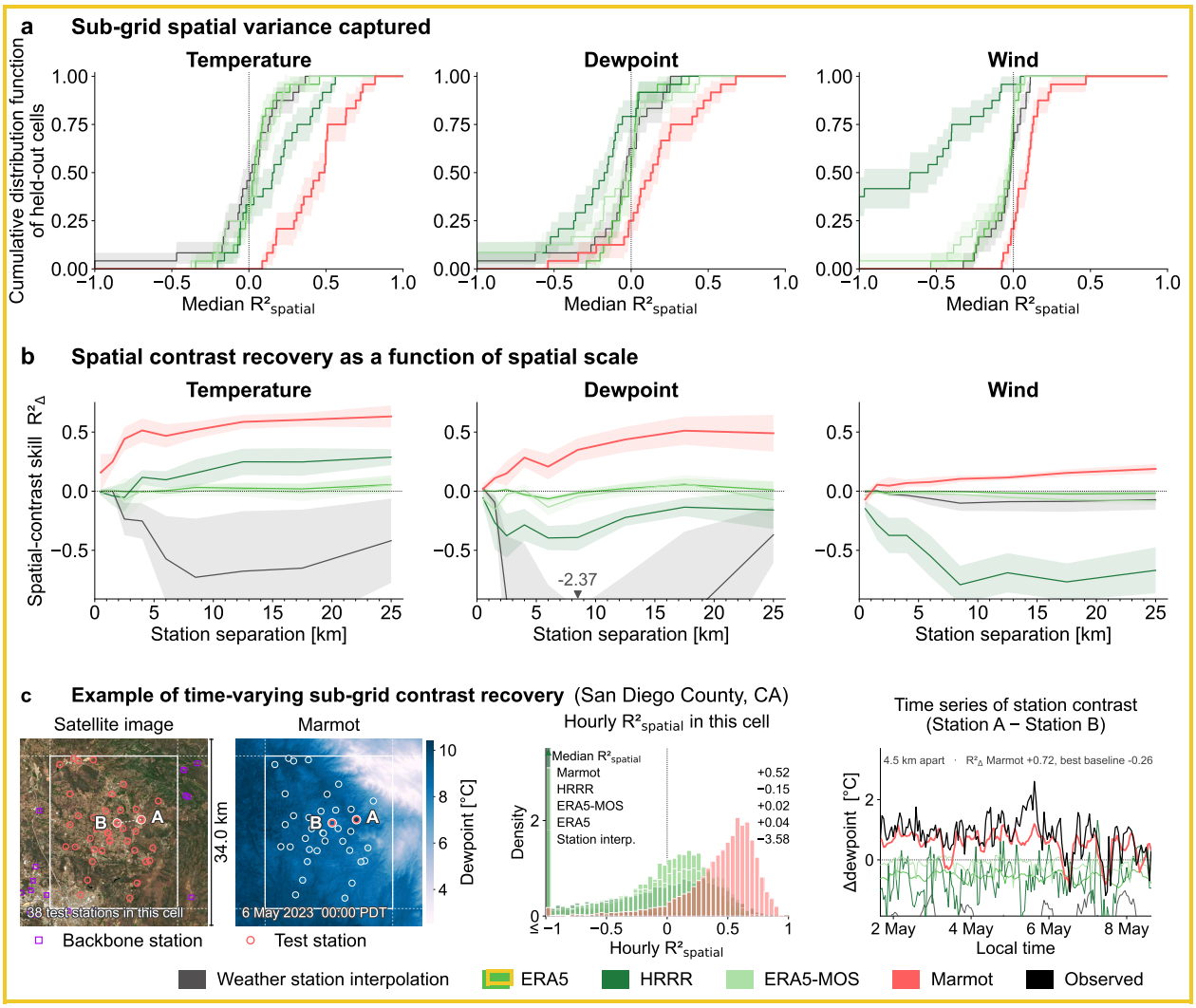


Figure 3: **Recovery of sub-grid spatial weather variability. (a)** Cumulative distribution functions (CDFs) of median R^2_{spatial} across held-out ERA5 grid cells. For each grid cell and hourly timestep, R^2_{spatial} measures the spatial variance captured by each method relative to observations across stations within the grid cell; values are then summarized per-cell by the median across hourly timesteps. **(b)** Recovery of spatial contrasts in near-surface weather as a function of spatial separation. Spatial contrast skill (R^2_Δ) measures how well each method recovers observed differences between pairs of held-out stations and is shown as a function of the distance between stations. **(c)** Example of time-varying sub-grid contrast recovery in San Diego County, California. The left panels show satellite imagery and Marmot's inferred dewpoint field within a held-out ERA5 grid cell containing 38 test stations. The center panel shows the distribution of hourly R^2_{spatial} within the cell over 2023. The right panel compares the observed and predicted dewpoint contrast between two held-out stations 4.5 km apart; Marmot captures the changing contrast with $R^2_\Delta = 0.72$, compared with -0.26 for the next-best baseline.

baseline. Dewpoint and wind variability is more challenging to recover, with Marmot achieving median R^2_{spatial} of 13% and 9%, respectively. Nevertheless, this is an improvement upon baseline methods, which have median R^2_{spatial} at or below zero for both variables, indicating that their sub-grid predictions are, on average, no better than assigning all stations within a grid cell a constant value. Notably, HRRR captures some sub-grid temperature variability, but it has the poorest performance for dewpoint and wind among the baselines.

Next, to determine how recovery varies with spatial scale, we evaluate contrasts between pairs of held-out stations as a function of their separation (Fig. 3b). Spatial contrast skill, R^2_Δ, measures recovery relative to predicting identical conditions at the two stations (Section 8.5). At separations of 1 km or greater, Marmot has the highest point estimate among the evaluated methods in every distance bin and for all three variables. In the 1–2 km bin, Marmot achieves R^2_Δ of 25%, 11%, and 6% for temperature, dewpoint, and wind, respectively. Below 1 km, the temperature point estimate remains positive ($R^2_\Delta = 16\%$), but uncertainty is large because such closely spaced station pairs are rare ($n \leq 13$); dewpoint and wind skill is near or below zero. Thus, contrast recovery generally declines at shorter separations and is substantially stronger for temperature than for dewpoint or wind. Pair-level analyses reveal substantial variation in recovery across locations (Fig. S5).

We further examine sub-grid spatial recovery in example grid cells for each weather variable. In Warren County, New Jersey, Marmot captures 50% of the observed within-cell spatial variance in temperature, compared with 15% for HRRR and 16% for ERA5-MOS (Fig. S3). The observed temperature contrast between two stations 12 km apart varies through time: Station A is warmer than Station B during the day but cooler at night. Marmot reproduces much of this changing spatial contrast ($R^2_\Delta = 53\%$), whereas the baseline methods do not capture the temporal pattern. In San Diego County, California, the most densely observed test grid cell with 38 held-out stations, Marmot captures 52% of the observed spatial variance in dewpoint, compared with 4% for the next-best method, ERA5 interpolation (Fig. 3c). Marmot likewise recovers the changing dewpoint contrast between two stations 4.5 km apart ($R^2_\Delta = 72\%$, versus -26% next-best). Wind remains more challenging: in Austin, Texas, Marmot captures little of the aggregate within-cell spatial variance (median $R^2_{\text{spatial}} = 3\%$), but recovers a substantial fraction of the time-varying wind-vector contrast between two stations ($R^2_\Delta = 41\%$, versus -6% next-best; Fig. S3). Together, these examples show that Marmot does not simply learn persistent offsets between locations, but recovers differences that evolve dynamically through time.

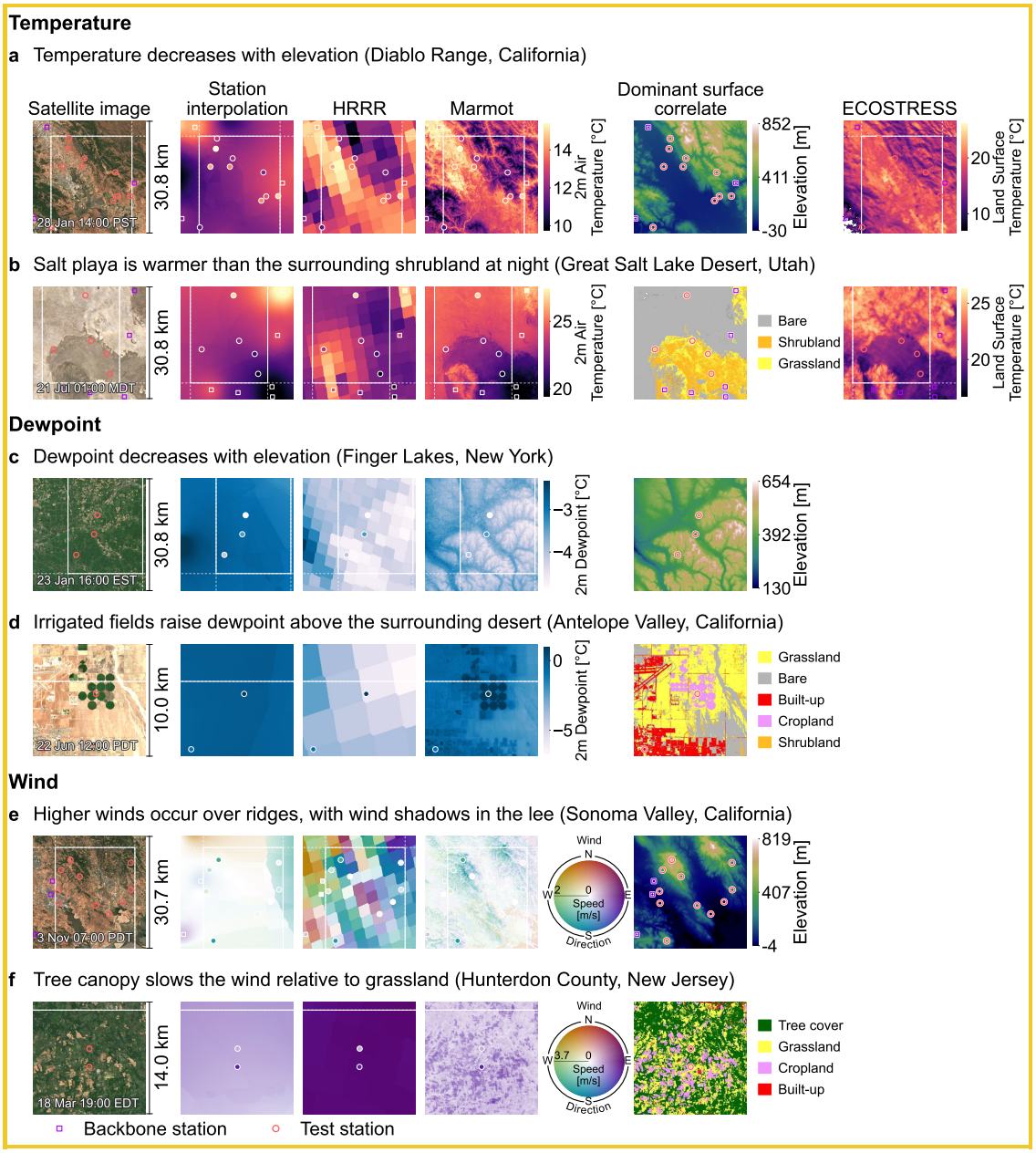


Figure 4: **Physical structure of inferred micro-weather fields.** Case studies comparing inferred 30 m resolution fields with station interpolation and HRRR. Satellite imagery[1] is shown for context, and terrain elevation or land cover is shown according to the dominant surface correlate of inferred spatial variability. For temperature panels, land surface temperature from ECOSTRESS is included for qualitative comparison. **(a)** Temperature decreases with elevation (Diablo Range, California). **(b)** Bare ground is warmer than the surrounding shrubland at night (Great Salt Lake Desert, Utah). **(c)** Dewpoint decreases with elevation (Finger Lakes, New York). **(d)** Irrigated cropland has higher humidity than the surrounding desert (Antelope Valley, California). **(e)** Higher winds occur over ridges; a southwesterly wind corresponds to windward-leeward contrasts in wind speed (Sonoma Valley, California). **(f)** Wind speed is higher over grass and cropland and lower over tree cover (Hunterdon County, New Jersey).

5 Physical structure of recovered micro-weather

Having established that Marmot recovers sub-grid spatial variability, we next examine whether the inferred fields exhibit physically coherent structure (Fig. 4). For a variety of case studies, we validate the learned spatial structure against held-out stations and examine the surface variable — either elevation or land cover — associated most strongly with Marmot's predicted field (Table S6; Section 8.5).

Marmot learns to recover the lapse rate effect. For instance, in the complex terrain of the Diablo Range surrounding Gilroy, California (Fig. 4a), inferred temperatures decrease with elevation over short horizontal distances. Although nearby backbone stations are located at low elevations, Marmot recovers cooler temperatures at high elevations, which is validated by held-out stations. HRRR captures similar broad spatial patterns but, at 3-km resolution, smooths over the narrow warmer valleys between peaks. Elevation explains 44% of the spatial variance in Marmot's inferred temperature field, compared with 20% for land cover (Table S6).

Land cover is more strongly associated with the inferred temperature structure in the Great Salt Lake Desert, Utah (Fig. 4b). At night, Marmot predicts warmer air temperatures over the barren salt playa and cooler temperatures over the surrounding shrubland, with temperature contrasts reaching several degrees Celsius and consistent with differences observed between held-out stations. An independent, contemporaneous ECOSTRESS land surface temperature (LST) observation exhibits the same fine-scale spatial pattern. Although LST and 2-m air temperature are distinct quantities, previous work has shown that they are strongly coupled at night and that nighttime LST captures spatial variation in air temperature associated with land cover (*35, 36*). HRRR, by contrast, predicts the opposite spatial contrast, with cooler temperatures over the playa than the shrubland.

For dewpoint, Marmot similarly recovers fine-scale gradients associated with topography and vegetation. In the hilly terrain of the Finger Lakes, New York (Fig. 4c), lower dewpoint values are inferred at higher elevations. These gradients are smoothed in HRRR and cannot be recovered through station interpolation given the sparsity of observations. In the arid Antelope Valley, California (Fig. 4d), Marmot delineates higher dewpoint over irrigated crop fields relative to the surrounding desert. HRRR appears to under-predict dewpoint across the broader desert region and does not resolve the more humid crop fields.

[1] All the high resolution image sources used for contextualization in the document were taken from Tiles © Esri – Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User

Lastly, fine-scale wind variability is more challenging to recover, but the inferred fields nevertheless exhibit coherent structure in both wind speed and direction. In the Sonoma Valley region of California (Fig. 4e), predicted wind speed varies with exposure, with higher wind speeds along ridgelines and lower speeds in valleys. Under predominantly southwesterly winds, Marmot predicts higher speeds on the southwestern sides of ridges and lower speeds on the northeastern sides. Despite the lack of a nearby backbone station, it also correctly predicts a shift to northerly winds observed at a held-out station approaching San Pablo Bay. HRRR exhibits much more varied wind directions across the region and over-predicts wind speeds. In the varied land cover of central New Jersey (Fig. 4f), the inferred field shows lower wind speeds over forested areas and higher speeds over grassland and cropland. These patterns are consistent with different surface roughness and drag coefficients across land cover types. Again, neither station interpolation nor HRRR can recover these patterns due to sparsity and coarseness, respectively.

We also find physically interpretable structure associated with temperature inversions, urban heat, and terrain-channeled wind across additional regions (Figs. S8 to S10). At the same time, Marmot does not recover micro-weather structure equally well across all conditions. Recovery is less reliable when nearby backbone observations provide limited information about conditions at the target location (Fig. S1). This occurs in some remote locations, where the nearest backbone stations are tens of kilometers away and experience very different conditions, and in nearshore settings, where backbone stations lie predominantly landward and poorly represent marine-influenced conditions (Figs. S11 and S12).

6 Performance under extreme weather

We next ask whether Marmot's improvements extend to extreme and rapidly changing weather, which have disproportionate impacts on human health and infrastructure despite their relative rarity. We evaluate performance across the tails of the observed weather distribution, at absolute extreme thresholds, and during rapid changes in weather (Section 8.5).

