



Deep learning-based high-resolution downscaling of terrain-driven near-surface winds

Nora Helbig¹, Florian Hammer¹, Gert-Jan Duine², Leila M.V. Carvalho², Charles Jones², Hugo Merzisen³, Isabelle Gouttevin³, and Sarah Barber¹

¹Eastern Switzerland University of Applied Sciences, Rapperswil, Switzerland

²University of California Santa Barbara, Santa Barbara, U.S.

³Université Grenoble Alpes, Université de Toulouse, Météo-France, CNRS, CNRM, Centre d'Études de la Neige, Grenoble, France

Correspondence: Nora Helbig (norahelbig@gmail.com)

Abstract. Spatio-temporal wind fields in mountainous terrain are complex due to strong flow modifications when wind interacts with topography. Machine-learning (ML)-based wind downscaling models offer a computationally efficient alternative to high-resolution atmospheric model simulations and to dynamical downscaling, but their transferability beyond the training regions and conditions in which they were developed remains largely untested. We present an enhanced version of DEVINE, a deep-learning-based downscaling model for near-surface terrain-forced wind flow, being trained on atmospheric model ARPS simulations over Gaussian Random Field synthetic topographies with a wide range of slopes and feature shapes. Incorporating wind direction-dependent input features enables rotational data augmentation and facilitates preprocessing. Using enhanced DEVINE, we systematically assess its performance across previously unseen terrain (synthetic, alpine, coastal mountain environments), wind regimes (terrain-forced flow, sea-breeze circulation, downslope windstorm), and driving model data, using both high-resolution atmospheric wind and ground station data as reference.

We find that DEVINE generalizes well to unseen mountain regions when flow is primarily terrain-forced. Its spatial wind speed mean absolute error (MAE) and correlation in the alpine and non-storm coastal environments are lower and higher, respectively, than the station-based results reported by previous ML wind downscaling studies. DEVINE's own station-based performance remains broadly comparable to, or slightly improves on, these studies' station-based results (e.g., MAE of 1.13 m s^{-1}), despite simultaneously transferring across new topography, wind regime, and driving data. Performance degrades under complex atmosphere-terrain interactions such as a downslope windstorm (e.g., MAE of 5.6 m s^{-1}); under non-storm conditions, however, DEVINE outperforms the dynamical model WRF-LES at stations for MAE (1.13 m s^{-1} versus 1.23 m s^{-1}) and matches it for normalized root mean square error (13 %). We further show that station-based and spatial-pattern evaluations can diverge substantially, including in the sign of bias, indicating that the evaluation basis itself shapes apparent performance, and that spatial evaluation may better separate genuine model limitations from local representativity effects.

Our findings clarify strengths and limitations of the enhanced DEVINE wind downscaling model under out-of-sample conditions, and more broadly underscore the value of systematically evaluating ML wind downscaling methods across scale and under unseen conditions.



1 Introduction

25 Wind fields in mountainous terrain exhibit strong spatial and temporal nonlinearities. These nonlinearities arise from interactions between terrain-forced flows, generated by the mechanical modification of large-scale winds by topography, and thermally driven flows, induced by horizontal temperature gradients (Whiteman, 2000). While very high-resolution atmospheric model simulations at few meter resolutions can represent these processes in detail (e.g. Raderschall et al., 2008; Mott and Lehning, 2010; Gerber et al., 2017, 2019), their high computational cost restricts applications to small domains and short time periods. 30 However, with 12-26 % of the Earth's land surface being mountainous (Körner et al., 2017), highly spatially and temporally resolved wind predictions in mountainous terrain are critical for informed decision-making in applications such as wind energy, wildfire and avalanche forecasting, snow hydrology, and atmospheric transport and dispersion modeling.

The spatial and temporal requirements of these applications vary widely, from long-term wind energy assessments to short-term wildfire forecasts. Yet across this range, only high-resolution atmospheric modeling Reynolds-averaged Navier-Stokes 35 (RANS) approaches at 5-25 m (Raderschall et al., 2008; Mott and Lehning, 2010) and large-eddy simulations (LES) at about 50 m (Vionnet et al., 2017; Gerber et al., 2018, 2019; Umek et al., 2021; Goger et al., 2022) could realistically represent mountain wind dynamics. Such resolutions, however, entail substantial computational cost and numerical challenges over steep, well-resolved terrain (e.g. Lundquist et al., 2010; Goger et al., 2019; Arthur et al., 2020; Serafin et al., 2024), motivating wind downscaling approaches that balance computational efficiency and spatio-temporal predictive detail across this range of 40 mountain wind applications.

Downscaling of low-resolution modeled wind fields to fine scales can be achieved using dynamical, diagnostic (physically based), and statistical or data-driven approaches. While downscaling has been applied across a wide range of spatial scales, this study focuses on downscaling to fine spatial resolutions of a few tens of meters, required to resolve realistic wind fields in complex mountainous terrain for the applications described above. Dynamical downscaling employs a prognostic atmospheric 45 model, such as a numerical weather prediction (NWP) or LES model, at progressively finer resolutions through grid nesting (Vionnet et al., 2017; Gerber et al., 2018, 2019; Umek et al., 2021; Goger et al., 2022). Diagnostic downscaling approaches instead use low-resolution forcing from atmospheric models or observations, either relying on conceptual representations of key processes with terrain-based scalings (Ryan, 1977; Liston and Sturm, 1998), or applying mass-conserving flow models such as WindNinja (Forthofer et al., 2014; Wagenbrenner et al., 2016) and the High-resolution Intermediate Complexity Atmospheric 50 Research (HICAR) model (Reynolds et al., 2023, 2024). These approaches diagnose high-resolution atmospheric flow from prescribed low-resolution forcing rather than prognostically solving the full atmospheric dynamics.

Statistical downscaling in mountainous terrain spans approaches from physically guided regression models to fully data-driven machine-learning (ML) methods that learn nonlinear relationships from training data. Regression-based methods have been used to extrapolate station observations using terrain exposure and wind predictors (e.g. Winstral et al., 2009). These 55 methods generate gridded wind fields through multilevel linear models incorporating terrain, low-resolution wind, and seasonal information (Winstral et al., 2017), or derive nonlinear regression models from large ensembles of atmospheric simulations over synthetic topographies to generalize wind-terrain interactions (Helbig et al., 2017, 2024). Data-driven deep-learning models,



particularly convolutional neural networks (CNNs), have proven well suited for high-resolution wind downscaling in mountainous terrain. Examples include WindTopo (Dujardin and Lehning, 2022), trained on station observations and NWP model COSMO (Consortium for Small-scale Modeling) output to produce 50 m near-surface u , v wind components over Switzerland, as well as DEVINE (Le Toumelin et al., 2023) and SNOWstorm (Saigger et al., 2026), both trained on controlled wind simulations over synthetic topographies to predict near-surface u , v , and w wind components. DEVINE was trained on several thousands of nonhydrostatic, compressible ARPS model (Advanced Regional Prediction System) (Xue et al., 2000) simulations at 30 m resolution over Gaussian Random Field (GRF) synthetic topographies (Helbig et al., 2017). SNOWstorm was trained on several hundreds of Weather Research and Forecasting (WRF) simulations (Skamarock et al., 2019) with LES diagnostics (Göbel et al., 2022) at 50 m resolution over power law spectra-based synthetic topographies.

Data-driven methods generally perform best when applied to conditions similar to their training data, i.e., for wind downscaling in a comparable wind regime and geographical region. WindTopo, trained on Swiss station observations and COSMO-1 output, is therefore likely most accurate within Switzerland, but offers no guarantee of transferability elsewhere and has a fixed grid spacing of 53 m. DEVINE and SNOWstorm, in contrast, are trained on controlled simulations over synthetic topographies, making them more generalizable surrogates for high-resolution downscaling in other mountainous regions, though their performance still depends on the diversity of that training data. DEVINE’s GRF topographies generalize across a wide range of real slope and shape combinations with local slopes up to 70° (Helbig and Löwe, 2012; Le Toumelin et al., 2023) and are not tied to a fixed grid resolution; the underlying ARPS simulations ran under neutral stratification to isolate topography-induced non-turbulent mean flow (Wood and Mason, 1993), representative of snow-covered terrain (Cullen et al., 2007). SNOWstorm instead used WRF-LES simulations with atmospheric conditions typical of mid- to high-latitude winters resolving large turbulent motions absent from DEVINE’s training data; SNOWstorm’s synthetic topography spectral characteristics are constrained from several glaciated mountain ranges in mid- to high latitudes and have maximum local slopes of 40° .

These training design choices reflect different priorities: WindTopo gains regional accuracy within Switzerland at the cost of a fixed 53 m grid resolution and untested transferability elsewhere; DEVINE gains resolution- and topography-flexibility at the cost of a quasi-steady, non-turbulent mean flow under neutral conditions, while SNOWstorm resolves large-scale turbulence and a wider range of atmospheric conditions at the cost of a fixed 50 m grid resolution and specific spectral real terrain characteristics. Beyond the training-data differences, a common limitation applies to all current CNN-based high-resolution wind models at up to 50 m grid spacing (Dujardin and Lehning, 2022; Le Toumelin et al., 2023, 2024; Saigger et al., 2026): they primarily capture near-surface terrain-forced winds, sheltering, speed-up, deflection, and channeling, while thermally-driven or synoptically driven flows can only be represented if already resolved in the low-resolution forcing. Systematic errors in that forcing can likewise persist or be amplified during downscaling, motivating bias correction as a mitigation strategy (e.g., Le Toumelin et al., 2024).