We first bin observations by their percentile within each held-out station's local distribution and compute conditional bias and RMSE within each bin. All three statistical methods — station interpolation, ERA5-MOS, and Marmot — exhibit regression toward the mean: low extremes are systematically over-predicted and high extremes under-predicted (Fig. 5a). However, Marmot ex-

Community

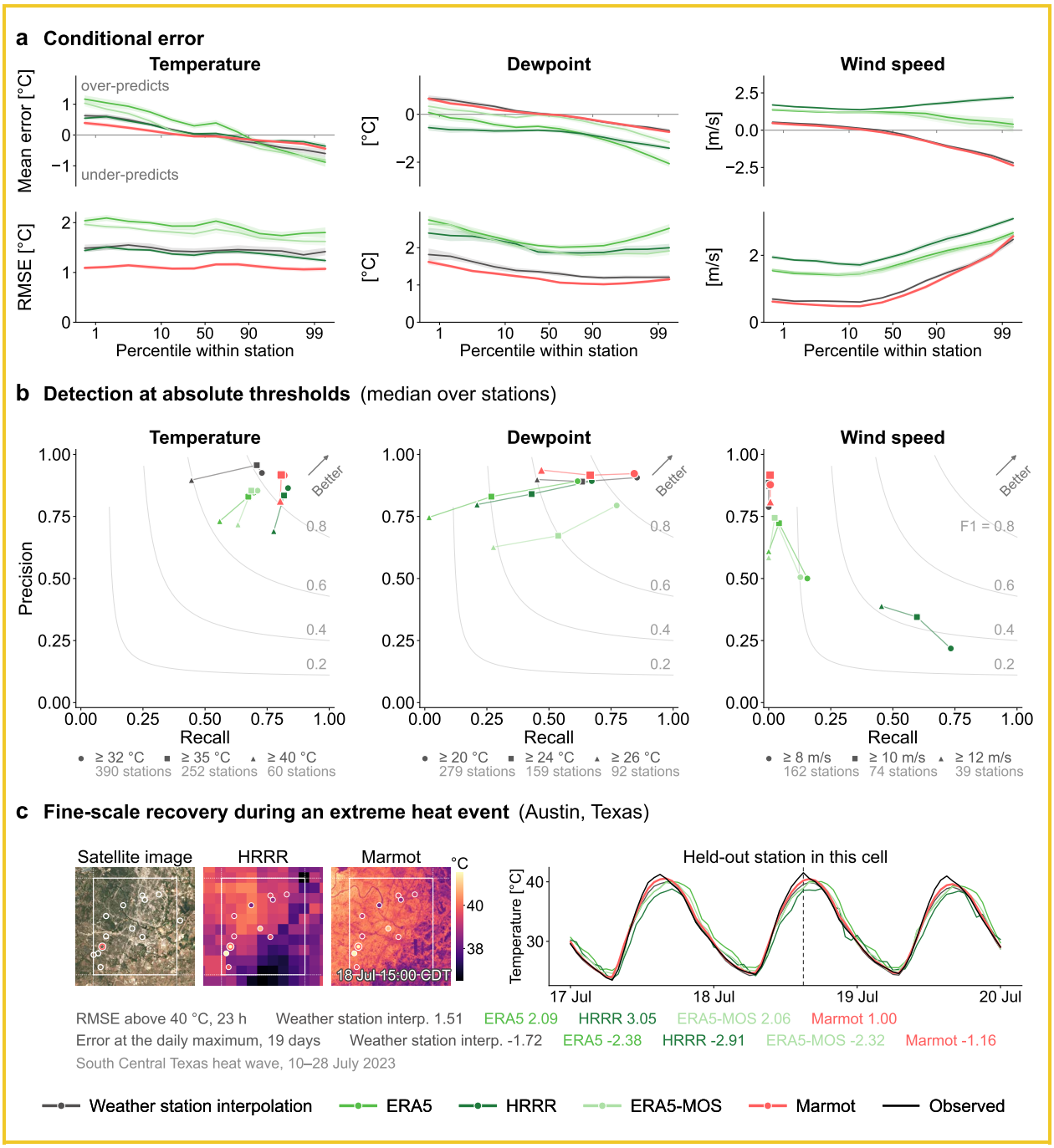


Figure 5: **Performance under extreme weather. (a)** Conditional mean error and RMSE as a function of observed percentile within each held-out station. **(b)** Precision and recall for detection of temperature, dewpoint, and wind speed extremes at fixed absolute thresholds, summarized by the median across stations. Points denote different thresholds; gray curves indicate lines of constant F1 score. **(c)** Fine-scale temperature recovery during the July 2023 South Central Texas heat wave. Maps compare HRRR and Marmot with held-out station observations; the time series shows observed and predicted temperature at a held-out station.

hibits the least severe over- and under-prediction for temperature, and among the least severe for dewpoint and wind. Temperature estimates from ERA5 and HRRR also regress toward the mean, but their errors for dewpoint and wind are dominated by overall negative and positive biases, respectively. Conditional RMSE is quite different across variables. For all methods, temperature error remains nearly uniform across the distribution, whereas dewpoint error increases toward the lowest percentiles and wind-speed error rises sharply toward the highest percentiles. Marmot achieves the lowest error for all three variables at nearly every percentile; the one exception is the highest wind-speed bin, where station interpolation is marginally better.

We next evaluate detection of extreme heat, humidity, and wind at fixed absolute thresholds (Fig. 5b). For all methods, detection becomes more difficult at increasingly extreme thresholds, with precision and/or recall declining as events become rarer. Nevertheless, Marmot lies on the precision-recall Pareto frontier for temperature and dewpoint across all thresholds, with its largest advantages over the baselines at the most extreme thresholds. For temperatures above 40 °C, for example, Marmot achieves an F1 score of 0.81, compared with 0.73 for the next-best method, HRRR. A similar advantage is observed for dewpoints above 26 °C.

The South Central Texas heat wave of July 2023 illustrates this advantage (Fig. 5c). All stations within the example grid cell are held out from Marmot. During temperatures above 40 °C, Marmot achieves an RMSE of 1.00 °C, compared to station interpolation at 1.51 °C and HRRR at 3.05 °C. All methods underestimate the daily temperature maximum, but Marmot has the smallest mean bias (−1.16 °C versus −1.72 °C for station interpolation and −2.91 °C for HRRR). The time series further shows that Marmot more closely tracks the magnitude and timing of the daily temperature peaks.

Wind is a different story. HRRR has by far the highest F1 score for detecting high wind events, with relatively high recall but low precision. Marmot, station interpolation, and ERA5-MOS instead have high precision but very low recall. In other words, Marmot rarely predicts extreme winds when they do not occur, but misses many observed high-wind events.

Finally, we evaluate rapid increases and decreases over three-hour intervals (Fig. S7). Ramp detection is more challenging than detection at absolute thresholds, especially for dewpoint and wind, although the relative behavior of the methods is similar. Marmot lies on the precision-recall Pareto frontier for temperature and dewpoint ramps. Wind ramps are difficult for all methods, with HRRR achieving the highest F1 scores through a similar high-recall, low-precision tradeoff.

7 Discussion

Our results demonstrate that a substantial component of near-surface micro-weather can be statistically recovered without explicitly resolving fine-scale atmospheric dynamics. Across the contiguous U.S., Marmot improves estimation of temperature, dewpoint, and wind relative to both statistical and dynamical baselines, and, critically, recovers observed spatial variance within ERA5 grid cells and between nearby stations. This recovery extends to time-varying weather contrasts between locations and to extreme heat and humidity. The inferred high-resolution fields exhibit physically coherent relationships with topography and land-surface characteristics, while the ablation shows that nearby station observations provide a strong foundation for inference and that elevation and high-resolution Earth observation contribute substantial additional information.

These results suggest a useful way to conceptualize the statistically recoverable component of micro-weather. Weather stations sparsely observe the instantaneous atmospheric state, while topography and Earth observation provide spatially dense information about how locations differ from one another. This combination allows spatial structure in near-surface weather to be inferred between observations. For example, stations may reveal that bare ground is warmer than nearby shrubland under a particular atmospheric state, while continuous Earth observation allows the same relationship to be applied at unobserved locations with similar surface characteristics. Although our analyses do not directly establish the learned mechanism, the Great Salt Lake Desert example (Fig. 4b), among others, is consistent with this interpretation. Coarse reanalysis provides additional atmospheric context, but its modest marginal contribution in our ablations suggests large redundancy with station observations for the present inference task. We note that this differs from forecasting, where future station observations are unavailable and dynamical forecasts must instead provide information about the evolving atmospheric state.

The extent of statistical recovery varies substantially by variable, spatial scale, and region. Temperature contrasts remain more recoverable than dewpoint or wind contrasts across spatial scales (Fig. 3b), but pair-level performance varies considerably across locations (Fig. S5b). At sub-kilometer separations, aggregate temperature-contrast skill remains positive but uncertain, whereas dewpoint and wind skill approaches zero. Rapid changes in dewpoint and wind are similarly challenging, and wind direction remains difficult even though Marmot improves over the baselines (Table S1). Geographic gains are also heterogeneous, with smaller improvements across parts of the central U.S. and in some sparsely observed or relatively homogeneous environments. These

patterns may reflect limited station information, insufficient environmental context, or variability generated by fine-scale atmospheric dynamics not represented in the model. We see these limitations in additional case studies, such as in terrain-confined valleys where Marmot only partially recovers observed wind channeling (Fig. S10), as well as in remote and nearshore locations, where nearby backbone stations poorly represent conditions at the target location (Figs. S11 and S12).

The comparison with HRRR provides further evidence that dynamical modeling is important for some of this remaining variability. HRRR resolves atmospheric dynamics at substantially finer scales than ERA5, whereas Marmot infers spatial variation from contemporaneous observations and environmental information. Their differing performance is particularly informative for wind. Marmot estimates average wind conditions more accurately and recovers some surface-associated spatial structure, but HRRR detects substantially more extreme winds and rapid wind changes. This suggests that higher-resolution dynamical modeling contains information about rapidly evolving weather that is absent from Marmot's current inputs. Conversely, Marmot could complement existing NWP diagnostics by revealing where regional models fail to resolve surface-associated spatial structure. To maximize micro-weather recovery, future work could apply a Marmot-type model to higher-resolution analyses or forecasts like HRRR.

The evaluation framework developed here separates aggregate accuracy from recovery of spatial variance within coarse grid cells. Its within-cell and pairwise metrics test whether finer-scale statistical or dynamical products recover observed sub-grid structure rather than merely reduce bias or increase nominal resolution. Validation at the shortest spatial scales nevertheless remains limited by the scarcity of closely spaced observations. Only 13 station pairs are available below 1 km, and their performance varies substantially across locations. Evidence approaching the 30 m grid spacing is therefore indirect, based on consistency with known physical relationships and, in selected temperature examples, independent ECOSTRESS land surface temperature observations. Denser networks and targeted field campaigns will be necessary to quantify recovery at these scales more comprehensively.

The ability to reconstruct near-surface micro-weather from existing observing systems could improve characterization of weather exposure in applications ranging from urban heat and wildfire behavior to wind energy and pollutant transport. More broadly, many Earth system processes are observed through a similar combination of sparse measurements of dynamic states and spatially dense observations of environmental structure. Our results show that combining information in these observing systems can reveal variability far below the resolution at which the dynamic state

is directly observed or modeled.

8 Methods

8.1 Data

Point-based Weather Observations We obtained point-based weather observations from the NOAA Meteorological Assimilation Data Ingest System (MADIS), which aggregates observations from governmental, academic, private, and volunteer station networks (*37*). We used the LDAD Mesonet source and retained 2 m temperature, 2 m dewpoint, and 10 m wind speed and direction.

In this work, we focus on stations over the contiguous United States (CONUS) from 2020–2023. We calculate hourly-averaged observations and keep observations with the MADIS quality flag "Screened" or "Verified". We then retain only stations with at least 60% of data of sufficient quality over the study period, leaving 11,849 stations (Fig. 7), each with a maximum of 35,064 hourly observations.

During both training and evaluation, missing values are masked out. For test stations, we applied additional variable-specific filters intended to remove records with conspicuous indications of measurement artifacts while retaining unusual but potentially valid observations. Temperature and dewpoint stations were excluded if they contained at least 24 consecutive hours with an identical value, a threshold intended to identify possible sensor freezing without excluding shorter periods of limited variability or measurement quantization. Wind stations were excluded if both wind components were exactly zero in more than 90% of quality-controlled hours, provided the station reported at least 100 such hours. Both criteria were evaluated on the 2023 evaluation year rather than over the full 2020–2023 study period. Although zero wind is physically plausible and some anemometers report zero below a minimum operating threshold (e.g., 2 knots), records with an unusually high proportion of zeros may be unreliable for evaluation. These additional filters were not applied to training stations because moderate observation noise can be tolerated during model fitting, whereas such artifacts could bias evaluation.

Kilometer-scale Weather Products The ECMWF Reanalysis v5 (ERA5) dataset (*32*) from the European Centre for Medium-Range Weather Forecasts (ECMWF) is a global gridded reanalysis product that integrates physical model outputs with observations. ERA5 provides hourly data on a $0.25^\circ \times 0.25^\circ$ grid (about 28 km at the equator) covering the period from 1940–present. It includes

weather data both at the surface and at different pressure levels. For this work, we curated 4 years (2020–2023) of surface variables over CONUS: 10 m wind u, 10 m wind v, 2 m temperature, and 2 m dewpoint temperature. ERA5 serves as both an input to the Marmot model and as a baseline; it is interpolated to each backbone station and target location using bicubic interpolation.

The High-Resolution Rapid Refresh (HRRR) system (*34, 38*) is an hourly updated, convection-allowing NWP system developed by the NOAA for the contiguous U.S. HRRR assimilates observations into a regional atmospheric model and provides weather fields on an approximately 3 km grid. We extracted the corresponding 10 m wind u, 10 m wind v, 2 m temperature, and 2 m dewpoint temperature fields for the 2023 evaluation period. HRRR is used only as an evaluation baseline, providing a comparison with a higher-resolution regional dynamical model. It is not an input to Marmot.

Meter-scale Satellite Imagery In our main model specification, we use embeddings from the AlphaEarth geospatial foundation model (*31*) to provide information about land cover, terrain, and other surface characteristics to Marmot. We also construct conventional Earth observation inputs from raw Sentinel-2 imagery, digital elevation model (DEM), digital surface model (DSM), and a land cover map, to test whether the foundation model embedding adds information beyond them. Below we describe the data used in both approaches.

The AlphaEarth foundation model ingests optical, radar, lidar, DEM, land cover, climatic, and other geospatial information into a 64-dimensional annual embedding at 10 m resolution (*31*). We access AlphaEarth embeddings through Google Earth Engine and downsample the embeddings to 30 m for our main model specification. We use AlphaEarth embeddings from 2019, the year preceding our study period, to ensure no temporal leakage.

Separately, we also prepare raw and derived satellite image chips for each station location, from which we construct conventional Earth observation inputs as an alternative to the AlphaEarth embedding (Fig. 6). These image chips are static in time and consist of 30 m-resolution DSM and DEM and 10 m-resolution land cover and Sentinel-2 imagery from December 2023 to December 2024. We construct 32×32 pixel 30 m-resolution chips and 96×96 pixel 10 m-resolution chips centered on each weather station. Each chip thus covers 960×960 meters surrounding each station. The DSM data are from the JAXA ALOS DSM: Global 30m v3.2 (*39*), and the DEM data are from Copernicus DEM GLO-30: Global 30m Digital Elevation Model (*40*). We use ESA WorldCover 10 m v200 for land cover data (*41*) and retain 7 categories: Tree Cover, Grassland/Herbaceous wet-

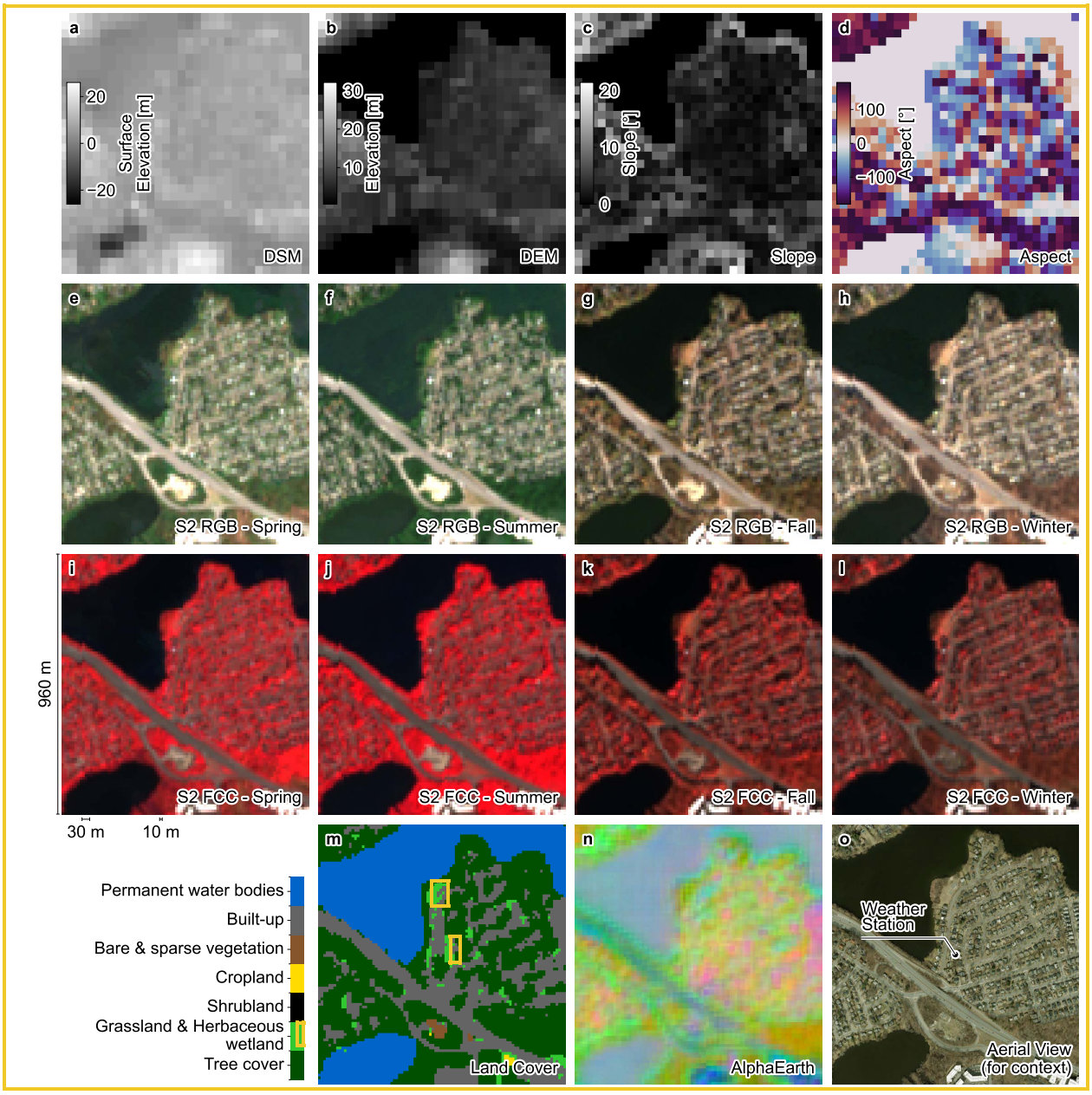


Figure 6: **Example of Earth observation inputs for a given location. (a–d)** DSM, DEM, and derived indices. **(e–h)** Sentinel-2 for spring, summer, fall, and winter seasons visualized in RGB. **(i–l)** Sentinel-2 for the four seasons visualized as a False Color Composite (FCC; NIR, R, G). **(m)** Land cover map with seven classes selected (of which only six are present at this location). **(n)** Geospatial foundation model embeddings produced by AlphaEarth (*31*), visualized via 3 embedding values. **(o)** Aerial image of the same location, not used in the model, for context only.

land, Shrubland, Cropland, Bare/Sparse Vegetation, Built-up and Permanent Water Bodies. For every location, we also create four seasonal Sentinel-2 composites, each comprised of the four 10 m resolution bands: Blue, Green, Red, and NIR. For each season, we select all the images that fall within the given time period, mask out clouds (*42*), and compute the median value for each pixel. For each band and season we retain the station's own pixel together with the mean and standard deviation over the chip, giving the 48 Sentinel-2 columns of Table S4.