Given its high generalizability across mountain regions and flexibility in grid design, we focus on DEVINE (Le Toumelin et al., 2023) in this study and present an enhanced version of the model. DEVINE downscales near-surface low-resolution wind components (u , v , w) to high-resolution topographies. The original model effectively represents wind flow when evaluated against station observations throughout the French Alps, showing reduced biases at elevated and exposed sites compared



to low-resolution forcing (Le Toumelin et al., 2023), and performed comparably well when evaluated against station observations in the French Pyrenees and Corsica, leading Le Toumelin et al. (2024) to conclude that DEVINE requires no additional calibration to be applied across Alpine-like topographies. Its performance in other wind regimes, and when evaluated against independent high-resolution atmospheric simulations, has not yet been systematically investigated. An understanding of the wind regimes DEVINE can represent is essential for quantifying its uncertainties and assessing its suitability across diverse environmental applications in mountainous terrain. We therefore broaden the model's applicability: we introduce two wind-direction-dependent features to enable rotational data augmentation, reformulate elevation as a relative feature variable, retrain the model on an augmented dataset, and migrate DEVINE to PyTorch, improving computational efficiency, flexibility, and GPU compatibility. Extending the point-wise station evaluations of the original model, we evaluate the enhanced model against spatio-temporally resolved high-resolution atmospheric simulations in two contrasting environments: first, using independent ARPS simulations over a real Alpine topography under neutral conditions to assess terrain-forced flow predictions, and second, using WRF-LES simulations in Southern California under mixed terrain- and synoptic, pressure-driven wind conditions to understand DEVINE's limitations. In Southern California, we further evaluate simulated wind fields against ground station data from two station networks as well as data collected during a field campaign. Section 2 describes training and evaluation data, Section 4 details the new technical developments enhancing DEVINE, and Section 5 presents the spatio-temporal evaluations highlighting strengths and limitations of the enhanced DEVINE downscaling model. Discussion and conclusion sections follow (Sections 6 and 7).

2 Training data

2.1 GRF-ARPS dataset

Helbig et al. (2017) generated a large wind field dataset over synthetic Gaussian Random Field (GRF) topographies using the atmospheric model ARPS, originally designed to develop a subgrid terrain parameterization and statistical downscaling framework applicable across mountain regions and spatial scales. The near-surface winds of this GRF-ARPS dataset were later used directly to train the wind downscaling model DEVINE (Le Toumelin et al., 2023), while its 3D wind fields also supported the development of a statistical downscaling model for snowfall deposition (Helbig et al., 2024). The GRF-ARPS dataset, described in the following subsections, is split into training and validation subsets used to train and validate the original (Le Toumelin et al., 2023) and enhanced DEVINE model, with the two models using different splits. The validation subset also serves as the synthetic evaluation environment described in Section 3. Data augmentation and the training/validation split procedure for the enhanced DEVINE model are both described in the Methods section.

2.1.1 Synthetic GRF topographies

To study wind-terrain interactions, a large variability of terrain characteristics is required in the underlying topographies used for wind simulations, i.e., to build a training dataset diverse enough to achieve applicability across mountain regions. Synthetic



topographies allow terrain properties to be varied systematically and independently of confounding real-world factors. Gaussian statistics were found to best reproduce the slope statistics of real topographies (Helbig and Löwe, 2012). The synthetic topographies of the GRF-ARPS dataset were therefore generated using Gaussian random fields (GRF), which implement such statistics through a prescribed Gaussian covariance. For a given domain mean slope, GRF topographies can be described by two characteristic length scales: a typical valley-to-peak elevation difference σ (feature height) and a correlation length ξ (feature width), allowing topographies of similar mean slope angle with different length-scale realizations.

The GRF dataset covers five domain mean slope angles (10° - 36°), each combined with nine width-height feature combinations (ξ : 200 m-1000 m, σ : 18 m-364 m) (see Table 2 in Helbig et al., 2017). Each of the resulting 45 slope-length-scale combinations was repeated with 199 random realizations, yielding 8955 synthetic topographies. Each topography (digital elevation model, DEM) has a square domain size length L of 3 km, with horizontal grid spacing Δx of 30 m. Across the full dataset, local slopes range from 0° to 70° . For further details on choice and generation of synthetic topographies, see Helbig et al. (2009); Helbig and Löwe (2012); Helbig et al. (2017), and for visual examples, see Fig. S1 or Le Toumelin et al. (2023).

2.1.2 ARPS simulations

Helbig et al. (2017) used the nonhydrostatic, compressible Advanced Regional Prediction System (ARPS) (Xue et al., 2000, 2001), originally developed for storm-scale numerical weather prediction, to simulate terrain-forced near-surface wind flow over the GRF topographies. ARPS has previously been shown capable of reproducing small-scale flow features such as speed-up, separation, and recirculation over steep terrain at spatial resolutions as fine as 5 m, in good agreement with observations (Raderschall et al., 2008; Mott and Lehning, 2010).

The simulations of Helbig et al. (2017) were designed to generate wind fields in which the mean flow is adapted to topography alone, i.e., representing terrain-forced flow only. To isolate the influence of topography on wind flow, thermally-induced circulations, such as mountain-valley breezes, and turbulent flow were suppressed by neglecting radiative forcing, setting atmospheric stability to neutral, and using a short total integration time of 30 s. The underlying surface was set to represent snow-covered terrain using a spatially homogeneous surface roughness length for snow of 0.01 m. ARPS was initialized with a spatially homogeneous vertical wind profile of 3 m s^{-1} and a fixed wind direction from the west, representing low-resolution forcing. Approximately 80 % of all ARPS simulations remained numerically stable. Near-surface ARPS wind speeds were extracted from the first model layer, with an average height of 2.95 m above the ground. The resulting near-surface, quasi-steady, non-turbulent 30 m wind fields for u , v and w were used to train the wind downscaling model DEVINE (Le Toumelin et al., 2023). A small number of simulations were excluded due to insufficient topographic variability or excessively steep domain mean slopes, which compromise numerical simulation stability. As a result, the exact number of usable ARPS simulations varies slightly across studies using this dataset, for example, 7232 in Helbig et al. (2017), 7279 in Le Toumelin et al. (2023), and 7859 in this study. Le Toumelin et al. (2023) report dataset characteristics such as wind speed accelerations of up to four times the initial forcing wind speed, reductions to near-zero wind speeds, and angular deviations between 0° and 82° from the prescribed forcing direction, and show corresponding distributions and example maps.



Table 1. Overview of models and data used across the three independent evaluation environments.

	Synthetic	Alpine	Coastal mountains
Low-resolution driving input	Prescribed uniform 3 m s^{-1} , West	COSMO-1-derived wind speed, AWS wind direction	1 km–WRF
High-resolution reference wind	30 m–ARPS (GRF-ARPS validation set, 20 %)	25 m–ARPS (Helbig et al., 2024)	111 m–WRF-LES (spatial); 56 stations (point)
DEVINE variant(s) evaluated	DEVINE PyTorch, DEVINE retrained, DEVINE low-bias, DEVINE low-scatter	DEVINE low-bias	DEVINE low-bias

3 Independent evaluation data

To understand the strengths and limitations of the enhanced DEVINE wind downscaling model, we evaluate its wind field predictions systematically across synthetic, alpine, and coastal mountain environments of increasing wind complexity (Table 1).
 160 First, in the synthetic environment, we assess model performance on a GRF-ARPS validation dataset representing terrain-forced flow only. Second, in the alpine environment, we extend this evaluation to real topography in the Eastern Swiss Alps, using independent ARPS simulations to test whether terrain-forced flow can be reproduced spatially in a real alpine setting, still without other wind regimes. Third, in the coastal mountain environment in Southern California, we evaluate the model against WRF-LES simulations and station observations, using an atmospheric model different from the one used for training, under
 165 mixed wind regimes beyond terrain-forced flow, largely dominated by synoptic, pressure-driven conditions during downslope windstorms.

3.1 Synthetic evaluation environment

For an independent evaluation in the synthetic environment, we employ an unused validation subset (20 %) of the GRF-ARPS dataset, randomly sampled within combinations of characteristic feature width ξ and domain-mean slope of the GRF topogra-
 170 phies before data augmentation (Section 4). This stratified sampling ensures that topographic characteristics are balanced across the training and validation subsets. An example topography from this dataset is shown in Fig. 1a.

3.2 Alpine evaluation environment

To evaluate enhanced DEVINE wind downscaling under realistic conditions in a real topography, we use an inner-Alpine topography in the Eastern Swiss Alps above Davos (Fig. 1b). The model topography is based on a 10 cm resolution WingtraOne
 175 unmanned aerial system (UAS) photogrammetric digital surface model (DSM) (Bühler et al., 2022), which was resampled to a grid spacing of $\Delta x = 25 \text{ m}$, covering $2.05 \times 2 \text{ km}^2$. This alpine topography has a domain mean slope angle of 37° (local slope angles range from 0° to 78°) and topographic characteristic length scales σ and ξ of 131 m and 262 m, respectively, falling well within the width-height feature scale and slope range of the GRF training topographies.



3.2.1 ARPS simulations

180 ARPS wind simulations at 25 m grid spacing were generated by Helbig et al. (2024) over the inner-Alpine topography (Fig. 1b). For the full input choices and preprocessing procedure, we refer to that study. Relevant to our evaluation of enhanced DEVINE is that five time steps were run using low-resolution input (one single forcing over the domain per time step) assembled for near-surface wind speed from the COSMO-1 ensemble forecasting system (analysis), after bias correction (Winstral et al., 2019), and for wind direction from an exposed automatic weather station (AWS) within the domain. The resulting wind speed and direction combinations were: $2.3 \text{ m s}^{-1} / 200^\circ$, $1.3 \text{ m s}^{-1} / 235^\circ$, $1.3 \text{ m s}^{-1} / 248^\circ$, $1.1 \text{ m s}^{-1} / 221^\circ$, and $1.4 \text{ m s}^{-1} / 290^\circ$. Helbig et al. (2024) used the same ARPS setup as for the GRF-ARPS dataset: they excluded thermal effects, suppressed turbulence, initialized with a spatially homogeneous low-resolution wind field, and assumed snow-covered terrain by setting the roughness length to that of snow (0.01 m). To exclude unrealistic ARPS wind speeds on steep slopes near the model domain boundaries, we cropped five grid cells from each edge, following Helbig et al. (2017). This existing dataset therefore extends the controlled ARPS experiments for terrain-forced flow performed on synthetic topographies to a real topography, under different low-resolution forcing wind speeds and directions than used in training. We use these high-resolution ARPS wind fields as evaluation data to assess enhanced DEVINE, downscaling the same five low-resolution inputs and comparing the resulting predictions against the corresponding spatial ARPS simulations.