ECOSTRESS Land Surface Temperature (LST) ECOSTRESS is an instrument aboard the International Space Station that senses land surface temperature and emissivity (LST&E) at 70 m resolution with 4 day revisit (*43*). As LST and air temperature are correlated (*35*), we use the LST data product to qualitatively assess the micro-weather model's ability to infer sub-grid temperature variations.

However, the 2 m air temperature inferred in this work differs from ECOSTRESS LST in several important ways, and their comparison requires careful interpretation. First, our air temperature is averaged hourly, while ECOSTRESS-sensed LST is instantaneous. Second, the relationship between ground temperature and 2 m air temperature is highly non-linear (*44, 45, 46, 47*). A diurnal cycle exists: at night, LST and air temperature are more closely correlated, whereas during the day, the differences can become substantial (*45*). Over vegetated areas, LST is generally lower than air temperature during the day due to evapotranspiration (*45*). Bare and sparsely vegetated areas, such as urban regions, exhibit the largest differences between LST and air temperature. Third, satellite-based LST retrievals have higher errors over complex terrain than flat terrain (*48*). Beyond measurement errors, the coupling between LST and air temperature weakens with elevation.

For these reasons, we limit comparison of our air temperature fields with ECOSTRESS LST to a qualitative assessment within case studies to establish general spatial correlations and physical plausibility.

8.2 Marmot Model Architecture

Marmot uses a transformer architecture to represent spatial relationships among weather stations and transfer information to arbitrary target locations (*49*). Let

$$\mathbf{y} = (T, D, u, v)^\top \in \mathbf{R}^4 \tag{1}$$

denote the four predicted weather variables: 2 m air temperature, 2 m dewpoint, and u, v components of 10 m wind. Each prediction is conditioned on same-timestamp observations of $\mathbf{y}$ from

the 64 nearest backbone stations, spatial relationships between those stations and the target, high-resolution surface information at the station and target locations, the spatial pattern in ERA5, and the time of prediction. Marmot predicts the four components of $\mathbf{y}$ jointly.

The architecture consists of three stages. First, Marmot produces a representation of the neighboring weather-station snapshot. Second, it uses AlphaEarth surface embeddings and station-to-target geometry to determine how informative each station is for the target and combines the stations into a target-specific summary. Third, it combines this summary with ERA5 and temporal information to predict $\mathbf{y}$. An auxiliary connection also allows an attention-weighted average of the station-level differences between observations and ERA5 to contribute directly to the corresponding outputs. In total, the Marmot model contains 2.33 million trainable parameters.

Representing Neighboring Stations For each target, Marmot selects the 64 nearest backbone stations by Haversine distance. Each station $i \in \{1, \ldots, 64\}$ is represented by a 77-dimensional token containing the four normalized weather observations $\mathbf{y}_i \in \mathbf{R}^4$; five spatial features; a 64-dimensional AlphaEarth embedding at the station location; and four ERA5 innovation terms. The spatial features are latitude and longitude offsets, Haversine distance, logarithm of Haversine distance, and elevation difference. Defining $\mathbf{e}_i \in \mathbb{R}^4$ as the corresponding ERA5 values at station i, its innovation vector is

$$\boldsymbol{\delta}_i = \mathbf{y}_i - \mathbf{e}_i. \tag{2}$$

Observations that are unavailable or fail quality control are masked out. When a backbone station is itself used as a training target, that station is excluded from its own neighbor set.

Marmot first processes the stations jointly using self-attention. The 77 station features are projected to a 128-dimensional representation and passed through four transformer blocks, each with four attention heads and a feed-forward layer. This stage allows the representation of each station to incorporate information from the surrounding station network. For example, it can represent whether an observation is part of a spatially coherent weather pattern or differs from the other nearby observations. A residual projection adds the resulting contextual information to the station representations.

Querying at the Target Location Marmot next assigns each neighboring station i a weight a_i according to how informative it appears for the target using cross-attention (*50, 51, 49*). The target query is generated from the AlphaEarth embedding at the target location. Each neighboring

station supplies a key constructed from its AlphaEarth embedding and its location relative to the target, and a value $\mathbf{v}_i \in \mathbf{R}^{256}$ constructed from its updated station representation. The scaled inner product of the target query and each station key determines its attention score. The resulting weights are normalized across the valid neighboring stations so that $\sum_{i=1}^{64} a_i = 1$. Marmot uses these weights to construct a neighboring-station summary $\mathbf{h} = \sum_{i=1}^{64} a_i \mathbf{v}_i$, with more relevant stations contributing more strongly. One set of station weights is used to construct this shared summary for all four components of $\mathbf{y}$. Subsequent prediction layers can nevertheless use the information in $\mathbf{h}$ differently for each output variable.

Adding ERA5 Spatial Information Marmot next combines $\mathbf{h}$ with information about how ERA5 varies around the target. For each weather variable, the model receives a 5×5 ERA5 neighborhood centered on the target. Let $\mathbf{E}_0 \in \mathbf{R}^4$ denote the ERA5 values at the center grid node and $\mathbf{E}_j \in \mathbb{R}^4$ the values at each of the other 24 grid cells. Marmot forms the differences

$$\mathbf{r} = [\mathbf{E}_1 - \mathbf{E}_0, \ldots, \mathbf{E}_{24} - \mathbf{E}_0] \in \mathbf{R}^{96}. \tag{3}$$

Thus, $\mathbf{r}$ describes the spatial pattern of the coarse ERA5 weather field around the target rather than its absolute level.

The 96 ERA5 features in $\mathbf{r}$ are concatenated with the 256-dimensional station summary $\mathbf{h}$ and projected back to a 256-dimensional representation $\mathbf{z}$. Marmot passes $\mathbf{z}$ through three residual gated feed-forward blocks and then maps the result to the four weather outputs. These prediction layers learn flexible relationships among the contemporaneous station observations, station and target environments, and the coarse spatial pattern in ERA5.

Representing Time Dependence The relationship between a target and its neighboring stations can vary by time of day and season. Marmot represents local solar hour and day of year using sine and cosine features, denoted collectively by $\boldsymbol{\tau}$. It combines $\boldsymbol{\tau}$ with the AlphaEarth embedding at the target, $\mathbf{s}_0$, to calculate a separate scale and offset for each component of $\mathbf{z}$:

$$\widetilde{\mathbf{z}} = \boldsymbol{\gamma}(\boldsymbol{\tau}, \mathbf{s}_0) \odot \mathbf{z} + \boldsymbol{\beta}(\boldsymbol{\tau}, \mathbf{s}_0), \tag{4}$$

where $\odot$ denotes elementwise multiplication. This adjustment is applied once, before the three feed-forward blocks. The same station and ERA5 information can therefore be interpreted differently depending on the target environment, local time, and season. For example, relationships between forested and bare locations may differ between day and night or across seasons.

Producing the Weather Predictions Let $F(\tilde{\mathbf{z}}) \in \mathbf{R}^4$ denote the output of the three feed-forward blocks and final prediction layer. An auxiliary connection uses the same station weights to calculate an average innovation,

$$\overline{\boldsymbol{\delta}} = \sum_{i=1}^{64} a_i \boldsymbol{\delta}_i, \tag{5}$$

and applies four learned, variable-specific scaling factors $\mathbf{g}(\mathbf{h}) \in \mathbf{R}^4$. The final prediction in the normalized output space is

$$\widehat{\mathbf{y}}_{\text{norm}} = F(\tilde{\mathbf{z}}) + \mathbf{g}(\mathbf{h}) \odot \overline{\boldsymbol{\delta}}. \tag{6}$$

The four components are then transformed back to physical units to obtain $\widehat{\mathbf{y}} = (\widehat{T}, \widehat{D}, \widehat{u}, \widehat{v})^\top$.

8.3 Marmot Training and Inference

We divided the 11,849 weather stations into 4 groups (Fig. 7): backbone, training, validation, and test sets. Stations were partitioned by whole 0.25° ERA5 grid cell, so that every station in a cell belongs to the same group and no cell is shared between groups. The backbone set (8,180 stations, corresponding to 245,562,977 valid station-time step observations) supplies information on local weather to nearby target locations for as many time steps as possible. The training set contains 2,260 stations (67,915,653 valid observations), and the validation and test sets contain 854 stations (25,651,801 valid observations) and 555 stations (16,632,860 valid observations), respectively.

The model weights were trained by tasking the model with predicting temperature, dewpoint, and the u, v components of wind at training-set and backbone stations (10,440 stations in total) for hourly time steps between 2020–2022. Models were trained for 120,000 steps with a batch size of 128 target station-hours (64 per GPU across two GPUs), using the AdamW optimizer (*52*) with a learning rate of 5×10^{-4} modified with cosine annealing and a weight decay of 10^{-2}. The model checkpoint used for test set evaluation and inference was selected based on the best validation set performance.

At inference time, Marmot can be queried at any location and timestamp for which an AlphaEarth embedding is available. Each output pixel is treated as an independent target and receives the same inputs as a target station during training. Spatially continuous fields at 30 m resolution are assembled from these per-pixel predictions, with the maximum region size constrained by available memory.

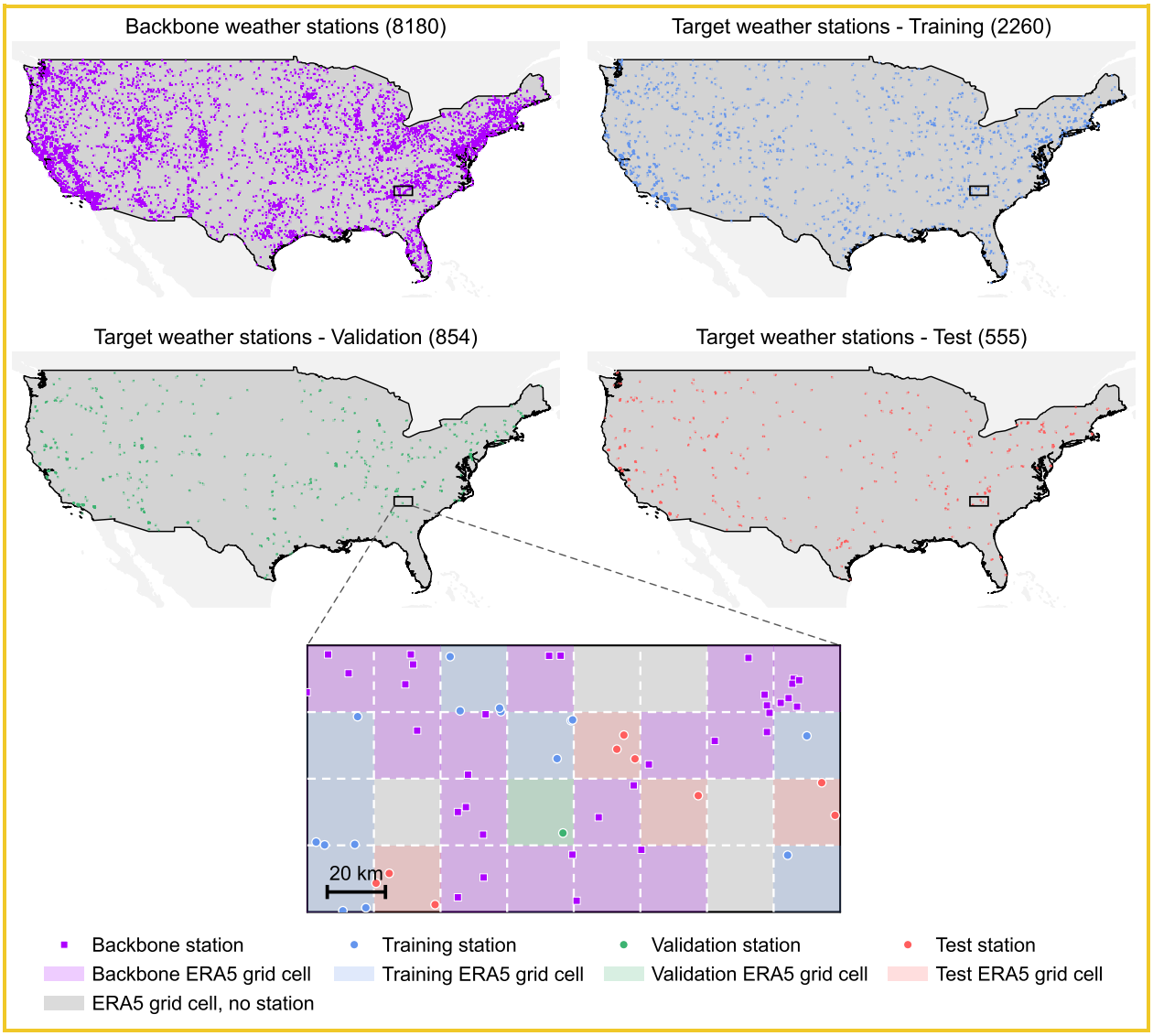


Figure 7: **Spatial partitioning of contiguous U.S. ERA5 cells and weather stations for training and evaluation.** ERA5 grid cells across the contiguous U.S. are split into four disjoint sets: backbone, training, validation, and test. Surface weather stations from NOAA MADIS inherit the assignment of the ERA5 grid cell in which they are located. Only test set stations are used for evaluation.

8.4 Baseline Models

In meteorology, the term "downscaling" is broad and encompasses methods with different prediction targets, spatial and temporal scales, and evaluation objectives. Because our task is to predict weather at point-level locations and hourly timesteps, we selected baselines that provide temporally aligned estimates that can be evaluated directly against held-out station observations. These baselines span four sources of local weather information: nearby station observations, the global reanalysis used as our model input, a higher-resolution regional NWP product, and statistical post-processing learned from station observations. Specifically, we compared against backbone station interpolation, ERA5, HRRR, and ERA5-Model Output Statistics (ERA5-MOS).

Other relevant methods address adjacent but different tasks. BCSD (*53*) and GenFocal (*54*) downscale gridded climate or regional-model fields, generally at coarser temporal scales. Aardvark passes off-grid observations through a gridded atmospheric representation and evaluates at station locations seen during training (*24*); MetNet-3 produces dense regional forecasts (*26*); and Yang et al. (*25*) forecast at instrumented stations using the target station's recent observations. None directly evaluates sub-grid spatial recovery at locations with no observation history. Applying these systems under Marmot's inputs and spatial split would therefore require methodological adaptation rather than direct evaluation of the published models.

Station Interpolation Baseline This baseline measures how well local conditions can be estimated from nearby station observations using standard spatial interpolation. We interpolated each weather variable from the backbone stations to every test station location by inverse-distance weighting ($1/d^2$ where d is Haversine distance) using the $k = 12$ nearest backbone stations, with k selected on the validation set and weights renormalized over whichever of those stations report at that hour. Where none of the twelve report, the estimate escalates to the nearest reporting stations among the 64 nearest backbone stations, and then to the whole backbone network, so that the baseline is defined on the same station-hours as the other methods; this affects fewer than 0.004% of values that are evaluated. We also evaluated radial-basis-function interpolation, but inverse-distance weighting performed better on the validation set and is therefore reported as the station-interpolation baseline.

ERA5 Baseline ERA5 is a leading global atmospheric reanalysis and provides the coarse weather fields used as inputs to our model; this baseline therefore measures both performance relative to

a widely used global weather product and the improvement provided by our model over its own meteorological inputs. We used bilinear interpolation of the ERA5 grid to obtain weather values at test station locations.

HRRR Baseline High-Resolution Rapid Refresh (HRRR) is an operational regional analysis built on the WRF-ARW dynamical core, running at 3-km convection-allowing resolution with hourly cycled data assimilation (*34*). It therefore provides a directly evaluable, observation-constrained WRF-based benchmark across CONUS. Comparing Marmot against HRRR places its improvements in the context of what can be achieved through higher-resolution dynamical modeling. We assigned each test station the HRRR value from the grid cell containing it at the corresponding hour. Because HRRR stores wind components relative to its Lambert conformal grid, we rotated them to true east and north before evaluation.

ERA5-MOS This baseline tests whether station-level accuracy can be improved through conventional statistical correction of ERA5 using observations and hand-crafted surface attributes. Model output statistics (MOS) is a statistical postprocessing approach that relates NWP output to observations to correct systematic model errors (*55*). We implemented a MOS baseline that learned the systematic difference between ERA5 and station observations (*33*). Separate models were trained for 2 m temperature, 2 m dew-point temperature, and the two 10 m wind components using the same training, validation, and test partitions as our model. The target was the hourly residual $e = x_{\mathrm{ERA5}} - x_{\mathrm{obs}}$, and the MOS prediction was obtained as $\hat{x} = x_{\mathrm{ERA5}} - \hat{e}$. Training targets were restricted to hours whose MADIS quality flag marks an observed value. Predictors comprised station and ERA5-grid surface attributes — including location, orographic variability, land-sea mask, vegetation characteristics, station elevation, and the elevation difference between the station and ERA5 grid cell — together with annual and local-time indicators. We evaluated second-order ridge regression, random forests, and multilayer perceptrons as the regression model. Their performance was similar, consistent with previous findings (*33*), and we report the specification with the best validation performance, a random forest.

8.5 Evaluation Protocol

We evaluate model and baseline performance at held-out test stations using metrics designed to measure overall predictive accuracy, recovery of spatial variability, recovery of local contrasts, and

performance under extreme or rapidly changing weather. Unless otherwise stated, metrics are computed using hourly observations from the held-out set. All methods are evaluated on the same available station-time observations.

Aggregate Error Metrics Temperature and dewpoint are evaluated using scalar root mean square error (RMSE) and mean absolute error (MAE). We report RMSE in the main text and MAE in additional tables. Wind is evaluated using the Euclidean distance between observed and predicted horizontal wind vectors, summarized by its root mean square error (vector RMSE) and mean absolute error (vector MAE). These metrics jointly account for errors in wind speed and direction. Scalar errors in wind speed and direction are reported separately.

National Spatial Variance Explained Spatial R^2 is computed to quantify how well the model captures spatial structure at a particular snapshot in time. At each time step t, we remove the across-station mean from both observations and predictions so that

$$\tilde{y}_{i,t} = y_{i,t} - \bar{y}_t, \quad \tilde{m}_{i,t} = m_{i,t} - \bar{m}_t,$$

where $\bar{y}_t, \bar{m}_t$ denote spatial means across stations at time t. Spatial R^2 is then defined as

$$R^2_{\text{spatial}}(t) \;=\; 1 - \frac{\sum_i (\tilde{y}_{i,t} - \tilde{m}_{i,t})^2}{\sum_i \tilde{y}^2_{i,t}}.$$

For wind, the multivariate R^2_{spatial} is calculated from the predicted and observed wind vectors using squared Euclidean distances in the numerator and denominator.

Because the observations and predictions are centered separately, this statistic measures agreement in their spatial deviations rather than differences in their spatial means. A value of one indicates perfect recovery of the observed spatial pattern, a value of zero indicates performance equivalent to predicting no spatial variation, and a negative value indicates poorer performance than that reference. When reporting a single R^2_{spatial} metric for a model, we average $R^2_{\text{spatial}}(t)$ across all test time steps.

Sub-grid Spatial Variance To evaluate spatial variability below the resolution of Marmot's ERA5 input, we compute R^2_{spatial} separately within each held-out ERA5 grid cell and hour.

Let $\mathcal{I}_{g,t}$ denote the set of test stations with valid observations in ERA5 grid cell g at time t. For a scalar variable, the within-cell observed and predicted anomalies are

$$\widetilde{y}_{i,g,t} = y_{i,t} - \overline{y}_{g,t}, \qquad \widetilde{m}_{i,g,t} = m_{i,t} - \overline{m}_{g,t},$$

where $\overline{y}_{g,t}$ and $\overline{m}_{g,t}$ are the respective means over stations in $\mathcal{I}_{g,t}$. Within-cell spatial variance explained is

$$R^2_{\text{spatial}}(g,t) = 1 - \frac{\sum_{i\in\mathcal{I}_{g,t}} (\widetilde{y}_{i,g,t} - \widetilde{m}_{i,g,t})^2}{\sum_{i\in\mathcal{I}_{g,t}} \widetilde{y}^2_{i,g,t}}.$$

For wind, the scalar differences are replaced with squared Euclidean distances between the centered horizontal wind vectors.

We compute this statistic for cell–time combinations in which at least five test stations report simultaneously and all five methods yield finite predictions, and we exclude combinations whose observed spatial variance is zero, for which R^2_{spatial} is undefined. Cells are eligible if they contain at least five test stations. Performance for each grid cell is summarized by the median of $R^2_{\text{spatial}}(g,t)$ over its eligible hours. Figure 3a reports the distribution of these cell-level medians across the 24 held-out ERA5 cells eligible for each variable, which contain a median of 6 test stations per cell (maximum 38) and a median of 7,991 eligible hours per cell for temperature and dewpoint and 8,148 for wind (minimum 227).

Sub-grid Spatial Contrast We also evaluate whether each method recovers differences in weather at the same timestamp between nearby locations. For a pair of test stations (i, j) observed at time t, we define the observed and predicted spatial contrasts as

$$\Delta y_{ij,t} = y_{i,t} - y_{j,t}, \qquad \Delta m_{ij,t} = m_{i,t} - m_{j,t}.$$

For a set $\mathcal{P}$ of valid station-pair-time observations, spatial-contrast recovery is quantified using

$$R^2_{\Delta} = 1 - \frac{\sum_{(i,j,t)\in\mathcal{P}} (\Delta y_{ij,t} - \Delta m_{ij,t})^2}{\sum_{(i,j,t)\in\mathcal{P}} (\Delta y_{ij,t})^2},$$

For wind, the contrasts are vector-valued, and the numerator and denominator are computed using squared Euclidean norms. A value of one indicates perfect recovery, zero indicates performance equivalent to predicting identical conditions at both locations, and negative values indicate performance worse than that reference.

For the national analysis, we form pairs of held-out stations within the same ERA5 grid cell, group them according to their geographic separation, and compute R^2_{Δ} separately within each distance bin. The distance bins are 0–1, 1–2, 2–3, 3–5, 5–7, 7–10, 10–15, 15–20 and 20–30 km. No pairs are excluded on distance grounds: the widest separation between two test stations sharing an ERA5 cell is 28.5 km, so the bins span the full range. Each station-pair-hour observation receives equal weight.

To characterize heterogeneity across station pairs, we additionally calculate R^2_Δ separately for each pair using all valid observations for that pair in 2023. For each variable, the observed contrast magnitude of pair (i, j) is summarized by its root-mean-square contrast, $\mathrm{RMS}(\Delta y_{ij}) = \sqrt{\frac{1}{n_{ij}} \sum_t \|\Delta y_{ij,t}\|^2}$. Figure S5 shows pair-specific scores as functions of observed RMS contrast and station separation.

Surface Correlates of Inferred Fields To quantify the association between inferred micro-weather structure and surface characteristics, we fit ordinary least-squares regressions separately for each case study in Fig. 4. The response variable is Marmot's predicted weather value at each 30 m pixel in the inferred field. We fit three specifications using (i) terrain, represented by elevation and slope magnitude derived from the DEM, (ii) land cover as indicator variables for each class covering at least 1% of the region, and (iii) terrain and land cover together. We report the coefficient of determination, R^2, which measures the fraction of spatial variance in the inferred field associated with each set of surface variables (Table S6).

Extreme Weather Evaluation We evaluate performance under extreme weather in three ways: conditional errors across the observed weather distribution, detection at fixed absolute thresholds, and detection of rapid weather changes.

First, to examine performance across the tails of the distribution, we assign each observation its percentile within the empirical distribution of its station. For wind, percentiles are computed using wind speed. Observations are grouped into twelve percentile bins with edges at 0, 1, 2, 5, 10, 25, 50, 75, 90, 95, 98, 99, and 100. Within each bin we compute RMSE and mean error at each station, then report the median across stations; a station contributes to a bin only if it has at least 20 observations in it, and stations with fewer than 2,000 valid hours over the test year are excluded entirely. Mean error indicates systematic over- or underprediction, with positive values denoting overprediction and negative values underprediction.

Second, we evaluate detection of extreme heat, humidity, and wind at fixed absolute thresholds. Temperature thresholds are 32, 35, and 40°C; dewpoint thresholds are 20, 24, and 26°C; and wind-speed thresholds are 8, 10, and 12m s^{-1}. An observed or predicted event occurs when the corresponding value meets or exceeds the threshold. We calculate precision and recall separately at each station and report the median across stations. Precision measures how often predicted extreme events occur, whereas recall measures how often observed extreme events are detected. Recall is

only calculated on stations with at least 10 observed events, and precision only at those of these stations where the method predicts at least one event, since precision is otherwise undefined. The number of stations eligible for recall at each threshold is reported in Figure 5b.

Finally, we evaluate rapid increases and decreases over 3-hour intervals. Ramps are defined as the difference between values 3 hours apart. Positive and negative ramp events are identified when the observed or predicted change exceeds the threshold in either direction. We evaluate thresholds of ± 5°C and ± 8°C for temperature, ± 5°C and ± 8°C for dewpoint, and ± 3 m s^{-1} and ± 5 m s^{-1}for wind speed. Ramp detection is evaluated using station-level precision, recall, and F1 score and summarized across stations in the same manner as detection at absolute thresholds.

Uncertainty Estimation We train five models with identical hyperparameters and different random seeds. All aggregate metrics reported for Marmot, and correspondingly for ERA5-MOS, are computed separately for each of the five models and then averaged. The spatially continuous 30 m fields, and the paired-station contrast statistics reported for individual case studies, are generated from a single trained model. Uncertainty in reported performance reflects variation in both the held-out test locations and the training of the learned methods. We use a bootstrap over the held-out ERA5 grid cells and the five independently trained models of Marmot and of ERA5-MOS. Each replicate resamples 0.25° ERA5 cell blocks rather than individual stations, because stations within a cell share weather; within that same draw it also resamples, with replacement, the five training seeds of ERA5-MOS and Marmot, the two methods that are trained, and averages over them. The other three methods are deterministic and carry no seed term. One cell draw is shared across methods, so differences between models are computed within a replicate. We report error bars for one standard deviation across 2,000 bootstrap replicates.

8.6 Ablation Studies

Information Source Ablation To understand the contribution of different information sources to Marmot's ability to recover micro-weather, we conducted an additive ablation and a leave-one-out ablation.

In the additive ablation, which is shown in Figure 2c, we constructed the model progressively from a station-only configuration. The initial learned model received the contemporaneous observations and locations of the 64 nearest stations, along with pre-computed distances between the stations and target. We then sequentially added station-target elevation difference, ERA5 features

(comprised of "innovation" — the difference between each neighboring station observation and the corresponding ERA5 value — and a 5×5 neighborhood of ERA5 cells around the target), AlphaEarth embeddings, and a temporal encoding and neighbor dropout. Each configuration inherited all components in preceding configurations, allowing the incremental contribution of each addition to be evaluated.

In the leave-one-out ablation, we removed components from the full model one at a time. All remaining components were held at their full-model settings.

We note that the additive ablation depends on the order in which components are introduced; for example, because station-target elevation differences and AlphaEarth embeddings contain overlapping information about topography, adding elevation differences first may reduce the subsequent improvement from AlphaEarth. The leave-one-out ablation is instead symmetric across components, but performance losses may be small when other components provide overlapping information.

Surface Information Ablation To identify which sources of surface information contribute most to model performance and whether the AlphaEarth embedding provides information beyond conventional surface products, we replaced the AlphaEarth embedding with features derived from terrain (the Copernicus GLO-30 digital elevation model and the ALOS AW3D30 digital surface model, with slope and aspect derived from the DEM), ESA WorldCover land cover, and seasonal Sentinel-2 reflectance composites, individually and in combination.

8.7 Sensitivity Analysis

We sought to understand Marmot's sensitivity to four major hyperparameters: the number of neighboring stations, the spatial resolution of the AlphaEarth embeddings, the spatial extent of the ERA5 input, and the neighbor-dropout rate. All sensitivity analyses were conducted on the validation set. Unless otherwise specified, each analysis varied one setting while holding the remaining architecture, inputs, and training procedure at the reference configuration: 64 neighboring stations, 30-m AlphaEarth embeddings, a 5×5 ERA5 neighborhood, and 25% neighbor dropout.

We varied the number of neighboring stations among 16, 32, 64, and 128. We compared AlphaEarth embeddings extracted at 10-, 30-, and 100-m spatial resolution. We varied the ERA5 neighborhood from the target cell alone (1×1) to 3×3, 5×5, 7×7, and 9×9 cells. To assess sensitivity to training-time neighbor dropout, we randomly withheld 0, 10, 25, 50, 60, or 75% of neighboring stations during training; all available neighbors were retained at inference.

Finally, we evaluated sensitivity to data partitioning by comparing our primary ERA5 grid-cell split against a random station-level split (Table S2). Because our primary evaluation holds out entire ERA5 grid cells, it measures spatial generalization to regions where no backbone stations are available. The random station split, by contrast, allows neighboring stations within the same grid cell to remain in the training set. Comparing the two quantifies the performance impact of spatial autocorrelation between nearby stations during evaluation.

References

1. S. Wang, *et al.*, Dual impact of global urban overheating on mortality. *Nature Climate Change* **15** (5), 497–504 (2025), doi:10.1038/s41558-025-02303-3, https://doi.org/10.1038/s41558-025-02303-3.

2. L. Li, *et al.*, Divergent urbanization-induced impacts on global surface urban heat island trends since 1980s. *Remote Sensing of Environment* **295**, 113650 (2023), doi:10.1016/j.rse.2023.113650, https://www.sciencedirect.com/science/article/pii/S0034425723002018.

3. G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, C. Draxl, The State-of-the-Art in Short-Term Prediction of Wind Power: A Literature Overview (2011), doi:10.11581/DTU:00000017.

4. M. Eghdami, P. A. Jiménez y Muñoz, A. DeCastro, Sensitivity to the Representation of Wind for Wildfire Rate of Spread: Case Studies with the Community Fire Behavior Model. *Fire* **8** (4) (2025), doi:10.3390/fire8040135, https://www.mdpi.com/2571-6255/8/4/135.

5. M. G. Cruz, M. E. Alexander, The 10% wind speed rule of thumb for estimating a wildfire's forward rate of spread in forests and shrublands. *Annals of Forest Science* **76** (2), 44 (2019).

6. A. Baklanov, S. Grimmond, A. Mahura, M. Athanassiadou, eds., *Meteorological and Air Quality Models for Urban Areas* (Springer, Berlin, Heidelberg), 1 ed. (2009), doi:10.1007/978-3-642-00298-4.

7. F. Giorgi, L. O. Mearns, Approaches to the Simulation of Regional Climate Change: A Review. *Reviews of Geophysics* **29** (2), 191–216 (1991), doi:10.1029/90RG02636.

8. M. Rummukainen, State-of-the-art with Regional Climate Models. *WIREs Climate Change* **1** (1), 82–96 (2010), doi:10.1002/wcc.8.

9. R. E. Chandler, Multisite, Multivariate Weather Generation Based on Generalised Linear Models. *Environmental Modelling & Software* **134**, 104867 (2020), doi:10.1016/j.envsoft.2020.104867.

10. C. Schär, *et al.*, Kilometer-Scale Climate Models: Prospects and Challenges. *Bulletin of the American Meteorological Society* **101** (5), E567–E587 (2020), doi:10.1175/BAMS-D-18-0167.1.

11. J. D. Jijón, K.-H. Gaudry, J. Constante, C. Valencia, Augmenting the Spatial Resolution of Climate-Change Temperature Projections for City Planners and Local Decision Makers. *Environmental Research Letters* **16** (5), 054028 (2021), doi:10.1088/1748-9326/abf7f2.

12. R. Lam, *et al.*, Learning Skillful Medium-Range Global Weather Forecasting. *Science* **382** (6677), 1416–1421 (2023), doi:10.1126/science.adi2336.

13. T. Kurth, *et al.*, FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators, in *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '23 (Association for Computing Machinery, New York, NY, USA) (2023), doi:10.1145/3592979.3593412, `https://doi.org/10.1145/3592979.3593412`.

14. T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, A. Grover, ClimaX: A foundation model for weather and climate, in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, *et al.*, Eds. (PMLR), vol. 202 of *Proceedings of Machine Learning Research* (2023), pp. 25904–25938, `https://proceedings.mlr.press/v202/nguyen23a.html`.

15. K. Bi, *et al.*, Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619** (7970), 533–538 (2023).

16. C. Bodnar, *et al.*, A Foundation Model for the Earth System. *Nature* **641** (8065), 1180–1187 (2025), doi:10.1038/s41586-025-09005-y.

17. J. Schmude, *et al.*, Prithvi WxC: Foundation Model for Weather and Climate (2024), doi:10.48550/arXiv.2409.13598.

18. I. Price, *et al.*, Probabilistic Weather Forecasting with Machine Learning. *Nature* **637** (8044), 84–90 (2025), doi:10.1038/s41586-024-08252-9.

19. F. Alet, *et al.*, Skillful Joint Probabilistic Weather Forecasting from Marginals (2025), doi:10.48550/arXiv.2506.10772.

20. M. Satoh, *et al.*, Global Cloud-Resolving Models. *Current Climate Change Reports* **5** (3), 172–184 (2019), doi:10.1007/s40641-019-00131-0, `https://doi.org/10.1007/s40641-019-00131-0`.

21. A. F. Prein, R. Rasmussen, G. Stephens, Challenges and Advances in Convection-Permitting Climate Modeling. *Bulletin of the American Meteorological Society* **98** (5), 1027 – 1030

(2017), doi:10.1175/BAMS-D-16-0263.1, https://journals.ametsoc.org/view/journals/bams/98/5/bams-d-16-0263.1.xml.

22. N. P. Wedi, *et al.*, A Baseline for Global Weather and Climate Simulations at 1 km Resolution. *Journal of Advances in Modeling Earth Systems* **12** (11), e2020MS002192 (2020), doi:10.1029/2020MS002192, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002192.

23. European Centre for Medium-Range Weather Forecasts, *ECMWF Strategy 2025–2034*, Tech. rep., European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK (2024), doi:10.21957/05dd9b657e, https://www.ecmwf.int/en/about/strategy.

24. A. Allen, *et al.*, End-to-End Data-Driven Weather Prediction. *Nature* (2025), doi:10.1038/s41586-025-08897-0.

25. Q. Yang, *et al.*, Local Off-Grid Weather Forecasting With Multi-Modal Earth Observation Data. *Journal of Advances in Modeling Earth Systems* **17** (12), e2025MS005207 (2025), doi:10.1029/2025MS005207.

26. M. Andrychowicz, *et al.*, Deep Learning for Day Forecasts from Sparse Observations. *arXiv preprint arXiv:2306.06079* (2023), doi:10.48550/arXiv.2306.06079.

27. M. Drusch, *et al.*, Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* **120**, 25–36 (2012), the Sentinel Missions - New Opportunities for Science, doi:10.1016/j.rse.2011.11.026, https://www.sciencedirect.com/science/article/pii/S0034425712000636.

28. L. Yang, *et al.*, A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing* **146**, 108–123 (2018), doi:10.1016/j.isprsjprs.2018.09.006, https://www.sciencedirect.com/science/article/pii/S092427161830251X.

29. T. G. Farr, *et al.*, The Shuttle Radar Topography Mission. *Reviews of Geophysics* **45** (2) (2007), doi:10.1029/2005RG000183, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183.

30. J. B. Fisher, *et al.*, ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station. *Water Resources Research* **56** (4),

e2019WR026058 (2020), doi:10.1029/2019WR026058, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019WR026058.

31. C. F. Brown, *et al.*, AlphaEarth Foundations: An Embedding Field Model for Accurate and Efficient Global Mapping from Sparse Label Data (2025), doi:10.48550/arXiv.2507.22291.

32. H. Hersbach, *et al.*, The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146** (730), 1999–2049 (2020), doi:10.1002/qj.3803.

33. Z. B. Bouallègue, *et al.*, Statistical Modeling of 2-m Temperature and 10-m Wind Speed Forecast Errors. *Monthly Weather Review* **151** (4), 897 – 911 (2023), doi:10.1175/MWR-D-22-0107.1, https://journals.ametsoc.org/view/journals/mwre/151/4/MWR-D-22-0107.1.xml.

34. D. C. Dowell, *et al.*, The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description. *Weather and Forecasting* **37** (8), 1371–1395 (2022), doi:10.1175/WAF-D-21-0151.1.

35. E. J. Good, An in situ-based analysis of the relationship between land surface "skin" and screen-level air temperatures. *Journal of Geophysical Research: Atmospheres* **121** (15), 8801–8819 (2016), doi:10.1002/2016JD025318, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025318.