3.2.2 Coastal mountain evaluation environment

195 Unlike the synthetic and inner-alpine environments, this evaluation targets a distinct wind regime and topographic setting. The study region lies in Santa Barbara County, Southern California, which exemplifies shallow (elevations below 1500 m), narrow, and steep coastal ranges (Fig. 1c). The 100 km east-west oriented Santa Ynez Mountains rise progressively toward the east, reaching elevations of roughly 1400 m, before descending sharply to the south to a narrow, populated coastal plain that meets the Pacific Ocean. Downslope windstorms are frequently observed along the southern slopes of the range. These gusty, dry, and warm lee-slope winds known as Sundowner winds (or "Sundowners") - for their typical onset near sunset - are primarily driven by cross-barrier pressure gradient and enhanced by reflected gravity-wave energy, via a self-induced critical level (Cannon et al., 2017; Duine et al., 2019; Carvalho et al., 2020), a mountaintop inversion (Blair, 1998; Seto et al., 2026), and/or a mean-state critical level in the approach flow (Cannon et al., 2017; Duine et al., 2025), with their relative importance varying by event (Blair, 1998; Cannon et al., 2017; Jones et al., 2021). These winds have been increasingly examined in recent years through modeling and observations and are recognized as a critical contributor to rapid wildfire spread in the region (e.g., Cannon et al., 2017; Duine et al., 2022; Zigner et al., 2022; Carvalho et al., 2024; de Orla-Barile et al., 2025). Although the Santa Ynez Mountains have only moderate elevations compared to alpine mountain ranges, the interaction of these downslope windstorms with the shallow, cool, and stable Pacific marine boundary layer adds further complexity to local wind regimes (Duine et al., 2019; de Orla-Barile et al., 2025; Seto et al., 2026).

210 The model topography used here is based on the USGS 3D Elevation Program (3DEP) 1/3 arc-seconds (approximately 10 m) DEM (U.S. Geological Survey, 2021), and was resampled to a grid spacing of $\Delta x = 100 \text{ m}$, covering $30 \times 30 \text{ km}^2$. All ocean

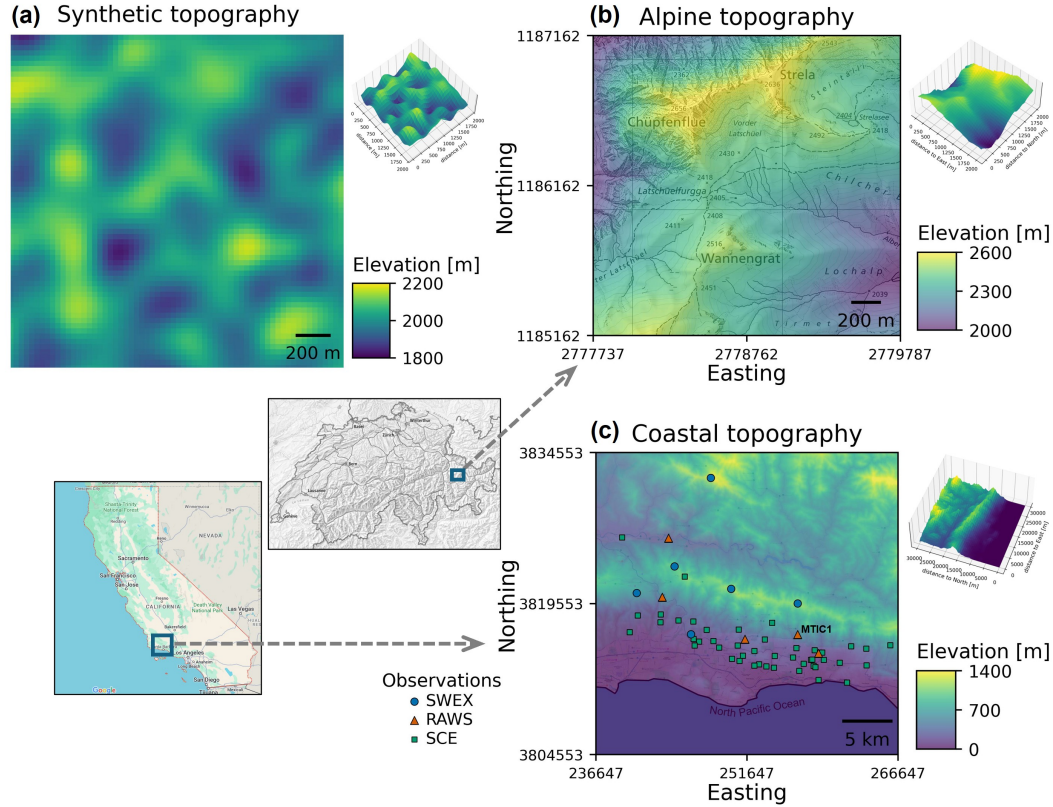


Figure 1. Overview of the topographies used to evaluate enhanced DEVINE: (a) Example synthetic topography from the GRF-ARPS validation dataset, with a domain mean slope angle of 36° , feature correlation length ξ of 200 m, and characteristic feature height σ of 73 m; (b) alpine topography in the eastern Swiss Alps; (c) coastal topography in the Santa Ynez Mountains, Southern California, with ground data locations used for evaluation indicated. Inset maps show the locations of the alpine and coastal study areas. Background maps: © swisstopo (Switzerland map and overview inset); © 2024 Google (California) and © 2026 Google (Santa Barbara), Google Maps.

grid cells are masked out for the evaluation. The topography has a domain mean slope angle of 18° (local slope angles range from 0° to 46°) and topographic characteristic length scales σ , ξ of 327 m and 654 m, respectively. The domain's mean slope angle and typical feature width ξ fall well within the range covered by the GRF training topographies; however, the typical feature height σ is substantially larger here than in GRF samples with comparable slope and ξ values.

We focus this study on one intensive observation period (IOP) of the Sundowner Winds Experiment (SWEX), a comprehensive field campaign that deployed multiple instrument platforms across the Santa Ynez Mountains between April 1 and May 15, 2022 (Carvalho et al., 2024). Specifically, we examined IOP 10, spanning 24 hours from May 12-13, 2022, during a Sundowner event that was particularly strong across the complex terrain of eastern Santa Ynez Mountains (de Orla-Barile et al., 2025). WRF-LES simulations and ground-based observations for this period are described in the following subsections.



3.2.3 WRF simulations

We ran WRF/WRF-LES simulations for the by one hour extended IOP 10 period described above, with grid spacings ranging from 1 km to 111 m, following a nested setup similar to Duine et al. (2019). WRF version 4.6.0 was used for the numerical simulations, with the European Centre for Medium-Range Weather Forecast Reanalysis' 5th Generation product (ERA5) as
 225 initial and boundary conditions (Hersbach et al., 2020), updated at the boundaries of the outermost domain every hour. We used three two-way nested domains to dynamically downscale from 9 to 1 km grid spacing. The mesoscale configuration domain follows previous numerical studies of Sundowner winds (Duine et al., 2019; Carvalho et al., 2020; Jones et al., 2021; Duine et al., 2021, 2022, 2025). The model was further downscaled from the 1 km domain to 111 m grid spacing offline using WRF's
 230 one-way nesting *ndown* procedure, generating initial and boundary conditions from the 1 km output for the subkilometer domain, subsequently run in LES mode. More configuration details are described in Table S1, and other parameterizations used can be found in Duine et al. (2019) and Jones et al. (2021).

All near-surface wind output (u , v and w) is at 10 m above ground, at 10-minute intervals, yielding 150 time steps for the period. For evaluating downscaled 100 m winds, we resample the 111 m WRF-LES output to 100 m grid spacing, referred to as WRF-LES_{111m} in the following. We excluded 106 rare WRF-LES_{111m} wind outliers exceeding the 99.999th percentile of
 235 the v - and w -components, corresponding to approximately 0.001 % of valid land grid cell values.

Using enhanced DEVINE, we statistically downscale low-resolution 1 km WRF near-surface wind over the Southern California model domain to a 100 m grid spacing. Inputs to this downscaling are the 1 km WRF wind data resampled to 100 m (hereafter WRF_{1km}), and the USGS 3DEP 1/3 arc-second (U.S. Geological Survey, 2021) DEM, both resampled to 100 m. We then evaluate the downscaled near-surface wind fields against the corresponding high-resolution WRF-LES_{111m} wind.

240 3.2.4 Ground observations

In addition to evaluating downscaled wind in the coastal environment against high-resolution WRF data, we perform point evaluations at 56 ground stations available within the domain (Fig. 1C), operated by three station networks. During the SWEX campaign, the National Center for Atmospheric Research/Earth Observing Laboratory (NCAR/EOL) deployed 18 Integrated Surface Flux System (ISFS) towers, most located in mountainous terrain, providing 10 m wind observations averaged over a
 245 5-minute window and reported every ten minutes. Six ISFS towers were located in the study domain. The U.S. Forest Service operates the permanent Remote Automatic Weather Station (RAWS) network (Zachariassen et al., 2003), of which five stations, mostly located in the foothills of the Santa Ynez Mountains, fall within our study domain, providing 6.1 m wind observations as 10-minute averages reported hourly. The Southern California Edison (SCE) energy company operates a permanent mesonet, of which 45 stations in Santa Barbara County were used, measuring wind at various heights above ground every 10 minutes. A
 250 large part of SCE stations are located in urban areas of Santa Barbara and are frequently sited along roadways. Model output was extracted from the nearest grid cell to each observation station and compared with the corresponding observed wind speed and direction.



4 Methods

4.1 Enhancing DEVINE

255 Le Toumelin et al. (2023) developed DEVINE, a fully convolutional U-Net architecture (Ronneberger et al., 2015), designed to downscale near-surface wind components (u , v , and w) over complex topography. The model relies on low-resolution forcing wind fields and high-resolution DEM data to produce wind predictions consistent with fine-scale topography.

The following methods subsections describe the modifications introduced to broaden DEVINE's applicability and the steps required to generate high-resolution wind field predictions in mountainous terrain. In the following, we refer to the model of

260 Le Toumelin et al. (2023) as the "original" DEVINE model.

4.1.1 Additional input features for DEVINE

The ARPS wind simulations used to train the original DEVINE model assume a single westerly inflow direction. From a physical perspective, this assumption is not restrictive, as the GRF topographies span a wide range of terrain characteristics, making a single inflow direction sufficient for generating diverse flow realizations. However, deep-learning models trained on
 265 gridded, image-like inputs learn patterns from the data representation itself. In the DEVINE framework, which uses gridded topography and flow fields as inputs and outputs, this implies that applying the model to wind directions other than west required rotating the topography to align with the western inflow during pre-processing and rotating the predictions back afterward. Without enlarging the computational domain, this led to data loss in the corner regions (Helbig et al., 2024, see Fig. 11c). To remove these pre-processing steps, we introduce wind-direction-dependent topographic input features. This enables rotational
 270 data augmentation during training and allows the model to predict flow fields for multiple wind directions using low-resolution wind direction inputs, without requiring topography rotation or domain enlargement.

We selected two wind-direction-dependent features that have previously been used successfully to describe local wind-flow modifications induced by topography in diagnostic and statistical downscaling approaches. The first feature represents the local (i) wind-flow deflection, $\theta_{d,i}$, from the low-resolution wind direction wd (Ryan, 1977; Liston and Elder, 2006), and is adapted
 275 here as

$$\theta_{d,i} = -0.5 \sigma_{S,i} \operatorname{sgn}(\sigma_{S,i}) \sin(2(\psi_i - wd)), \quad (1)$$

where $\operatorname{sgn}(\cdot)$ denotes the sign function and ψ_i is the local terrain aspect angle. Equation (1) uses the terrain slope in the wind direction, $\sigma_{S,i}$ (Liston and Elder, 2006), defined as $\sigma_{S,i} = \zeta_i \cos(wd - \psi_i)$ where the local terrain slope angle ζ_i is computed from the partial derivatives of the elevation z in orthogonal horizontal directions as $\zeta_i = \tan^{-1}(\sqrt{(\partial_x z_i)^2 + (\partial_y z_i)^2})$. The
 280 second wind-direction-dependent feature represents a local downscaling factor for the vertical wind speed, $Y_{dsc,i}$, introduced by Helbig et al. (2024). It represents enhanced upward motion on windward slopes and downward motion on lee slopes:

$$Y_{dsc,i}(\Delta\psi_i, \mu_i) = [(a' - b'(\Delta\psi_i) + c' \operatorname{erf}(d'(\Delta\psi_i)))] (e' + \mu_i^{f'}), \quad (2)$$



with constant parameters $a' = -0.087122$, $b' = 0.4788$, $c' = 2.068$, $d' = 0.6298$, $e' = -0.046577$ and $f' = 0.72451$. $Y_{\text{dsc},i}(\Delta\psi_i, \mu_i)$ is a function of a local slope parameter $\mu_i = \sqrt{((\partial_x z_i)^2 + (\partial_y z_i)^2)/2}$ and a local relative terrain aspect parameter $\Delta\psi_i$,
285 defined as the local terrain aspect angle ψ_i relative to the low-resolution wind direction wd , with values between -90° to 0° for
rather wind-sheltered and from $>0^\circ$ to 90° for more wind-exposed terrain (further details in Helbig et al., 2024). In addition to
these two wind direction-dependent features, we introduce a third, direction-independent feature used for retraining the original
DEVINE. The third feature extends the local elevation z_i , which was the only topographic input feature in the original DEVINE
model, to a local relative elevation $z_{r,i}$. This feature normalizes local elevation by subtracting the domain mean elevation $\bar{z}$:
290 $z_{r,i} = z_i - \bar{z}$. We omit the subscript (i) for local features throughout the remainder of the manuscript.

4.1.2 Rotational data augmentation

Rotational augmentation, facilitated by the use of the two wind-direction-dependent features, was primarily applied to enable
DEVINE to operate without requiring an explicit terrain-alignment preprocessing step. This also exposes the DL model to a
wider range of orientations for a given topography during training, which can support robustness. Note, however, that rotated
295 versions of the same topography are not independent samples of the underlying terrain distribution, they represent the same
terrain viewed at different angles. This is accounted for in the stratified training/validation split by keeping all rotational
augmentations of a given topography within the training subset.

Horizontal rotations were applied to the input features as well as wind vectors, and all rotated fields were cropped to a central
64×64 region to ensure consistent input dimensions (for more technical details and an example see Supplementary material).
300 Rotational augmentation was applied at 5° increments over the full 360° range, expanding the total of 7859 GRF-ARPS grids
by a factor of 72 yielding 565 848 spatial wind fields; a 10° subset, every second rotation, was used for a separate training,
yielding 282 924 fields.

4.1.3 Adaptations to the original DEVINE implementation and training

Several adaptations were made to the original DEVINE model of Le Toumelin et al. (2023) and are detailed in the Supple-
305 mentary material. While the network architecture and most of the hyperparameters remained unchanged (see Table S2 for the
hyperparameters that differ from those used in the original DEVINE), DEVINE was reimplemented in PyTorch and enhanced
with automatic padding of stacking layers, replacing the manual padding and cropping required in the original implementa-
tion. DEVINE was also retrained on square 64×64 pixel input maps, which ensure consistent alignment between encoder and
decoder feature maps and avoid the separate padding and cropping issues caused by the original 79×69 input size. Further
310 changes to normalization and training convergence criteria were also made relative to the original version.

Beyond these technical adaptations, we changed the training and validation procedure, and trained several DEVINE variants.
The GRF-ARPS dataset was split into into 80 % training and 20 % validation data, stratified across combinations of ξ and
domain-mean slope of the GRF topographies (rotational augmentation is applied only to the training subset, see Section 4.1.2).
This differs from the original DEVINE model, which used random 90 %-10 % splits without stratified topographic sampling,
315 but employed 10-fold cross-validation. Owing to the substantially larger training dataset resulting from data augmentation



Table 2. New and original DEVINE variant specifics.

Model name	Loss function	Activation function	Data augmentation	Features
DEVINE PyTorch*	MSE	ReLU	None	DEM
DEVINE retrained	MSE	ReLU	None	z_r
DEVINE low-bias	MSE	ReLU	10°	θ_d, z_r, Y_{dsc}
DEVINE low-scatter	MSE	ReLU	5°	θ_d, z_r, Y_{dsc}

*DEVINE PyTorch is a PyTorch reimplementation of the original DEVINE model using the original trained weights and training split.

(more than 225 000 resp. 450 000 wind fields), 10-fold cross-validation was not applied here. Building on these changes, we considered four DEVINE model variants in total (see Table 2 for variant specifics): two new models, DEVINE low-bias and DEVINE low-scatter, trained on the augmented dataset; DEVINE retrained, a PyTorch reimplementation of the original DEVINE architecture, retrained on the unrotated dataset using relative elevation z_r only as input feature; and DEVINE PyTorch, a PyTorch reimplementation of the original DEVINE model using the original trained weights and training split, without retraining.

Training using the 6285 unrotated 64x64 grids took around 40 minutes, training on the rotational augmented datasets (5° and 10° increments) took around 60 and 30 hours, respectively. Predictions for 7859 64×64 grids take about 27 seconds, and for 150 300×300 grids about 10 seconds. All training and prediction were run on the institutional HPC's shared login node, using an NVIDIA Tesla V100 (16 GB) GPU and 4 cores of an Intel Xeon E5-2650 v4 CPU. Reported computational times are indicative and may vary with cluster utilization.

4.1.4 DEVINE predictions: Pre- and postprocessing

The steps required to deploy DEVINE were considerably reduced compared to the original DEVINE model. Pre-processing consists of computing the three features (Section 4.1.1) using high-resolution terrain data and low-resolution wind direction input. Post-processing involves linearly scaling the DEVINE wind components by the low-resolution wind speed and dividing by 3 m s^{-1} , which corresponds to the initial wind speed of the ARPS wind dataset. This post-processing step is unchanged from the original DEVINE model. Low-resolution wind input is interpolated to the fine-scale grid to prevent chessboard-like artifacts in the predicted wind fields. While different interpolation methods can be used, we apply bilinear interpolation for the predictions presented in this study.

4.2 Performance metrics

We evaluate model performance using bias, mean absolute error (MAE), the standard deviation of the signed error (σ_{error}), the Pearson correlation coefficient (r), and the normalized root mean square error (NRMSE). Bias is defined as the mean signed model error (model minus reference), MAE as the mean absolute model error, and σ_{error} as the standard deviation of the signed



errors, describing their dispersion around the mean error. NRMSE is computed as the root mean square error normalized by the
 340 full range of the reference data, expressed as a percentage. In addition we report normalized wind speed and normalized bias,
 computed by dividing by the corresponding low-resolution (input) wind speed, in order to express local errors relative to the
 driving forcing conditions. Wind direction errors are computed using circular statistics. To avoid instability in wind direction
 and normalized error metrics under near-calm conditions, we exclude atmospheric model reference or observed wind speed
 $\leq 0.5 \text{ m s}^{-1}$ from all error calculations. This threshold excludes a negligible fraction of samples in the synthetic environment
 345 (0.04 %), 0.9 % in the alpine environment, and 4.8 % in the coastal environment (6.6 % when also applied to the low-resolution
 input wind). For the station evaluation, the total excluded fraction was 9.3 % (9.9 % when also applied to the low-resolution
 input wind).

To assess whether DEVINE reproduces terrain-induced spatial variability in the wind field, we compute the horizontal
 gradient magnitude of horizontal wind speed wsp from the time-averaged wind fields as $|\nabla wsp| = \sqrt{\left(\frac{\partial wsp}{\partial x}\right)^2 + \left(\frac{\partial wsp}{\partial y}\right)^2}$
 350 for ARPS, WRF and DEVINE. Gradient-magnitude difference maps are obtained by subtracting the corresponding reference
 model (ARPS or WRF) fields from DEVINE. In addition, we compute the horizontal gradient magnitudes of the wind compo-
 nents, $|\nabla u|$, $|\nabla v|$, and combine their differences into u - v gradient magnitude difference maps for ARPS, WRF and DEVINE
 wind,

$\Delta_{uv} = \sqrt{(|\nabla u|_{\text{DEVINE}} - |\nabla u|_{\text{model}})^2 + (|\nabla v|_{\text{DEVINE}} - |\nabla v|_{\text{model}})^2}$. While the first metric quantifies how well DEVINE
 355 reproduces the spatial variability in wind speed magnitude, gradients of the individual u - and v -components additionally quan-
 tify how well DEVINE captures the spatial variability arising from both speed variations and flow turning.