36. J. W. Oyler, S. Z. Dobrowski, Z. A. Holden, S. W. Running, Remotely Sensed Land Skin Temperature as a Spatial Predictor of Air Temperature across the Conterminous United States. *Journal of Applied Meteorology and Climatology* **55** (7), 1441 – 1457 (2016), doi:10.1175/JAMC-D-15-0276.1, https://journals.ametsoc.org/view/journals/apme/55/7/jamc-d-15-0276.1.xml.

37. P. A. Miller, M. F. Barth, L. A. Benjamin, An Update on MADIS Observation Ingest, Integration, Quality Control, and Distribution Capabilities, in *Proceedings of the 21st International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology* (American Meteorological Society, San Diego, California) (2005), https://ams.confex.com/ams/pdfpapers/86703.pdf, paper J7.12.

38. S. G. Benjamin, *et al.*, A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. *Monthly Weather Review* **144** (4), 1669 – 1694 (2016), doi:

10.1175/MWR-D-15-0242.1, https://journals.ametsoc.org/view/journals/mwre/144/4/mwr-d-15-0242.1.xml.

39. J. Takaku, T. Tadono, K. Tsutsui, Generation of High Resolution Global DSM from ALOS PRISM. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XL-4**, 243–248 (2014), doi:10.5194/isprsarchives-XL-4-243-2014.

40. European Space Agency, Copernicus Global Digital Elevation Model (2024), doi:10.5069/G9028PQB.

41. D. Zanaga, *et al.*, ESA WorldCover 10 m 2021 V200 (2022).

42. V. J. Pasquarella, C. F. Brown, W. Czerwinski, W. J. Rucklidge, Comprehensive Quality Assessment of Optical Satellite Imagery Using Weakly Supervised Video Learning, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, Vancouver, BC, Canada) (2023), pp. 2125–2135, doi:10.1109/CVPRW59228.2023.00206.

43. S. Hook, G. Hulley, ECOSTRESS Swath Land Surface Temperature and Emissivity Instantaneous L2 Global 70 m V002 (2022).

44. M. Jin, R. E. Dickinson, Land Surface Skin Temperature Climatology: Benefitting from the Strengths of Satellite Observations. *Environmental Research Letters* **5** (4), 044004 (2010), doi:10.1088/1748-9326/5/4/044004.

45. E. J. Good, D. J. Ghent, C. E. Bulgin, J. J. Remedios, A Spatiotemporal Analysis of the Relationship between Near-Surface Air Temperature and Satellite Land Surface Temperatures Using 17 Years of Data from the ATSR Series. *Journal of Geophysical Research: Atmospheres* **122** (17), 9185–9210 (2017), doi:10.1002/2017JD026880.

46. Y. Mo, N. Pepin, H. Lovell, Understanding Temperature Variations in Mountainous Regions: The Relationship between Satellite-Derived Land Surface Temperature and in Situ near-Surface Air Temperature. *Remote Sensing of Environment* **318**, 114574 (2025), doi:10.1016/j.rse.2024.114574.

47. H. Zhang, *et al.*, Global Evaluation of High-Resolution ECOSTRESS Land Surface Temperature and Emissivity Products: Collection 1 versus Collection 2. *Remote Sensing of Environment* **326**, 114799 (2025), doi:10.1016/j.rse.2025.114799.

48. B. Dong, *et al.*, Quantitative Evaluation of Complex Terrain Effects on Land Surface Temperature: A Data-driven Coupling Model of Dynamic and Static Features (2025), doi:10.2139/ssrn.5384696.

49. A. Vaswani, *et al.*, Attention Is All You Need, in *Advances in Neural Information Processing Systems* (Long Beach, CA, USA), vol. 30 (2017).

50. H. Lin, X. Cheng, X. Wu, D. Shen, CAT: Cross Attention in Vision Transformer, in *2022 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, Taipei, Taiwan) (2022), pp. 1–6, doi:10.1109/ICME52920.2022.9859720.

51. C.-F. R. Chen, Q. Fan, R. Panda, CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, Montreal, QC, Canada) (2021), pp. 347–356, doi:10.1109/ICCV48922.2021.00041.

52. I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in *International Conference on Learning Representations* (2019), `https://openreview.net/forum?id=Bkg6RiCqY7`.

53. A. W. Wood, L. R. Leung, V. Sridhar, D. P. Lettenmaier, Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Climatic Change* **62** (1–3), 189–216 (2004), doi:10.1023/B:CLIM.0000013685.99609.9E.

54. Z. Y. Wan, *et al.*, Regional Climate Risk Assessment from Climate Models Using Probabilistic Machine Learning. *Nature Machine Intelligence* (2026), doi:10.1038/s42256-026-01308-7.

55. H. R. Glahn, D. A. Lowry, The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology and Climatology* **11** (8), 1203 – 1211 (1972), doi:10.1175/1520-0450(1972)011⟨1203:TUOMOS⟩2.0.CO;2, `https://journals.ametsoc.org/view/journals/apme/11/8/1520-0450_1972_011_1203_tuomos_2_0_co_2.xml`.

Acknowledgments

The authors express their gratitude to Daniel Salles Civitarese, Johannes Jakubik, and Björn Lütjens for inspiring discussions and to Chris Hill for his support and guidance.

Funding: J.G. was supported by an MIT-IBM Watson AI Lab award sponsored by Shell Information Technology. Q.Y. and R.H. were partially supported by the National Aeronautics and Space Administration (NASA) under award No. 80NSSC26K0189.

Author contributions: Conceptualization: J.G., Q.Y., R.H., S.W. Data curation: J.G., Q.Y., R.H. Formal analysis: J.G., Q.Y., R.H., S.W. Funding acquisition: J.V., D.H., C.W., S.W. Investigation: J.G., Q.Y., R.H. Methodology: J.G., Q.Y., R.H., E.S., A.C., S.W. Software: J.G., Q.Y., R.H., Y.Z. Supervision: S.W. Visualization: J.G., R.H., S.W. Writing – original draft: J.G., S.W. Writing – review and editing: J.G., Q.Y., R.H., S.W.

Competing interests: There are no competing interests to declare.

Data and materials availability: Processed data and code will be made publicly available upon publication. ERA5 data are publicly available from the Copernicus Climate Data Store (`DOI:10.24381/cds.adbb2d47`), and MADIS observations are available from NOAA (`https://madis.ncep.noaa.gov/`). Surface data were accessed through the Google Earth Engine Data Catalog using collections `ESA/WorldCover/v200`, `COPERNICUS/DEM/GLO30`, `JAXA/ALOS/AW3D30/V3_2`, `COPERNICUS/S2_SR_HARMONIZED`, and `GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL`. ECOSTRESS L2T Land Surface Temperature and Emissivity (LSTE) Version 2 data were obtained through the NASA Earthdata portal. HRRR data were obtained from NOAA's HRRR archive in the Registry of Open Data on AWS. We used the zero-hour analysis fields.

9 Supplementary Materials

9.1 Station Pair-level Variation in Spatial Contrast Recovery

Figure 3b measures aggregate contrast recovery by pooling station-pair-hour observations within each separation bin. Because pairs with larger observed contrasts contribute more to this pooled statistic, we additionally calculate contrast skill separately for each station pair (Fig. S5). Pair-specific scores vary substantially at all separations. Median temperature skill generally declines toward shorter separations, while dewpoint and wind approach zero more rapidly. Scores are also generally higher for pairs with larger observed contrasts, consistent in part with the greater relative penalty imposed by R^2_Δ when the true contrast is small (Section 8.5). At sub-kilometer separations, the limited station sample includes both substantial temperature recovery and negative scores associated with incorrect inferred contrasts. Thus, the positive aggregate temperature estimate below 1 km reflects recovery of some observed spatial contrasts rather than uniform skill.

9.2 Temporal Recovery at Held-out Stations

To complement the spatial analyses in the main text, we examine the temporal behavior of Marmot's predictions at held-out test stations. We first revisit the station pairs highlighted in the sub-grid analysis and characterize their observed and predicted contrasts over diurnal and seasonal cycles (Fig. S4). We then examine individual weather events at held-out stations where Marmot improves upon the baselines (Fig. S6); failure cases are examined separately in Section 9.4.

The station-pair contrasts in Warren County, San Diego, and Austin (Fig. S3) exhibit systematic temporal variation rather than persistent offsets. In Warren County, the observed contrast reverses sign over the diurnal cycle. Marmot reproduces this reversal and much of its magnitude, whereas baselines predict weak contrasts. Marmot also recovers the seasonal trend in temperature contrast most closely among the methods. In San Diego County, Marmot captures the dewpoint diurnal evolution better than baselines, whereas the seasonal magnitude is overall best reproduced by HRRR. Lastly, the Austin wind-speed contrast is partially recovered by Marmot, but HRRR recovers the seasonal contrast better.

Selected weather events provide additional evidence of temporal recovery (Fig. S6). In a valley in central Idaho, Marmot captures a large diurnal temperature range that is underestimated by the baselines, reducing RMSE to 1.76°C compared to 3.22°C for the next-best method. During a Santa Ana event in San Diego, observed dewpoint falls rapidly by more than 20°C; Marmot captures both

the pre-event dewpoint and the subsequent drying, reducing RMSE to 1.97°C versus 7.15°C for the next-best baseline. Finally, during a cold-front passage along the Texas Gulf Coast, Marmot follows the frontal wind shift with the lowest vector error (1.43 m s^{-1} versus 2.18 m s^{-1} next-best), although its timing is delayed relative to HRRR.

9.3 Additional Physical Structure of Recovered Micro-weather

The main text presents examples of inferred micro-weather associated with topography and land cover. Here, we examine additional cases in which these relationships vary with atmospheric conditions, including temperature inversions, urban heat, and terrain-channeled wind.

In Sonoma Valley, California, the relationship between temperature and elevation varies with atmospheric conditions (Fig. S8). Across 4,380 nighttime hours in 2023, 57% exhibit a temperature inversion — defined here as a positive relationship between temperature and elevation — across 17 stations spanning 8–723 m elevation, compared with 24% during daytime. The occurrence of nighttime inversions is strongly related to wind speed: 79% of nighttime hours with winds below 1 m s^{-1} are inverted, compared with only 1% above 3 m s^{-1}. During a calm night on 23 August, observations show a positive lapse rate of +0.54°C per 100 m, with warmer temperatures at higher elevations; Marmot reproduces this inversion and its spatial structure. During a daytime example under stronger winds, observations instead show the conventional decrease in temperature with elevation, which Marmot also recovers. Contemporaneous ECOSTRESS observations exhibit similar spatial structure in both examples (Fig. S8). Together, these cases show that Marmot does not impose a fixed relationship between temperature and elevation, but can recover changes in the sign and magnitude of this relationship under different atmospheric conditions.

A second example illustrates recovery of urban temperature contrasts (Fig. S9). In Syracuse, New York, a held-out station in a 53%-built environment is on average 1.49 °C warmer at night than a held-out station 6.8 km away in an 11%-built environment at the same elevation, a contrast that falls to 0.74 °C by day. Marmot partially recovers this spatial contrast and more closely estimates temperature at the urban station than the baseline methods, although all methods underestimate its magnitude. The inferred field delineates fine-scale temperature structure associated with the urban surface that is absent from station interpolation and the coarser atmospheric products, with qualitatively similar structure visible in contemporaneous ECOSTRESS land surface temperature.

We additionally test whether Marmot recovers terrain channeling of near-surface wind (Fig. S10). We first identify valley stations among the 504 held-out stations with sufficient wind obser-

vations, using the DEM to determine whether a station is bounded by higher terrain and, if so, the orientation and degree of topographic confinement of the valley. The observed average wind direction increasingly aligns with the valley orientation as confinement increases, with the median angular difference decreasing from 38° in weakly confined valleys (channel walls are < 50 m height; $n = 228$) to 16° in strongly confined valleys (> 150 m height; $n = 29$). We therefore focus on the most strongly confined stations where observed winds clearly follow the terrain. In 9 out of the 15 most highly confined stations, Marmot achieves the lowest wind-vector RMSE. However, even in those cases, it only partially recovers the strong directional confinement seen in the observations. Wind channeling through narrow valleys therefore exposes a remaining limitation shared by all methods. Figure S10 shows two of Marmot's relative successes and two cases in which another method performs better.

9.4 Failure Modes

Marmot does not recover fine-scale weather equally well across all locations and atmospheric conditions. We examine cases in which Marmot is outperformed by baseline methods to better understand its failure modes.

One recurring failure occurs at remote locations that are far from backbone stations and experience large departures from surrounding conditions. In the Camas Prairie, Idaho, the nearest backbone station is 24 km from the held-out station. During a winter cold-pool event, Marmot substantially underestimates the magnitude of local cooling, with an RMSE of 6.23°C compared with 3.41°C for HRRR (Fig. S11); dewpoint at the same station exhibits a similar pattern. At a high-elevation site in the southern Sierra Nevada, California (2,743 m), Marmot likewise misses repeated declines in dewpoint that are captured more accurately by HRRR (Marmot RMSE 6.77°C versus HRRR 3.80°C). These examples could reflect limited information from distant backbone observations, limited training stations in comparable settings, difficulty in recovering locally generated atmospheric processes, or combinations of these.

A second failure mode occurs at some nearshore locations in the East Bay of California, where weather transitions sharply between marine and inland conditions and neither nearby backbone stations nor ERA5 estimate the weather well (Fig. S12). At a station in the Berkeley Hills, California, where the neighboring backbone stations are landward, Marmot overestimates temperature with an RMSE of 4.75°C, while HRRR achieves an RMSE of 2.04°C. At a nearby station on the Richmond shoreline, Marmot also underestimates wind speed, while HRRR is more accurate (vector

RMSE 4.60 versus 1.82 m s^{-1}). We hypothesize that the asymmetric placement of stations relative to the bay (since this work does not include measurements over bodies of water), steep gradients in weather in the East Bay, and/or the poor baseline performance of nearby stations and ERA5 may make these conditions difficult to infer.

9.5 Information Source Ablation Results

The additive ablation in the main text evaluates how performance changes as information sources are progressively introduced to a station-only model (Fig. 2c; Section 8.6). We additionally perform subtractive ablations from the full Marmot model, removing individual information sources while holding the remaining model components fixed (Table S3). The results are broadly consistent: station observations provide the foundation for prediction, while elevation, Earth observation, ERA5, and temporal information each contribute additional information, with ERA5 and the temporal encoding providing the smallest marginal improvements. The relative contributions of elevation and Earth observation depend on the other information available to the model. Adding explicit elevation differences to the station-only model produces a large improvement, whereas in the full model, removing AlphaEarth embeddings has a slightly larger effect than removing elevation differences. This difference likely reflects overlapping information about topography contained in the two representations.

9.6 Sensitivity Analysis Results

Marmot is generally robust to the hyperparameter choices examined (Table S5; Section 8.7). Performance improves as the number of neighboring stations increases, with diminishing returns or slightly worse performance beyond 64 neighbors. Performance is similarly stable across neighbor-dropout rates, especially between 10% and 50%, with no single setting consistently optimal across temperature, dewpoint, and wind. Varying the ERA5 spatial context from a single grid cell to a 9×9 neighborhood also produces only small differences, with stencil sizes between 3×3 and 7×7 performing slightly better overall. The best AlphaEarth resolution — 10, 30, or 100 m — depends on the weather variable and metric, but differences between the three resolutions are within 1%. These results indicate that performance is not strongly dependent on the particular hyperparameter configuration used for the primary Marmot specification.

Beyond model hyperparameters, we also examine sensitivity to the spatial partitioning used for evaluation. As expected, Marmot performs better under the random station split, where training

and backbone stations can occur in the same ERA5 grid cells as held-out stations: RMSE decreases by 6% for temperature, 6% for dewpoint, and 2% for wind relative to the by-cell split (Table S2). The random split provides a complementary estimate of performance when nearby station information is not deliberately excluded at test locations, whereas our primary grid-cell split tests the setting of spatial extrapolation beyond the backbone network.

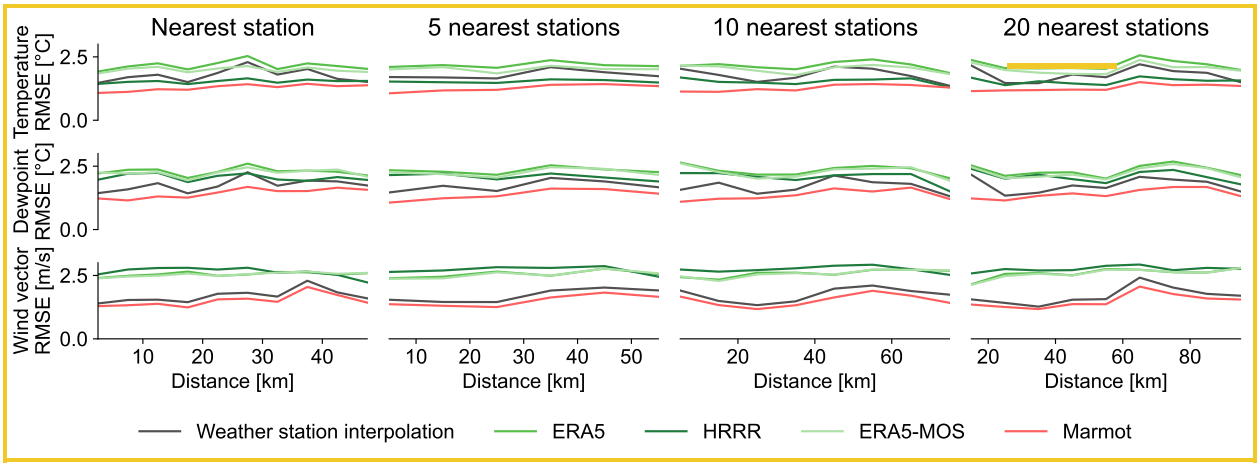


Figure S1: **Prediction error as a function of distance to the nearest k backbone stations,** for $k = 1, 5, 10$, and 20. For $k > 1$ the distance is the mean over the k nearest backbone stations. Across the 555 test stations, these distances span 2.1–71.6 km ($k = 1$; median 15.7 km), 4.0–101.3 km ($k = 5$; median 23.2 km), 6.5–158.9 km ($k = 10$; median 28.7 km), and 10.0–228.2 km ($k = 20$; median 39.6 km). Distance bins are 5 km wide from 0–50 km for $k = 1$; 10 km wide from 0–60 km for $k = 5$; 10 km wide from 0–80 km for $k = 10$; and 10 km wide from 10–100 km for $k = 20$.