5 Results

We evaluate four DEVINE model variants: the PyTorch reimplementation (DEVINE PyTorch) and retrained version (DEVINE
 retrained) of the original model, and two newly trained variants using our rotation-augmented dataset (DEVINE low-bias,
 360 DEVINE low-scatter). We first assess all four variants in the synthetic environment against GRF-ARPS simulations, using
 DEVINE PyTorch and DEVINE retrained as benchmarks for the original model's performance. Based on this evaluation, we
 select DEVINE low-bias as the default model, and use it to demonstrate strengths and limitations in two contrasting real-
 world environments, alpine and coastal mountains, by comparing downscaled wind against atmospheric model wind in both
 environments, and additionally against ground observations in the coastal environment. A cross-environment comparison with
 365 atmospheric model wind (Section 5.4) brings the results together.

5.1 Evaluation in synthetic environment

For this evaluation on the unused GRF-ARPS validation dataset (using splits that differ from those of Le Toumelin et al.
 (2023)), we did not expect gains over DEVINE PyTorch and DEVINE retrained, since our preprocessing changes aimed to
 simplify, not improve on, the original approach.



Table 3. Performance metrics of the four DEVINE model variants for the synthetic environment relative to high-resolution atmospheric model wind components (u , v , w), horizontal wind speed (wsp), and wind direction (wd). Bias, MAE and σ_{error} are reported in the unit of each variable.

Variable	DEVINE PyTorch*					DEVINE retrained				
	Bias	MAE	σ_{error}	NRMSE [%]	r [-]	Bias	MAE	σ_{error}	NRMSE [%]	r [-]
u	0.02	0.21	0.32	3.1	0.92	-0.02	0.17	0.25	2.4	0.95
v	0.03	0.15	0.23	3.0	0.90	-0.03	0.15	0.21	2.8	0.92
w	0.01	0.10	0.16	1.7	0.98	0.01	0.07	0.12	1.3	0.99
wsp	0.01	0.21	0.33	3.3	0.91	-0.02	0.18	0.26	2.6	0.95
wd	-	3.2	4.9	2.8	-	-	3.0	4.6	2.6	-
Variable	DEVINE low-bias					DEVINE low-scatter				
	Bias	MAE	σ_{error}	NRMSE [%]	r [-]	Bias	MAE	σ_{error}	NRMSE [%]	r [-]
u	< 0.01	0.14	0.20	2.0	0.97	0.01	0.14	0.20	1.9	0.97
v	0.01	0.13	0.19	2.5	0.94	-0.03	0.13	0.18	2.5	0.94
w	< 0.01	0.06	0.10	1.1	0.99	< 0.01	0.06	0.10	1.1	0.99
wsp	< 0.01	0.15	0.22	2.1	0.97	< 0.01	0.15	0.21	2.1	0.97
wd	-	2.7	4.0	2.2	-	-	2.7	4.0	2.2	-

Integrated bulk error metrics (MAE, bias, correlation) for DEVINE retrained are slightly worse across all metrics compared to those reported by Le Toumelin et al. (2023), though still within the same order of magnitude. DEVINE PyTorch shows the weakest performance of all four variants. Performance from the two new DEVINE variants is of almost identical magnitude to that of the original DEVINE model (as in Le Toumelin et al., 2023), albeit derived here for a different GRF-ARPS validation split (Table 3). Overall, all performance metrics improve progressively from DEVINE PyTorch, through DEVINE retrained, to the two new DEVINE variants. Between the new variants, performance differences are marginal. Nevertheless, we differentiate between those two models due to their slight differences in bias and scatter, motivating our model naming convention. Table 3 shows that DEVINE low-bias achieves smaller or comparable bias magnitudes than DEVINE low-scatter across all wind components, wind speed wsp , and wind direction wd , most notably for v (0.01 m s^{-1} vs. 0.03 m s^{-1}). Conversely, DEVINE low-scatter shows lower or comparable bias standard deviation across all wind components and wsp and wd compared to DEVINE low-bias, consistent with its marginally lower NRMSE for these variables despite larger bias.

Consistent with the overall bulk error metrics, probability distributions of all DEVINE variants show almost identical agreement with ARPS simulations for all wind components within the $[q_5, q_{95}]$ range. Beyond this range, DEVINE low-bias agrees best with ARPS, except for $v \leq q_5$, whereas DEVINE PyTorch most often shows the poorest agreement (Fig. 2, upper row).

MAEs of the original DEVINE models are largest on steep, sheltered slopes, especially for u and w , and additionally for v on steep windward slopes (Fig. S2). These MAEs reduce considerably from DEVINE PyTorch to the two new DEVINE variants, by about 0.1 m s^{-1} , but largest MAE remain on steep, sheltered slopes. Bias spread for DEVINE low-bias and DEVINE low-scatter increases with increasing slope steepness and for convex terrain reaching 5th-95th percentile values up to $\pm 0.43 \text{ m s}^{-1}$.



DEVINE low-bias additionally slightly underestimates modeled wind in concave terrain (Figs. S3, S4). *wd* bias for both variants has likewise a larger spread on steeper slopes, though the spread is more balanced across terrain curvature, reaching 5th-95th percentile values up to $\pm 8^\circ$. Median *wdbiases* show slight, systematic clockwise shifts relative to ARPS for DEVINE low-scatter across all terrain characteristics, whereas median *wd* biases for DEVINE low-bias are close to zero for *wd* overall, aside from marginal counterclockwise shifts on the steepest slopes. As with the terrain-conditioned patterns found later in the alpine and coastal evaluations, MAE and bias diverge in where they are largest: MAE peaks on steep, sheltered slopes, where terrain-flow interactions are most variable, while bias spread instead grows mainly with slope steepness and convexity, without the same sheltered-terrain concentration, a divergence that recurs, more pronounced, in the alpine and coastal environments.

In summary, both new DEVINE variants reproduce wind components without pronounced systematic bias across terrain types and additionally show slightly improved performance compared to the DEVINE PyTorch and DEVINE retrained benchmarks. Because even small systematic biases persist and can accumulate over repeated or long-term predictions, we chose DEVINE low-bias as the default DEVINE model. DEVINE low-scatter, in turn, may be preferable where a narrow prediction spread matters more than eliminating a small systematic offset, for instance, in short-term, single-event applications. In the following sections, DEVINE refers to DEVINE low-bias unless otherwise stated. Both variants are available on GitHub (see Code and Data Availability).

5.2 Evaluation in alpine environment

Moving from the synthetic to the alpine environment tests DEVINE's generalization to real, independent topography. Although GRF topographies show comparable slope characteristics to real terrain (Helbig and Löwe, 2012), we expect a modest performance decrease given this genuine out-of-sample test.

Bulk error metrics between DEVINE and ARPS in the alpine environment show that model performance decreases relative to the synthetic environment (Table 4), though this decrease remains modest: MAE and NRMSE stay very low in absolute terms across all variables, while correlation coefficients decrease for *u*, *w*, and *wsp*, but remain high overall ($r \geq 0.84$); only *v*'s correlation is essentially unchanged, showing a marginal improvement. The most notable change is in bias, which grows more than MAE or NRMSE, most clearly for *u* and *v*, though overall MAE and NRMSE remain small ($\leq 0.23 \text{ m s}^{-1}$, $\leq 7^\circ$, $\leq 5\%$). Since bias was close to zero for all variables in the synthetic environment, this suggests that alpine errors carry a modest systematic bias component that was largely absent from the synthetic evaluation, while the increase in random scatter σ_{bias} remains comparatively smaller. Notably, *v* deviates from this pattern: despite having the largest bias increase among wind components, its NRMSE and correlation are essentially unchanged (or marginally improved) relative to the synthetic environment.

Probability distributions of DEVINE wind components agree well overall with those from ARPS within the $[q_5, q_{95}]$ range, though agreement is weakest for the *u* component (Fig. 2, middle row), consistent with *u* also showing comparatively weaker performance among wind components (Table 4). Beyond this range, distributions deviate more clearly from ARPS than in the synthetic environment (Fig. 2, upper row), though the direction of this deviation follows a similar pattern across components: an overestimation in the lower tail (below q_5) and an underestimation in the upper tail (above q_{95}).

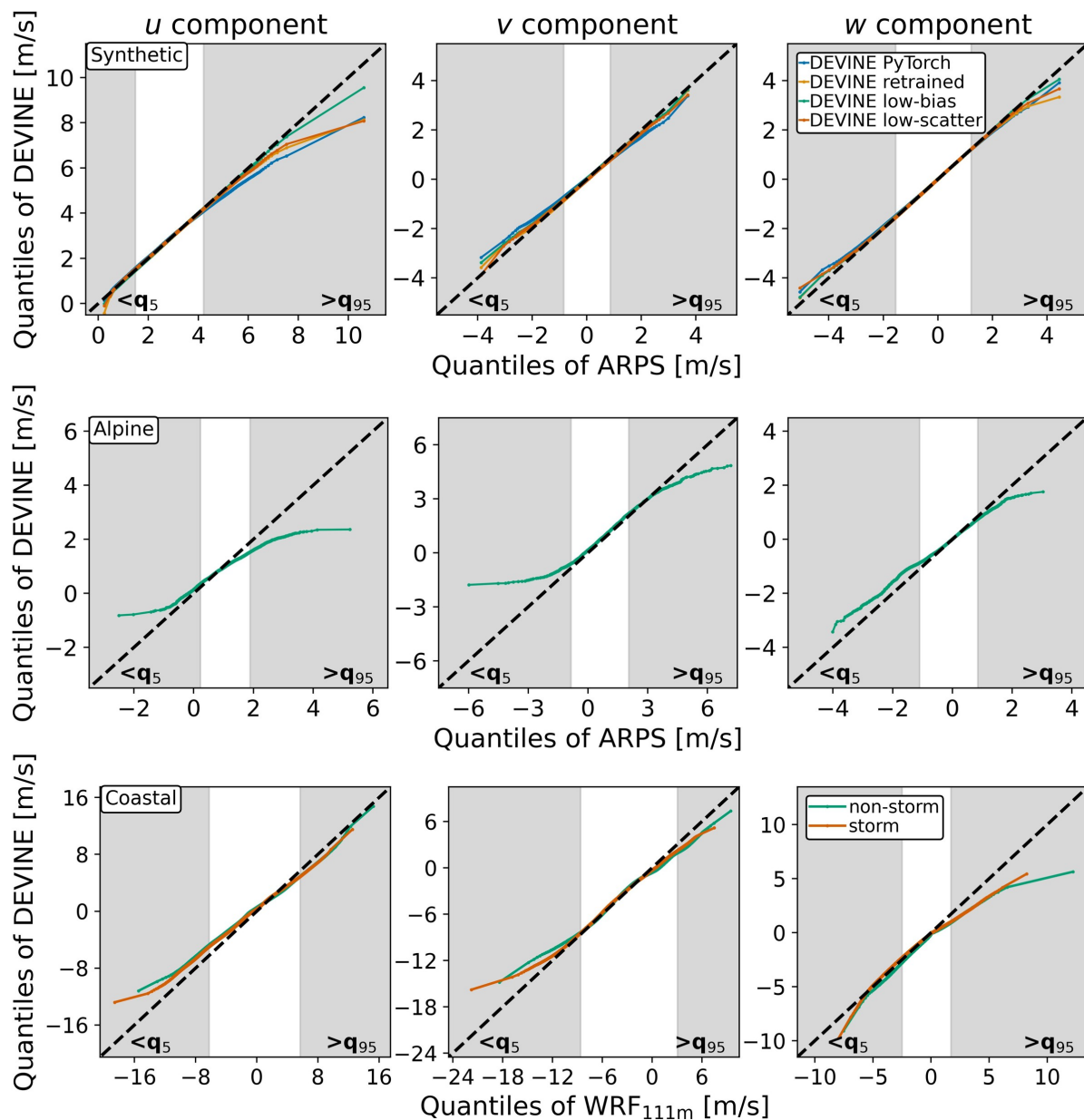


Figure 2. Quantile-quantile plots comparing DEVINE to the reference wind field (ARPS or WRF-LES_{111m}, depending on environment) for each wind component (u , v , w), using 10 000 uniformly spaced percentile levels (0th-100th). Row 1 compares all four DEVINE variants (Table 2) in the synthetic environment; rows 2-3 show the selected DEVINE low-bias model in the alpine and coastal environments, respectively, with coastal distributions shown separately for non-storm and storm periods. Shaded regions indicate the q_5 - q_{95} range of the respective reference dataset.



Table 4. Performance metrics for the alpine and coastal environments relative to high-resolution atmospheric model wind components (u , v , w), horizontal wind speed (wsp), and wind direction (wd), with bias, MAE and σ_{error} reported in the unit of each variable. For coastal, upper/lower cell values correspond to non-storm and storm periods, respectively (see Section 5.3).

Variable	Alpine environment					Coastal environment				
	Bias	MAE	σ_{error}	NRMSE [%]	r [-]	Bias	MAE	σ_{error}	NRMSE [%]	r [-]
u	-0.08	0.21	0.29	3.9	0.85	0.06	1.18	1.61	5.3	0.86
						0.53	1.50	1.93	6.5	0.80
v	0.15	0.20	0.25	2.2	0.96	-0.55	1.36	1.76	6.6	0.77
						-0.04	1.59	2.24	7.7	0.80
w	0.04	0.15	0.23	3.4	0.94	-0.32	0.55	0.76	4.2	0.67
						-0.07	0.60	0.89	5.5	0.76
wsp	-0.06	0.23	0.33	4.8	0.84	-0.70	1.18	1.46	8.4	0.75
						-0.44	1.60	2.15	9.6	0.77
wd	-	7	-	5	-	-	30	-	24	-
						-	34	-	29	-

DEVINE's spatial wind fields agree well with ARPS in the alpine environment (Fig. S5), as in the synthetic environment. DEVINE's largest MAE, however, occur on the steepest slopes ($>54^\circ$) at strongly negative relative terrain aspect angles (approximately -51° to -80° , i.e. wind-sheltered terrain), with maximum MAE values of 0.75 m s^{-1} for u , 0.55 m s^{-1} for v , and 0.62 m s^{-1} for w (Fig. 3). Unlike in the synthetic environment, wsp shows a slight low bias that grows toward gentler terrain (Fig. S6). A weaker version of this same low-bias pattern also appears in concave terrain, consistent with the synthetic environment (Figs. S3 vs. S6). Wind direction errors show a clear counterclockwise shift across all slope and aspect classes, which increases toward exposed terrain with maximum median biases of about -5° , alongside increased scatter on steeper slopes. This contrasts with the synthetic environment, where wind direction bias was minimal (counterclockwise for DEVINE low-bias and clockwise for DEVINE low-scatter) and remained balanced across terrain exposure.

Overall, DEVINE reproduces terrain-induced spatial wind speed gradients (speed-up, sheltering) reasonably well, with a median and mean relative difference of -23% and -9% , respectively (Fig. 4a-c). This difference shows a U-shaped dependence on slope and curvature, quantified via the Laplacian following Helbig et al. (2017), where positive values indicate concave and negative values convex terrain, most negative at the extremes (shallowest/steepest slopes; most concave/convex terrain) and least negative near the mid-range. The median underestimation also grows more consistently with terrain exposure (Fig. 4b), from -12% (most sheltered) to -27% (most exposed).

In contrast, DEVINE shows substantially larger discrepancies in spatial gradients of the horizontal wind components (median and mean combined component gradient magnitude difference of 48% and 51% , respectively; Fig. S7), increasing modestly at the steepest slopes but otherwise largely terrain-independent. The larger discrepancies suggest fine-scale spatial variations of the vector wind field (e.g., channeling, flow deflection) are reproduced less reliably than spatial wind speed magnitude gradients (e.g., speed-up/sheltering) alone.

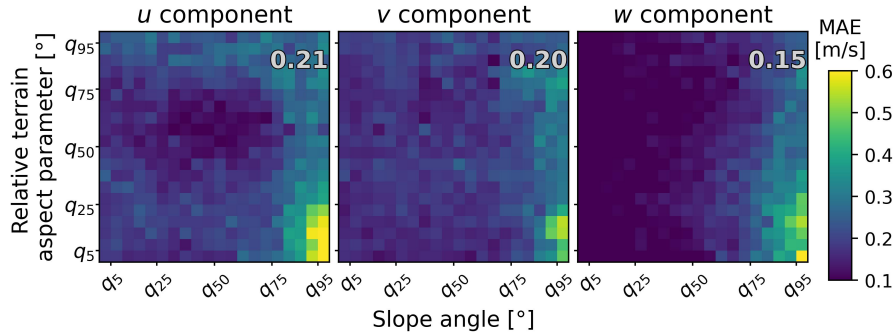


Figure 3. Mean absolute error (MAE) between ARPS and DEVINE as a function of terrain parameter quantiles (5 % bins) in the alpine environment. Shown quantiles (q_5 , q_{25} , q_{50} , q_{75} , q_{95}) correspond to terrain slope angles of 10° , 20° , 29° , 36° , and 54° , and relative terrain aspect parameters $\Delta\psi$ of -80° , -42° , -3° , 34° , 76° . Corners values show overall MAE per wind component (m s^{-1}).

Bringing alpine terrain-conditioned analyzes together, the largest MAEs occur over steep, sheltered terrain relative to the incoming flow (Fig. 3), whereas bias in w_{sp} (Fig. S6) and in w_{sp} gradient magnitude (Fig. 4a-c), instead is mostly largest on low-slope and exposed terrain. Since MAE reflects both systematic bias and error scatter, its peak occurring on terrain where bias does not similarly peak indicates that scatter, rather than systematic over- or underestimation, is the dominant contributor to the larger MAE on steep, sheltered slopes, consistent with greater uncertainty in complex, sheltered flow conditions, where terrain-flow interactions are more variable.

The evaluation in the alpine environment shows that DEVINE predicts wind speed and wind direction with spatially varying error patterns similar to those in the synthetic environment, though with larger bias magnitudes overall.

5.3 Evaluation in coastal mountain environment

In the coastal Santa Ynez Mountains of Southern California, we evaluate DEVINE's performance in a distinctly different mountain environment (Section 3.2.2), using WRF-LES simulations, rather than ARPS, the model used during training, together with independent station observations. The evaluation period includes a warm, gusty downslope wind event, producing mixed wind regimes that consist of complex atmospheric-terrain flow interactions. To isolate the influence of wind regime, we separate the downslope storm event into "non-storm" and "storm" periods.

5.3.1 Comparison with WRF wind

This downslope windstorm represents a wind regime DEVINE never encountered during training, extending the out-of-sample evaluation introduced in the alpine environment. A diurnal sea-breeze circulation dominates during the day, reversing at sunset into strong, gusty cross-barrier flow toward the ocean during the storm, and shifting back by morning to the typical sea-land breeze pattern (see Figs. S8, S9 for examples). Since the observed onset and end of the downslope event vary across the region (de Orla-Barile et al., 2025), we define a single onset and end time by visually inspecting modeled wind direction patterns from