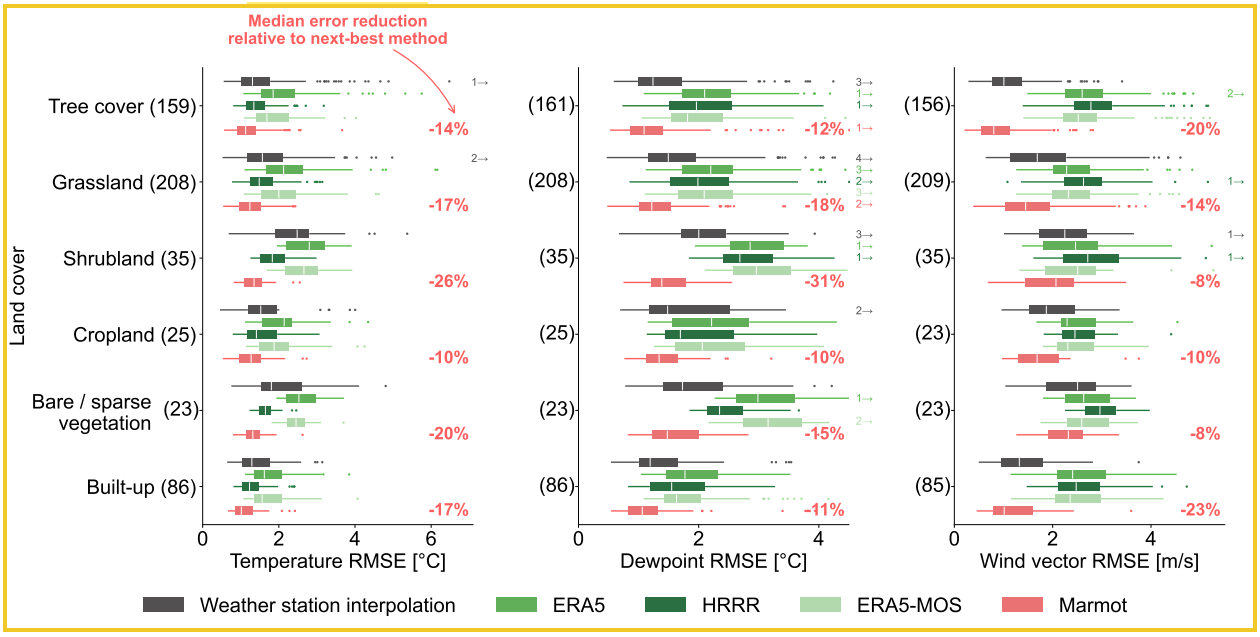


Figure S2: **Distribution of prediction error by land cover type.** Land cover labels are derived from the ESA WorldCover 10 m v200, and we retain the 6 land cover categories with sufficient stations to summarize. Percentage labels show the median error reduction by Marmot relative to the next-best baseline method for each land cover type and weather variable.

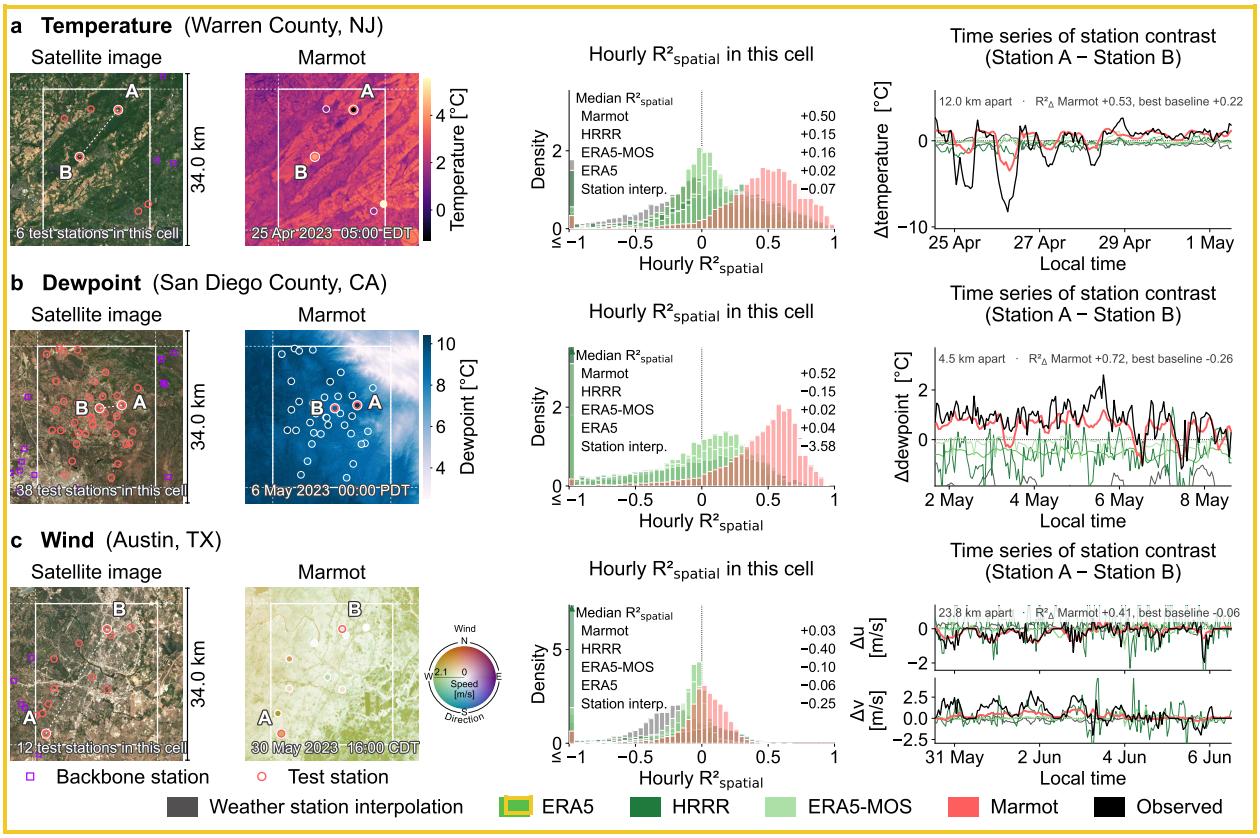


Figure S3: **Case studies of sub-grid spatial variance recovery.** Examples are shown for **(a)** air temperature in Warren County, New Jersey, **(b)** dewpoint in San Diego County, California, and **(c)** wind in Austin, Texas. For each case study, panels show satellite imagery and the inferred micro-weather field within a single ERA5 grid cell, the distribution of hourly R^2_{spatial} within the grid cell across 2023, and a time series of the observed and predicted weather contrast between two selected held-out stations (A and B). Station contrasts are calculated as the difference in weather conditions between stations A and B and evaluate each method's ability to reproduce time-varying sub-grid spatial differences.

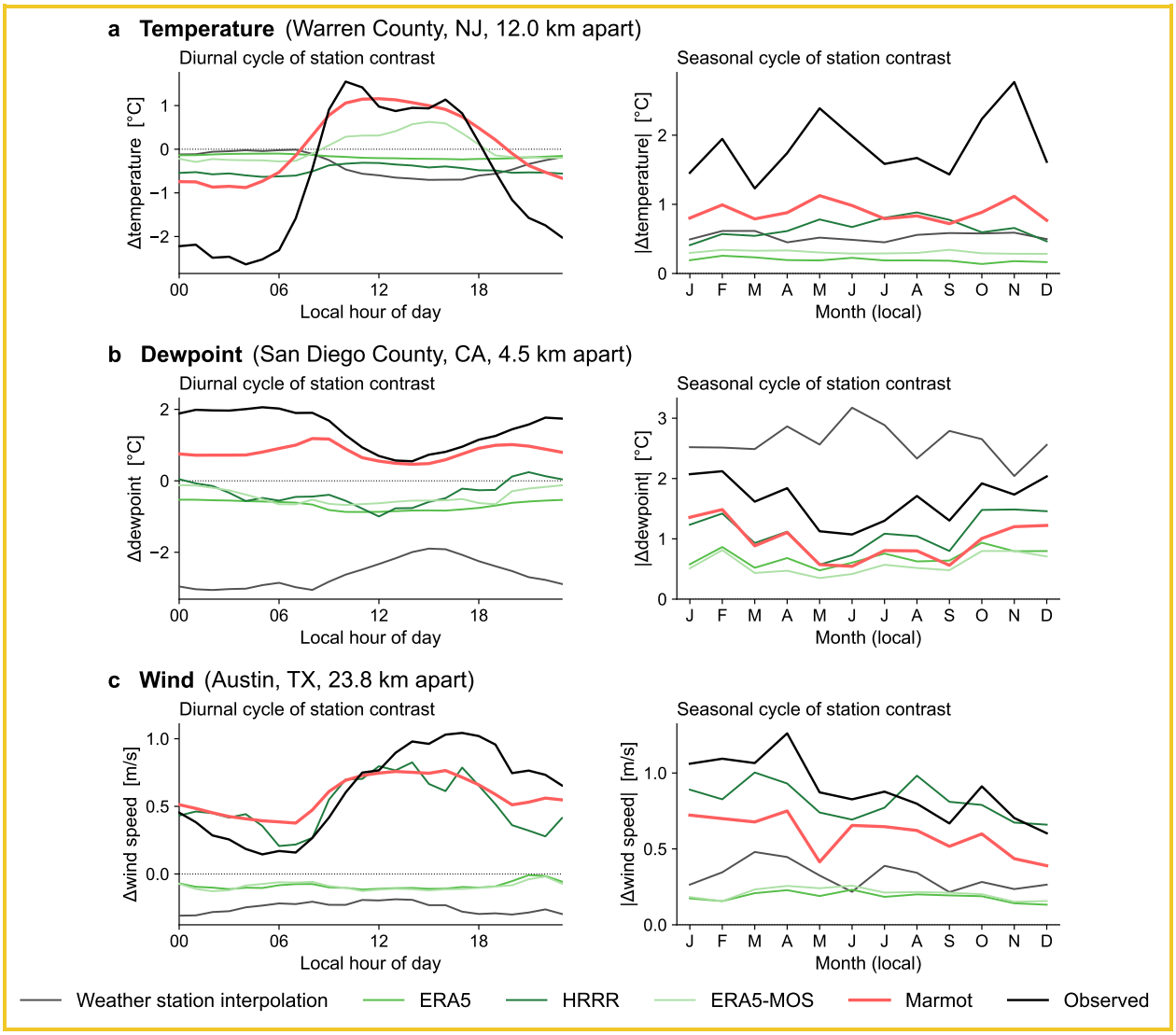


Figure S4: **Diurnal and seasonal patterns in local weather contrasts.** Observed and modeled contrasts are aggregated by local hour of day (left) and local month (right) for the three station pairs shown in Figure S3: **(a)** temperature in Warren County, New Jersey; **(b)** dewpoint in San Diego County, California; and **(c)** wind speed in Austin, Texas. Diurnal panels show signed station contrasts, whereas seasonal panels show absolute contrasts.

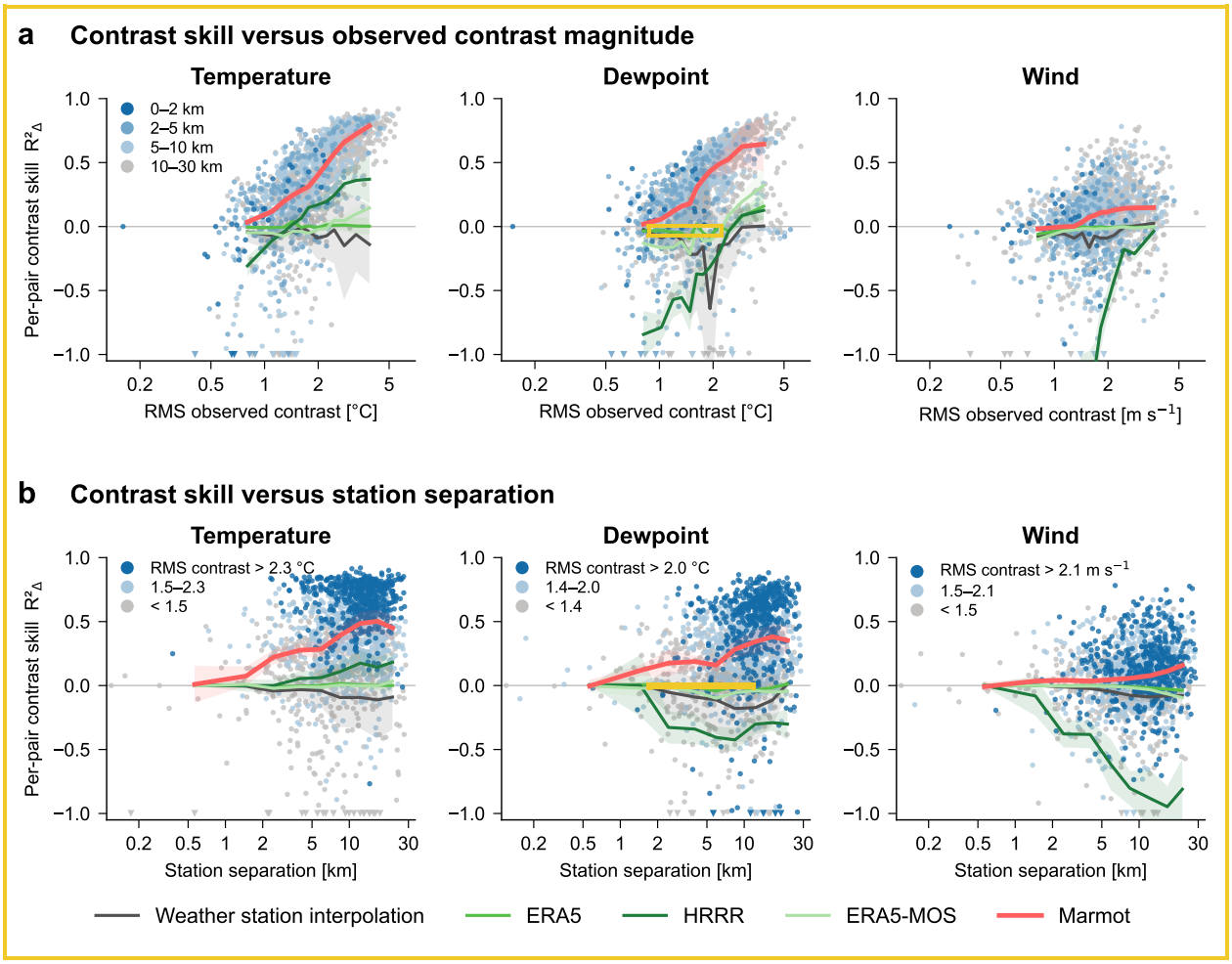


Figure S5: **Station pair-level contrast skill as a function of observed contrast magnitude and station separation.** Contrast skill is calculated separately for each pair of held-out stations using valid observations for that pair in 2023. Each point therefore represents one station pair. **(a)** Scatter plot of pair-specific Marmot skill as a function of the RMS observed contrast for temperature, dewpoint, and wind. Point color denotes station separation. Curves show running median pair-level skill for Marmot and the baseline methods. **(b)** Scatter plot of pair-specific Marmot skill as a function of station separation. Point color denotes the observed RMS contrast. Downward triangles at the lower axis boundary indicate scores below the displayed range.

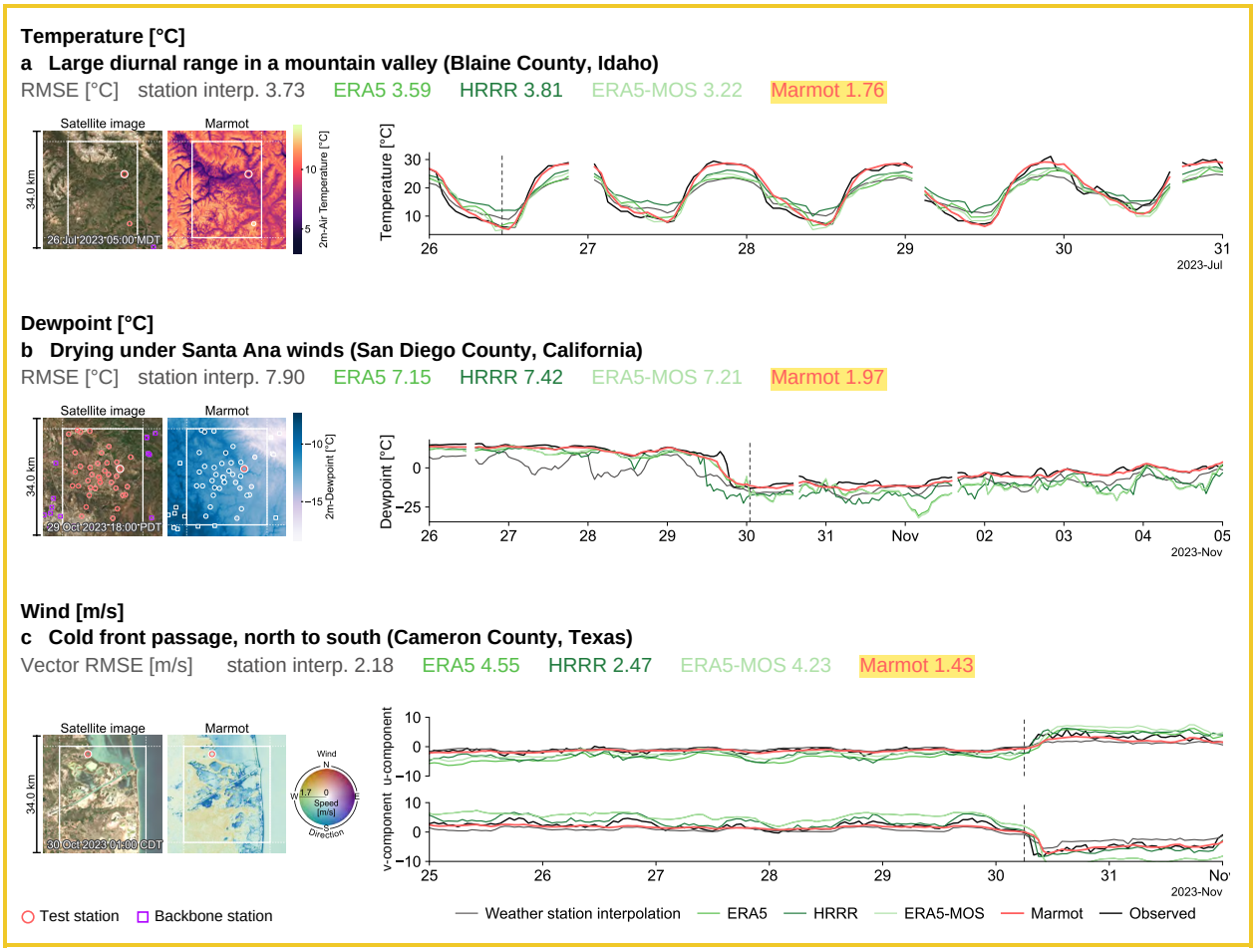


Figure S6: **Temporal behavior of inferred micro-weather at example held-out stations.** Hourly time series compare weather station interpolation, ERA5 reanalysis, HRRR analysis, ERA5-MOS and Marmot against station observations during 2023. **(a)** Station located on a valley floor between mountain ranges (Blaine County, Idaho). **(b)** Station located in inland terrain during a Santa Ana wind event (San Diego County, California). **(c)** Station located on the coastal plain during a cold-front passage (Cameron County, Texas).