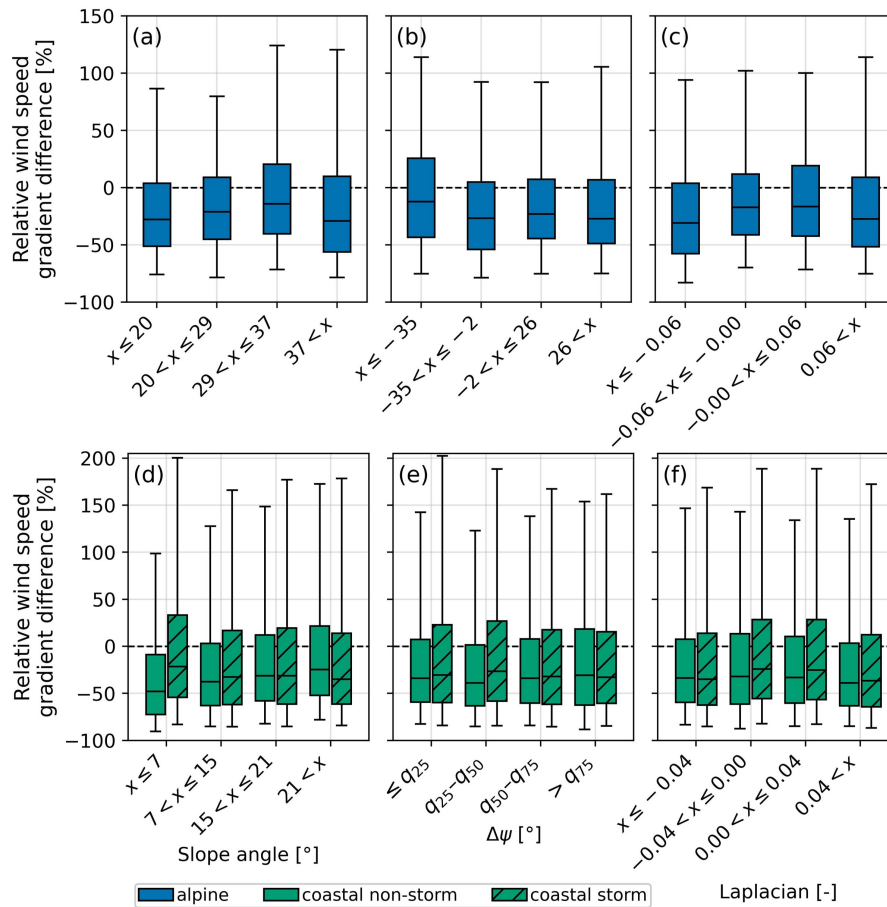


Figure 4. Relative wind speed gradient difference between ARPS and DEVINE in the alpine environment (a-c) and between WRF-LES_{111m} and DEVINE in the coastal environment (d-f), grouped by terrain parameter quantiles: (a, d) slope angle, (b, e) relative terrain aspect parameter $\Delta\psi$, and (c, f) Laplacian. Values are normalized by the respective ARPS and WRF-LES_{111m} gradient magnitudes. Quantiles thresholds correspond to q_{25} , q_{50} , and q_{75} ; shown $\Delta\psi$ quantiles are -13° , 13° , 36° for coastal non-storm and -41° , -13° , 10° for coastal storm. Boxplots show the median, interquartile range (box), and 5th/95th percentile whiskers; outliers are omitted.



WRF_{1 km}, informed by the detailed event descriptions and station observations in de Orla-Barile et al. (2025). Our extended IOP 10 period spans 12 May, 09:00 PDT (16:00 UTC), to 13 May, 09:50 PDT (16:50 UTC). Within this period, we define the storm start at 18:00 PDT (01:00 UTC) and end at 08:00 PDT (15:00 UTC) the following day. Of the 150 available 10-minute time steps, this yields 65 time steps outside the downslope event (before and after combined: non-storm) and 85 time steps during the event (storm).

Bulk error metrics between DEVINE and atmospheric model wind show a substantial decrease in performance for the coastal environment relative to the alpine environment (Table 4). Considering first the non-storm period, where flow is dominated by sea-breeze and terrain forcing, most comparable to the terrain-forced flow in the synthetic and alpine environments, MAE increases by roughly four- to sevenfold relative to alpine across wind components, wind speed, and wind direction (e.g. u : 0.21 to 1.18 m s⁻¹; wd : 7° to 30°), a substantially larger increase than the 1.5- to threefold increase from synthetic to alpine. As with the bias increase from synthetic to alpine, bias again grows clearly for wsp , w and v , while u 's bias magnitude decreases, contrasting with the alpine environment, where bias worsened for all wind variables and v showed the largest relative increase. Correlation decreases relative to alpine for v , w , and wsp (most for w : 0.94 to 0.67) but is essentially unchanged for u . Correlations overall are lower and more variable (0.67-0.86) for coastal than for alpine. NRMSE increases across all wind variables in the non-storm period, moderately for u and w (factor 1.3) and more markedly for wsp , v and wd (factor 1.75, 3.0, and 4.8, respectively; wd : 5% to 24%). Unlike when moving from the synthetic to the alpine environment, σ_{bias} also increases substantially here (roughly 3- to 7-fold), indicating that both systematic and random error components grow in the coastal environment for non-storm conditions. During the downslope-storm period, MAE and σ_{bias} increase further, consistent with greater unresolved flow variability and turbulence, though this further increase is modest compared to the increase from alpine to coastal non-storm. NRMSE also increases further, with only w exhibiting a larger increase compared to the increase from alpine to coastal non-storm, while the increases for u , v , wsp , and wd remain comparatively modest. In contrast, bias and correlation improve for all wind variables except u , which is at first counterintuitive, since one might expect storms to drive larger systematic deviations.

Consistent with the substantial decrease in bulk performance relative to alpine, DEVINE's distributions deviate more clearly from WRF-LES_{111 m} beyond $[q_5, q_{95}]$ for all wind components (Fig. 2, bottom row), while agreement within this range remains close for both periods. The quantile range itself also extends to markedly higher speed magnitudes than in the synthetic and alpine environments. Beyond $[q_5, q_{95}]$, u and v show the same asymmetric pattern as in the alpine environment (Fig. 2, middle row): overestimation $< q_5$, underestimation $> q_{95}$. This asymmetry is absent for w , which instead underestimates at both tails. Differences between non-storm and storm distributions are small within $[q_5, q_{95}]$, increasing toward both extremes, most visibly for v and w .

Terrain-conditioned MAE in the coastal environment shows a MAE pattern not present in the alpine or synthetic environments, with MAE increasing toward the lowest slopes ($< 10^\circ$), most dominantly in wind-sheltered terrain (Fig. 5), and far more pronounced during the storm than non-storm. A weaker, less consistent secondary increase also appears at the steepest slopes during the storm. The largest binned MAE occurs for v during non-storm (1.96 m s⁻¹) and for u during storm (2.03 m s⁻¹),

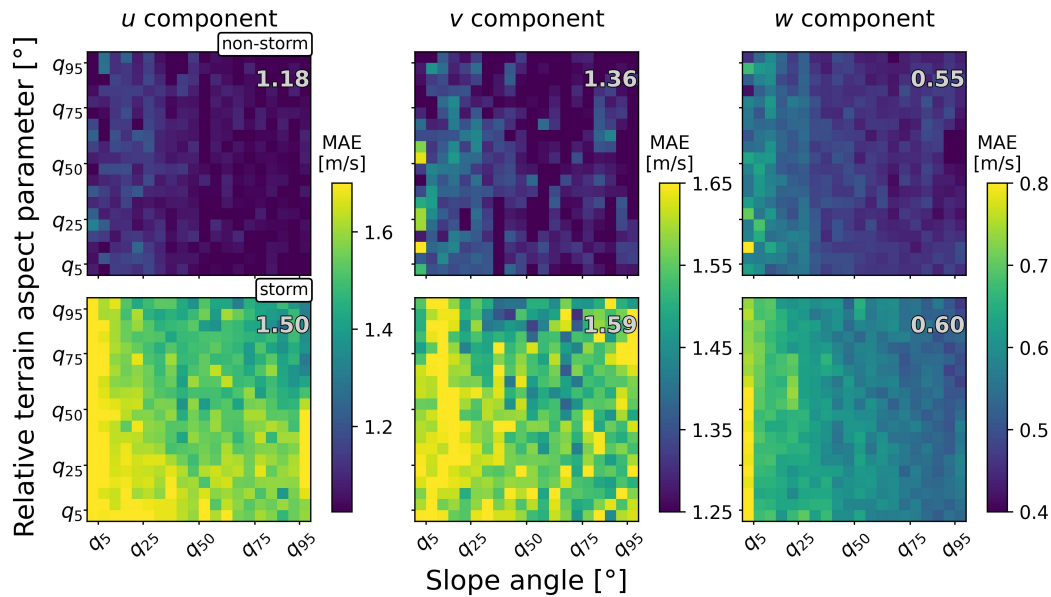


Figure 5. MAE between WRF-LES_{111m} and DEVINE, as a function of terrain-parameter quantiles (5 % bins) in the coastal environment: non-storm (top row) and storm (bottom row). Shown quantiles (q₅, q₂₅, q₅₀, q₇₅, q₉₅) correspond to slope angles of 1°, 7°, 15°, 21°, 30°, and relative terrain aspect parameters $\Delta\psi$ of -78°, -32°, 16°, 55°, 83° (non-storm) and -84°, -59°, -21°, 26°, 76° (storm). Corner values show overall MAE per period (m s^{-1}).

both at the lowest slope bins in sheltered terrain. This is the opposite spatial pattern to the alpine environment, where MAE instead peaks at the steepest, sheltered slopes, with the lowest slopes performing comparatively well (Fig. 3).

For both periods, DEVINE underestimates WRF-LES_{111m} wsp , similar to the low bias in the alpine environment, albeit that it is larger in the coastal environment (Fig. S10 vs. S6). Wind speed bias grows with slope, terrain exposure and convexity (Fig. S10A-C). During storm, wsp bias scatter increases markedly, with similar or slightly reduced median bias. Non-storm wind direction bias is consistently positive (clockwise) across all terrain classes, in contrast to the counterclockwise shift found in the alpine environment (Fig. S10D-F). During storm, wd bias shifts toward zero and becomes counterclockwise, with wider p5-p95 scatter, particularly at gentle slopes.

In the coastal environment during non-storm conditions, DEVINE underestimates terrain-induced spatial wsp gradient magnitude with a median and mean relative difference of -35 % and -11 %, respectively, on average only 2,% more than in the alpine environment (Fig. 4d-f). This underestimation grows toward gentler slopes, with a substantially wider p5-p95 range than in the alpine environment. During storm, the median and mean relative difference reduces to -31 % and -0.5 %, respectively, likely associated with the wider 25th-75th percentile range (IQR) compared to non-storm. Besides the non-storm slope dependence, relative wind speed differences are largely terrain-independent for both periods, unlike the alpine environment, which showed a U-shaped slope/curvature dependence and exposure trend.



Table 5. Performance metrics for the coastal environment relative to station horizontal wind speed (wsp) and wind direction (wd), reported for the non-storm and storm periods (see Section 5.3).

Variable	Metric	WRF _{1 km}		WRF-LES _{111 m}		DEVINE	
		non-storm	storm	non-storm	storm	non-storm	storm
wsp	Bias [m s^{-1}]	0.75	4.98	0.81	2.94	0.49	3.91
	MAE [m s^{-1}]	1.34	6.23	1.23	3.85	1.13	5.20
	σ_{error} [m s^{-1}]	1.90	5.15	1.43	3.73	1.60	4.69
	NRMSE [%]	16	26	13	17	13	22
	r [-]	0.37	-0.14	0.58	0.27	0.50	-0.12
wd	MAE [$^{\circ}$]	48.6	70.0	42.8	65.1	51.6	71.8
	NRMSE [%]	36	49	32	47	37	50

Spatial gradients of the horizontal wind components show larger, largely terrain-independent discrepancies for both periods (median and mean combined component gradient magnitude difference of 55 % and 59 % non-storm, 56 % and 61 % storm, respectively; Fig. S11), about 10 % higher than in the alpine environment. This modest increase indicates DEVINE reproduces coastal terrain-forced spatial wind speed structure (speed-up, sheltering) reasonably well even with different atmospheric driving model data, while fine-scale vector wind variations remain less reliably captured, consistent with the alpine result.