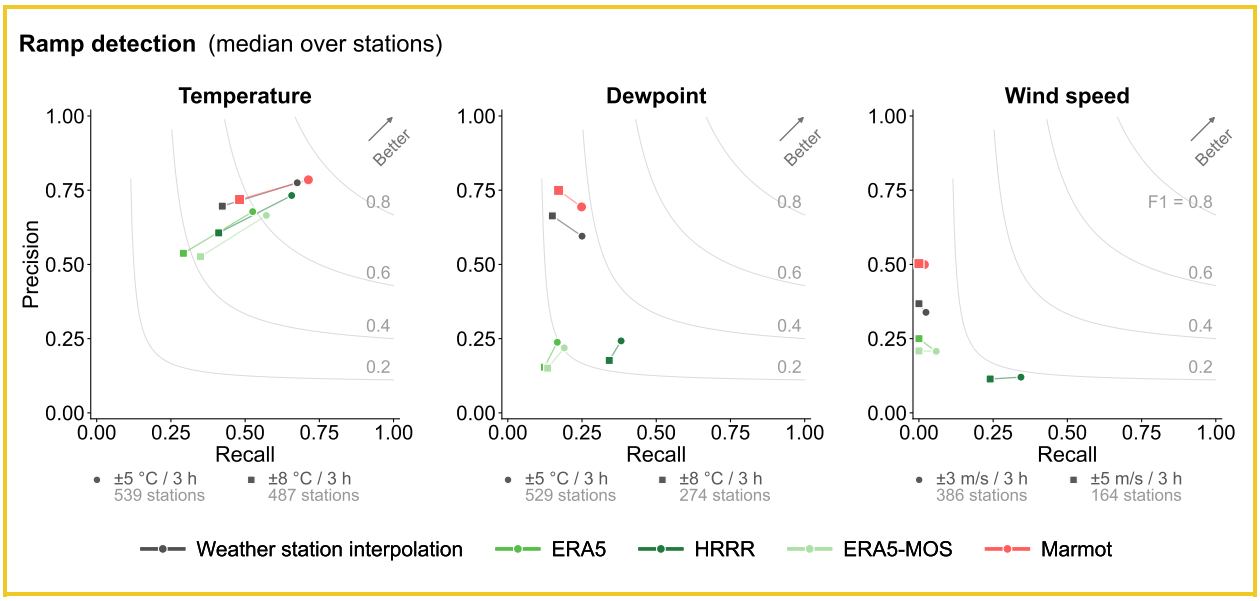


Figure S7: **Detection of ramps (changes) in weather.** Precision and recall for detection of changes over 3 hour intervals. Points denote different magnitude thresholds; gray curves indicate lines of constant F1 score.

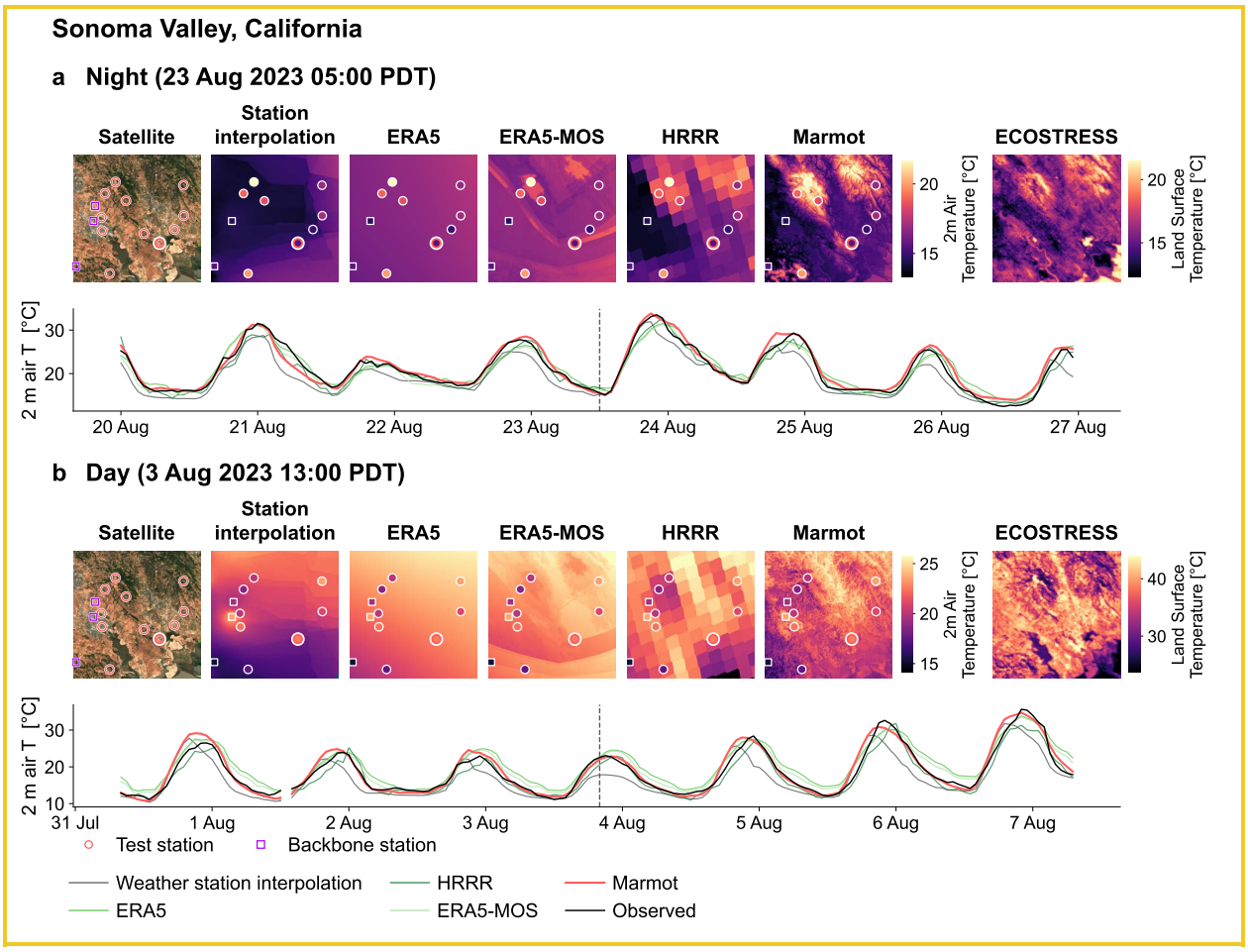


Figure S8: **Example of temperature inversion in Sonoma Valley, California. (a)** Nighttime temperatures observed at weather stations show a temperature inversion on August 23, 2023, in which higher elevations exhibit higher temperatures than lower elevations. **(b)** In the daytime on August 3, 2023, no temperature inversion is observed. Temperature fields generated by Marmot recover inversions at night and the lack thereof in the day. Satellite imagery, temperature fields predicted by baseline methods, and ECOSTRESS land surface temperature are also shown for comparison. Time series of temperature compare Marmot and baseline method predictions against observed values.

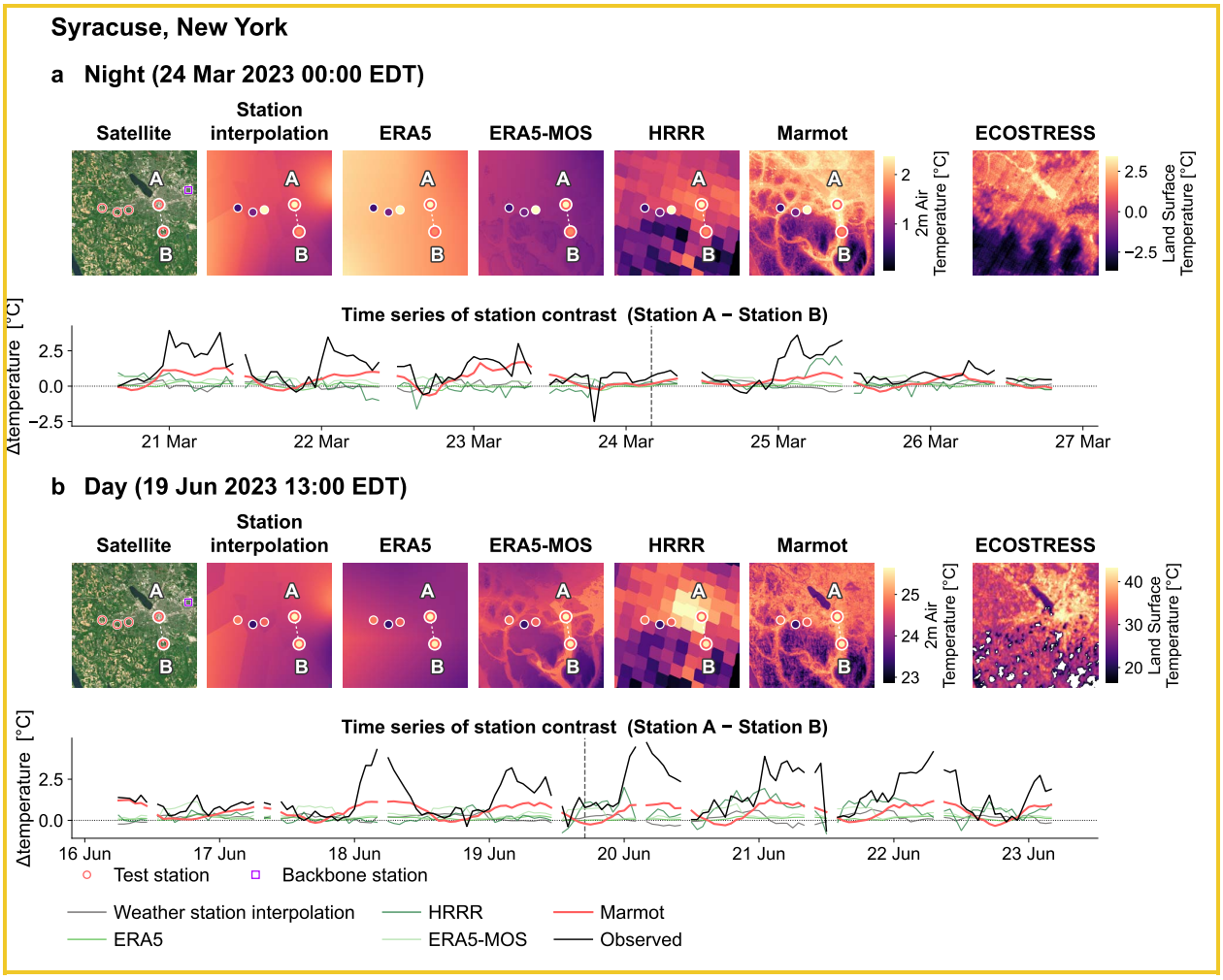


Figure S9: **Example of urban heat in Syracuse, New York. (a)** Nighttime and **(b)** daytime temperature contrast between Station A, a held-out station surrounded by built area, and Station B, a held-out station 6.8 km away at the same elevation surrounded by tree cover and grassland. Station A is the warmer of the two in 91% of the hours drawn in (a) and 95% in (b). Every method understates the contrast; Marmot recovers the largest share of it, 39% at night and 36% by day, against 32% and 30% for the best baseline.

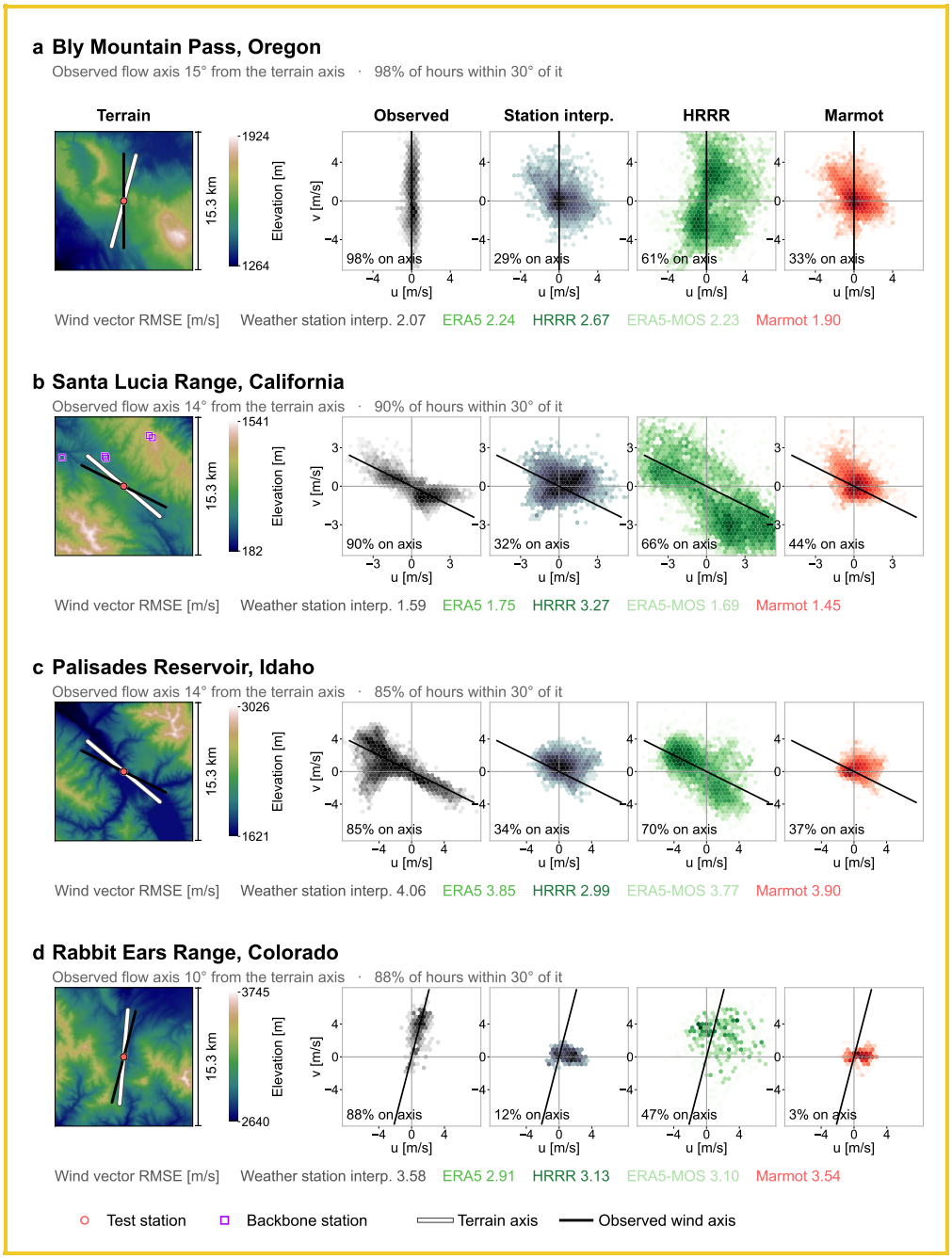


Figure S10: **Examples of wind channeling through valleys.** In **(a)** Bly Mountain Pass, Oregon and **(b)** Santa Lucia Range, California, Marmot's wind vector RMSE is lower than baseline methods'. In **(c)** Palisades Reservoir, Idaho and **(d)** Rabbit Ears Range, Colorado, one or more baseline methods outperform Marmot.

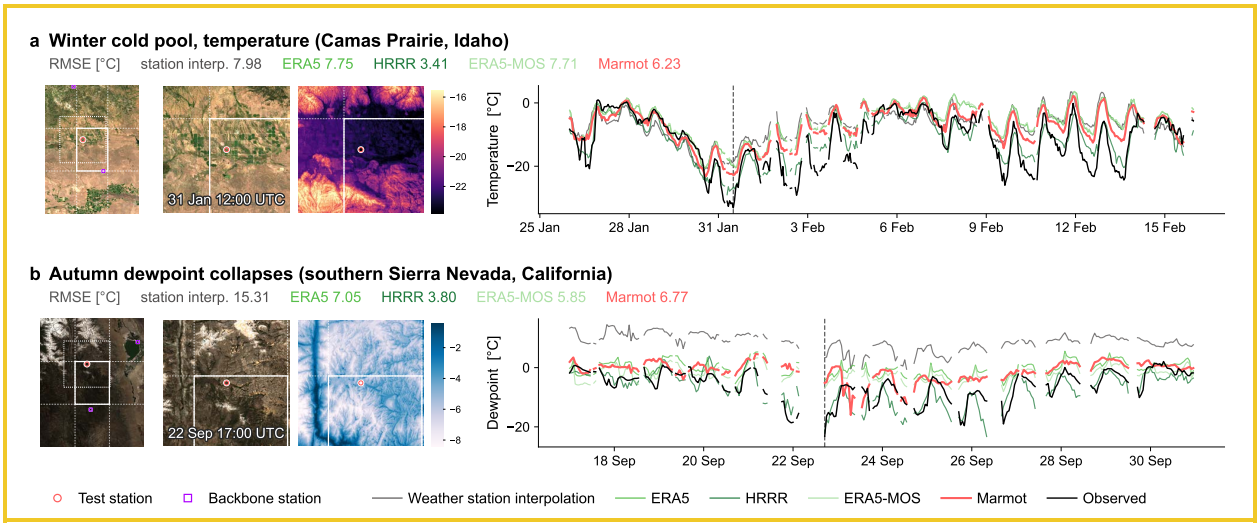


Figure S11: **Examples of Marmot failure in remote mountain valleys. (a)** Temperature during a winter cold-pool event in Camas Prairie, Idaho. **(b)** Dewpoint during repeated drying events in the southern Sierra Nevada, California. For each example, maps show the broader station context, satellite imagery, and Marmot prediction at the indicated time. Time series compare the held-out station observation with Marmot and baseline methods. RMSE is reported above each panel. In both examples, the held-out station is distant from the backbone network and experiences large declines that Marmot does not fully recover. HRRR has the lowest error.

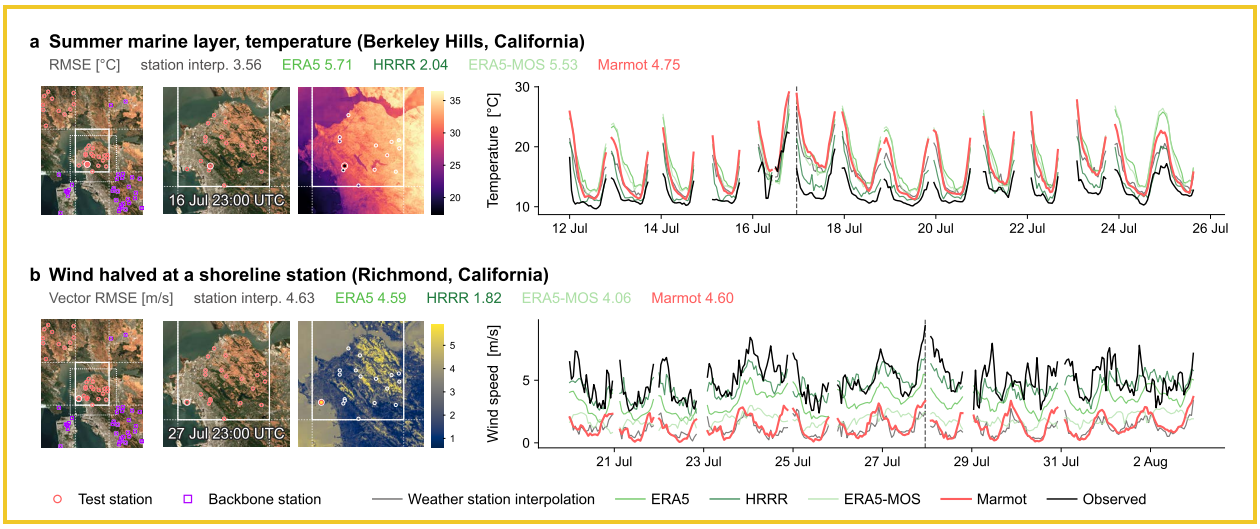


Figure S12: **Examples of Marmot failure in East Bay, California. (a)** Temperature at a held-out station in the Berkeley Hills, California, during summer marine-layer conditions. **(b)** Wind speed at a held-out station on the Richmond shoreline, California. For each example, maps show the broader station context, satellite imagery, and Marmot prediction at the indicated time. Time series compare the held-out station observation with Marmot and baseline methods. RMSE is reported above each panel. In both cases, nearby backbone stations lie landward of the target, and Marmot does not recover the cooler temperature or stronger winds observed at the nearshore stations. HRRR has the lowest error.

Table S1: **Model errors and spatial R^2 at held-out test stations.** All methods are scored on the 555 test stations in 2023, a year absent from training, so the evaluation is held out in space and in time. **(a)** RMSE and MAE of temperature, dewpoint, and the wind vector. **(b)** Wind error broken out into speed and direction. **(c)** Spatial variance explained is calculated for each time step over all test stations and then averaged over all time steps. We report one standard deviation over the 2,000 replicates of a nested bootstrap. Bold and underline mark the best and second-best point estimate in each column; overlapping intervals do not resolve an ordering.

(a) Error at held-out test stations

	Temperature [°C]		Dewpoint [°C]		Wind vector [m/s]	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE
Station interpolation	2.006 ±0.126	1.378 ±0.084	<u>2.008</u> ±0.192	<u>1.295</u> ±0.097	<u>1.720</u> ±0.055	<u>1.263</u> ±0.047
ERA5	2.277 ±0.066	1.680 ±0.048	2.419 ±0.077	1.749 ±0.051	2.613 ±0.085	2.232 ±0.083
HRRR	<u>1.574</u> ±0.040	<u>1.165</u> ±0.029	2.248 ±0.088	1.565 ±0.057	2.781 ±0.056	2.326 ±0.052
ERA5-MOS	2.093 ±0.065	1.535 ±0.047	2.329 ±0.085	1.655 ±0.055	2.570 ±0.081	2.195 ±0.079
Marmot	**1.295** ±0.034	**0.919** ±0.022	**1.455** ±0.063	**0.985** ±0.032	**1.539** ±0.056	**1.104** ±0.046

(b) Wind error at held-out test stations, broken out into speed and direction

	Wind speed [m/s]		Wind direction [°]	
Model	RMSE	MAE	RMSE	MAE
Station interpolation	<u>1.326</u> ±0.041	<u>0.898</u> ±0.030	67.38 ±1.40	49.54 ±1.36
ERA5	2.161 ±0.095	1.717 ±0.093	66.39 ±1.34	48.68 ±1.28
HRRR	2.319 ±0.058	1.837 ±0.059	<u>63.46</u> ±1.50	<u>45.53</u> ±1.36
ERA5-MOS	2.115 ±0.090	1.673 ±0.087	67.53 ±1.29	49.93 ±1.25
Marmot	**1.182** ±0.046	**0.780** ±0.033	**62.71** ±1.24	**45.22** ±1.16

(c) Spatial R^2 at held-out test stations

Model	Temperature	Dewpoint	Wind
Station interpolation	0.883 ±0.016	<u>0.903</u> ±0.023	<u>0.334</u> ±0.031
ERA5	0.863 ±0.012	0.874 ±0.012	−0.501 ±0.140
HRRR	<u>0.929</u> ±0.006	0.895 ±0.012	−0.675 ±0.135
ERA5-MOS	0.877 ±0.011	0.873 ±0.013	−0.465 ±0.134
Marmot	**0.951** ±0.004	**0.949** ±0.005	**0.464** ±0.021

Table S2: **Model and baseline errors under two station partitions.** The by-ERA5-cell split withholds whole grid cells, so held-out stations have no training or backbone station in their own cell; it measures spatial extrapolation and is the partition used throughout the main text. The random split withholds individual stations, so neighbors in the same cell may remain available. Because the two partitions are evaluated in different station-hours (n given per group), rows may be compared within a group but not across groups. We report one standard deviation over the 2,000 replicates of a nested bootstrap. Bold marks the best point estimate in each column.

	Temperature [°C]		Dewpoint [°C]		Wind vector [m/s]	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE
By-ERA5-cell split (used in paper; whole cells withheld together, no training station in target cell; n = 3,885,810)						
Marmot	**1.295 ±0.034**	**0.919 ±0.022**	**1.455 ±0.063**	**0.985 ±0.032**	**1.539 ±0.056**	**1.104 ±0.046**
Station interpolation	2.006 ±0.126	1.378 ±0.084	2.008 ±0.192	1.295 ±0.097	1.720 ±0.055	1.263 ±0.047
HRRR	1.574 ±0.040	1.165 ±0.029	2.248 ±0.088	1.565 ±0.057	2.781 ±0.056	2.326 ±0.052
ERA5	2.277 ±0.066	1.680 ±0.048	2.419 ±0.077	1.749 ±0.051	2.613 ±0.085	2.232 ±0.083
ERA5-MOS	2.093 ±0.065	1.535 ±0.047	2.329 ±0.085	1.655 ±0.055	2.570 ±0.081	2.195 ±0.079
Random split (stations drawn at random across the US; n = 3,928,815)						
Marmot	**1.213 ±0.027**	**0.847 ±0.016**	**1.372 ±0.033**	**0.950 ±0.022**	**1.509 ±0.044**	**1.066 ±0.030**
Station interpolation	1.717 ±0.063	1.158 ±0.035	1.671 ±0.045	1.150 ±0.030	1.744 ±0.058	1.243 ±0.035
HRRR	1.652 ±0.033	1.200 ±0.022	2.367 ±0.045	1.658 ±0.031	2.889 ±0.043	2.390 ±0.034
ERA5	2.438 ±0.065	1.765 ±0.043	2.581 ±0.051	1.868 ±0.034	2.684 ±0.046	2.273 ±0.035
ERA5-MOS	2.118 ±0.040	1.540 ±0.028	2.353 ±0.040	1.693 ±0.027	2.638 ±0.046	2.227 ±0.035

Table S3: **Component ablations of the micro-weather model. (a)** Additive ladder: the base row is a pure station-interpolation model; each row below adds one information source and inherits everything above it. The sources enter in the order: station elevation difference, two ERA5 features (innovation, then patch), Earth observation features (AlphaEarth), a temporal encoding, and a regularizer (neighbor dropout). **(b)** Leave-one-out: one component removed from the Marmot (full model), everything else kept, in that same order. Bold marks Marmot (full model), which is the last row of (a) and the first of (b). Both panels use the column header printed above (a). We report one standard deviation over the 2,000 replicates of a nested bootstrap.

	Temperature [°C]		Dewpoint [°C]		Wind vector [m/s]	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE
(a) Additive ladder: each row adds one information source on top of everything above						
Station interpolation	2.006	1.378	2.008	1.295	1.720	1.263
Station-only model	1.896 ±0.091	1.320 ±0.058	1.778 ±0.092	1.216 ±0.057	1.651 ±0.058	1.204 ±0.048
+ elevation difference	1.433 ±0.040	1.007 ±0.025	1.515 ±0.064	1.033 ±0.033	1.631 ±0.058	1.185 ±0.047
+ ERA5 innovation	1.428 ±0.038	1.002 ±0.024	1.507 ±0.063	1.028 ±0.032	1.603 ±0.056	1.175 ±0.045
+ ERA5 patch	1.426 ±0.037	1.002 ±0.024	1.503 ±0.063	1.027 ±0.033	1.598 ±0.055	1.177 ±0.045
+ AlphaEarth	1.337 ±0.034	0.946 ±0.023	1.479 ±0.064	1.000 ±0.033	1.545 ±0.055	1.110 ±0.046
+ temporal encoding	1.308 ±0.034	0.926 ±0.022	1.464 ±0.062	0.995 ±0.031	1.541 ±0.055	1.106 ±0.046
+ neighbor dropout (Marmot)	**1.295 ±0.034**	**0.919 ±0.022**	**1.455 ±0.063**	**0.985 ±0.032**	**1.539 ±0.056**	**1.104 ±0.046**
(b) Leave-one-out: one component removed from the Marmot (full model)						
Marmot	**1.295 ±0.034**	**0.919 ±0.022**	**1.455 ±0.063**	**0.985 ±0.032**	**1.539 ±0.056**	**1.104 ±0.046**
− station observations	1.682 ±0.036	1.244 ±0.026	1.904 ±0.066	1.350 ±0.043	1.626 ±0.064	1.172 ±0.052
− elevation difference	1.396 ±0.034	1.012 ±0.025	1.511 ±0.063	1.030 ±0.031	1.549 ±0.057	1.115 ±0.047
− ERA5 innovation	1.312 ±0.035	0.931 ±0.023	1.468 ±0.064	0.992 ±0.032	1.552 ±0.056	1.110 ±0.046
− ERA5 patch	1.295 ±0.034	0.916 ±0.022	1.461 ±0.063	0.990 ±0.031	1.540 ±0.055	1.102 ±0.046
− AlphaEarth	1.410 ±0.037	0.993 ±0.024	1.492 ±0.064	1.013 ±0.032	1.593 ±0.055	1.172 ±0.045
− temporal encoding	1.318 ±0.034	0.934 ±0.022	1.468 ±0.064	0.994 ±0.032	1.546 ±0.056	1.108 ±0.047
− neighbor dropout	1.308 ±0.034	0.926 ±0.022	1.464 ±0.062	0.995 ±0.031	1.541 ±0.055	1.106 ±0.046

Table S4: **Ablation of Earth observation information sources.** The first row begins with the model with no Earth observation information and the last row, marked in bold, is the Marmot model used throughout the paper. The rows are grouped into single hand-built sources, combinations of them, and the learned embedding. "Cols" is the number of feature columns each input contributes. We report one standard deviation over the 2,000 replicates of a nested bootstrap.

		Temperature [°C]		Dewpoint [°C]		Wind vector [m/s]	
Surface information	Cols	RMSE	MAE	RMSE	MAE	RMSE	MAE
No surface information	0	1.410 ±0.037	0.993 ±0.024	1.492 ±0.064	1.013 ±0.032	1.593 ±0.055	1.172 ±0.045
Single source							
Terrain	5	1.400 ±0.037	0.990 ±0.025	1.492 ±0.064	1.014 ±0.033	1.588 ±0.056	1.161 ±0.045
Land cover	7	1.390 ±0.038	0.984 ±0.025	1.490 ±0.063	1.011 ±0.033	1.581 ±0.057	1.147 ±0.047
Sentinel-2	48	1.368 ±0.035	0.965 ±0.023	1.482 ±0.063	1.012 ±0.032	1.576 ±0.054	1.140 ±0.046
Combined hand-built sources							
Land cover + terrain	12	1.390 ±0.039	0.981 ±0.025	1.486 ±0.063	1.008 ±0.032	1.578 ±0.057	1.143 ±0.046
All three	60	1.371 ±0.036	0.971 ±0.024	1.480 ±0.063	1.007 ±0.032	1.580 ±0.056	1.143 ±0.046
Learned embedding							
AlphaEarth 30 m (Marmot)	**64**	**1.295** ±0.034	**0.919** ±0.022	**1.455** ±0.063	**0.985** ±0.032	**1.539** ±0.056	**1.104** ±0.046

Table S5: **Sensitivity to model design choices.** Each row is the Marmot (full model) with one setting changed; the row marked "Marmot" is the configuration used throughout the paper and repeats identically in every panel. **(a)** Number of backbone neighbors k attended to. **(b)** Fraction of neighbors randomly dropped during training. **(c)** Size of the ERA5 stencil centered on the target. **(d)** Resolution of the AlphaEarth embedding. Bold marks the best point estimate in each column within a panel. We report one standard deviation over 2,000 replicates of a nested bootstrap, scored on the 854 validation stations over 2020–2022 rather than the test partition.

	Temperature [°C]		Dewpoint [°C]		Wind vector [m/s]	
Setting	RMSE	MAE	RMSE	MAE	RMSE	MAE
(a) Neighbor count k, with and without the embedding						
$k = 16$	1.402 ±0.033	0.998 ±0.025	1.577 ±0.039	1.104 ±0.026	1.663 ±0.049	1.188 ±0.037
$k = 32$	1.388 ±0.034	0.987 ±0.025	1.559 ±0.037	1.094 ±0.024	1.661 ±0.049	1.185 ±0.037
$k = 64$ (Marmot)	1.378 ±0.034	**0.981 ±0.025**	**1.547 ±0.036**	**1.085 ±0.023**	1.658 ±0.049	**1.183 ±0.038**
$k = 128$	**1.376 ±0.033**	**0.981 ±0.025**	1.549 ±0.036	1.088 ±0.024	**1.652 ±0.048**	1.185 ±0.038
(b) Neighbors dropped at training time						
0%	1.389 ±0.035	0.986 ±0.025	1.554 ±0.036	1.092 ±0.024	1.666 ±0.049	1.190 ±0.038
10%	1.380 ±0.034	**0.980 ±0.025**	1.551 ±0.036	1.087 ±0.024	1.661 ±0.049	1.188 ±0.038
25% (Marmot)	**1.378 ±0.034**	0.981 ±0.025	**1.547 ±0.036**	**1.085 ±0.023**	1.658 ±0.049	1.183 ±0.038
50%	**1.378 ±0.034**	0.983 ±0.025	1.548 ±0.035	1.086 ±0.023	**1.657 ±0.049**	**1.180 ±0.037**
60%	1.381 ±0.034	0.985 ±0.025	1.551 ±0.035	1.089 ±0.023	1.659 ±0.048	1.181 ±0.037
75%	1.387 ±0.033	0.992 ±0.024	1.563 ±0.035	1.101 ±0.024	1.664 ±0.049	1.187 ±0.038
(c) ERA5 stencil size						
1×1	1.389 ±0.035	0.986 ±0.026	1.554 ±0.036	1.092 ±0.024	1.663 ±0.049	1.185 ±0.038
3×3	**1.371 ±0.033**	**0.974 ±0.025**	1.549 ±0.037	1.088 ±0.024	1.658 ±0.049	1.184 ±0.038
5×5 (Marmot)	1.378 ±0.034	0.981 ±0.025	**1.547 ±0.036**	**1.085 ±0.023**	1.658 ±0.049	**1.183 ±0.038**
7×7	1.385 ±0.033	0.987 ±0.024	1.560 ±0.036	1.097 ±0.025	**1.657 ±0.049**	1.187 ±0.038
9×9	1.380 ±0.033	0.981 ±0.025	1.551 ±0.036	1.090 ±0.024	1.660 ±0.049	1.185 ±0.038
(d) AlphaEarth resolution						
10 m	1.382 ±0.034	0.982 ±0.025	1.550 ±0.036	1.088 ±0.024	**1.657 ±0.048**	1.185 ±0.037
30 m (Marmot)	1.378 ±0.034	0.981 ±0.025	**1.547 ±0.036**	**1.085 ±0.023**	1.658 ±0.049	**1.183 ±0.038**
100 m	**1.369 ±0.033**	**0.974 ±0.024**	1.549 ±0.036	1.089 ±0.023	1.663 ±0.049	1.189 ±0.037

Table S6: **Surface correlates of inferred micro-weather structure.** Fraction of spatial variance (R^2) in the inferred micro-weather field of each case study in Figure 4 explained by terrain elevation, land cover, and the two together.

Panel	Case	Terrain R^2	Land cover R^2	Together R^2
a	Diablo Range, California	**0.44**	0.20	0.49
b	Great Salt Lake Desert, Utah	0.05	**0.42**	0.44
c	Finger Lakes, New York	**0.91**	0.02	0.92
d	Antelope Valley, California	0.09	**0.34**	0.40
e	Sonoma Valley, California	**0.37**	0.08	0.49
f	Hunterdon County, New Jersey	0.19	**0.56**	0.59