Bringing the coastal terrain-conditioned results together, the largest MAE occurs at low-slope, and predominantly wind-sheltered terrain (negative relative terrain aspect parameter; Fig. 5), whereas wsp bias instead grows with increasing slope and terrain exposure (Fig. S10), the opposite terrain association found for MAE. As in the alpine environment, this divergence suggests MAE at low-slope, sheltered terrain is driven primarily by error scatter rather than systematic bias, though here concentrated in flat rather than steep terrain, consistent with storm-driven and sea-breeze flow variability dominating this region rather than terrain-forced flow complexity alone. The evaluation in the coastal environment shows that DEVINE's terrain-forced flow transfers well under a different atmospheric driving model and mixed flow regimes, though with larger, more scattered bias magnitudes overall than in the alpine environment, and with wind direction bias reversing sign between the non-storm and storm periods.

5.3.2 Comparison with ground observations

DEVINE bulk errors against station observations differ notably from the spatial evaluation against WRF-LES_{111 m} (Table 5 vs. 4). DEVINE's non-storm station wsp bias and MAE are slightly lower than in the spatial comparison, and σ_{error} only slightly larger, but station NRMSE is substantially higher for both variables: wsp and wd NRMSE are roughly 55 % higher during non-storm, and roughly 130 % (wsp) and 72 % (wd) higher during storm. This widening storm-period gap points to station-specific error sources (e.g., point-scale representativeness, station siting) compounding with the coastal terrain and regime effects already identified. Notably, the sign of station bias also differs from the spatial comparison: all three models



overestimate station wsp in both periods (Table 5), whereas the spatial comparison showed underestimation for both, coastal and alpine wsp , (Table 4), indicating the difference between point and grid-cell representativeness discussed further below.

Among the three models non-storm station errors, DEVINE has the lowest wsp MAE and bias, and matches WRF-LES_{111 m}'s lower NRMSE relative to the low-resolution driving model WRF_{1 km}. For r and σ_{error} , WRF-LES_{111 m} performs best, while DEVINE still improves on WRF_{1 km}. For wind direction, DEVINE's non-storm and storm station MAE and NRMSE are both larger than in the spatial WRF-LES_{111 m} comparison, and DEVINE does not improve on WRF_{1 km}'s station errors in either period, unlike WRF-LES_{111 m}, which shows consistent modest improvements. DEVINE's relative degradation from non-storm to storm (wsp MAE: 4.6-fold) closely tracks WRF_{1 km}'s own degradation (4.7-fold) rather than the smaller degradation seen in WRF-LES_{111 m} (3.1-fold). Correspondingly, DEVINE performs worse than WRF-LES_{111 m} on every metric during the storm, and on r and σ_{error} during non-storm as well. Together, these results suggest that DEVINE's increased error during storm periods is largely related to errors already present in its WRF_{1 km} driving data rather than being introduced by the downscaling step itself. Because DEVINE is trained to downscale terrain-forced flow in neutral conditions, it cannot recover local-scale storm characteristics or thermally- or stability-induced features that are not represented in its low-resolution input. In contrast, WRF-LES_{111 m} resolves these storm-related flow characteristics directly. We examine this distinction in more detail in the following terrain-conditioned station analysis.

During storm, terrain-conditioned wsp and wd MAE against station observations shows a broadly similar pattern to the spatial evaluation against WRF-LES_{111 m}: WRF_{1 km}, WRF-LES_{111 m} and DEVINE MAE increase toward gentler, mostly wind-sheltered slopes (Fig. 5 vs. 6). Since WRF-LES_{111 m} and DEVINE share the same driving model (WRF_{1 km}), for which this pattern is most pronounced, this supports the earlier finding that DEVINE's storm-period degradation is inherited in part from its driving data rather than added independently by the downscaling step. At low-slope, sheltered terrain DEVINE's storm wsp and wd MAE is larger than WRF-LES_{111 m}'s, consistent with DEVINE underperforming WRF-LES_{111 m} on every storm-period bulk metric (Table 4). During non-storm conditions, both models show comparatively low and only weakly terrain-dependent MAE, although wsp MAE increases slightly for steeper, less sheltered station slopes, resembling the terrain-dependent behavior found for the alpine and synthetic cases.

For storm conditions compared to non-storm, all three models WRF_{1 km}, WRF-LES_{111 m} and DEVINE show wider p5-p95 scatter in wind speed bias and larger storm median wsp bias across terrain (Fig. S12). Storm median wsp bias decreases with increasing slope and toward less sheltered terrain, consistent with terrain-dependent MAE patterns. As noted above, all models overestimate station wsp in both periods (Table 5), opposite in sign to the spatial comparison. Largest wsp storm bias is from WRF_{1 km}, the lowest from WRF-LES_{111 m}. Relative to WRF_{1 km}, WRF-LES_{111 m} reduces storm bias and MAE more than DEVINE does (Table 5), consistent with WRF-LES_{111 m}'s tighter p5-p95 storm scatter throughout Fig. 5. During non-storm conditions, wsp bias is largely terrain-independent, though increasing toward steeper slopes and more sheltered terrain. Except at the most sheltered terrain, DEVINE shows the lowest median non-storm wsp bias across terrain and compared to the WRF models, agreeing with the lowest bulk non-storm station bias (Table 5). Wind direction bias shows a mostly clockwise shift during non-storm across all terrain bins and models, while storm median biases are more balanced in sign and magnitude across terrain, with markedly wider p5-p95 scatter. WRF-LES_{111 m} shows the lowest wd bias magnitude overall.

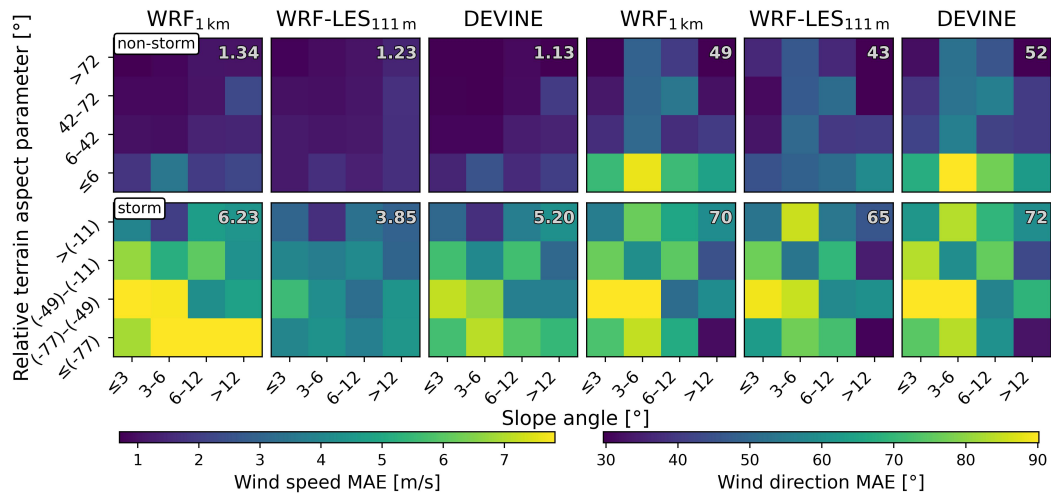


Figure 6. MAE between observed wind and WRF_{1km}, WRF-LES_{111m} and DEVINE, as a function of terrain-parameter quantiles: non-storm (top row) and storm (bottom row). Quantile thresholds correspond to q_{25} , q_{50} , and q_{75} . Corner values show overall MAE per period (m s^{-1}).

Across the normalized station wind speed range, DEVINE and WRF-LES_{111m} show broadly similar bias patterns against station wsp (Fig. 7), both overestimating at the lowest wsp quantiles in both periods, with bias increasing toward the lowest quantile. DEVINE shows a clear reversal to underestimation at the highest quantile, WRF-LES_{111m}'s median bias there remains near zero in both periods. WRF-LES_{111m}'s median bias is lower in magnitude than DEVINE's during storm across most of the range, while during non-storm DEVINE's median bias magnitude is lower than WRF-LES_{111m}'s, except at the largest wsp , where WRF-LES_{111m} is markedly less biased. Bias scatter of WRF-LES_{111m} is clearly larger than or similar (for largest wsp) to DEVINE's across the station wind range in both periods. Wind direction bias patterns are also similar between DEVINE and WRF-LES_{111m}: non-storm median biases are often lower for WRF-LES_{111m}, while storm medians are broadly similar between the two. The most striking result, however, is that storm-period wd bias scatter is extremely large, increasing toward the lowest station wind speed and reaching IQR exceeding 200° . Since station wsp is normalized by the low-resolution driving WRF_{1km} flow, quantiles ≤ 1 correspond to stations relatively sheltered from low-resolution wind, while quantiles > 1 correspond to stations relatively exposed.

To understand model performance at stations in more detail, we examine period-mean spatial wind speed around and at MTIC1, a station in the foothills of the Santa Ynez Mountains (Fig. 8). MTIC1 sits on a small, exposed ridge (491 m a.s.l., slope 21° , Laplacian parameter -0.18 (convex), topographic position index $TPI_{500m} = 52$ m; terrain characteristics from the nearest 100 m DEM grid cell). Both WRF-LES_{111m} and DEVINE introduce clear terrain-dependent local wind structure that aligns visually with the underlying DEM, in contrast to the smooth, largely homogeneous WRF_{1km} flow field, demonstrating the added value of downscaling low-resolution wind. During non-storm conditions, the spatial patterns of WRF-LES_{111m} and DEVINE around MTIC1 agree closely ($r = 0.73$, $\sigma_{\text{error}} = 0.4 \text{ m s}^{-1}$), whereas during the storm period DEVINE's wsp magnitude is locally lower than WRF-LES_{111m}'s, resulting in a lower spatial correlation ($r = 0.63$) and larger σ_{error} (1.5 m s^{-1}), consistent

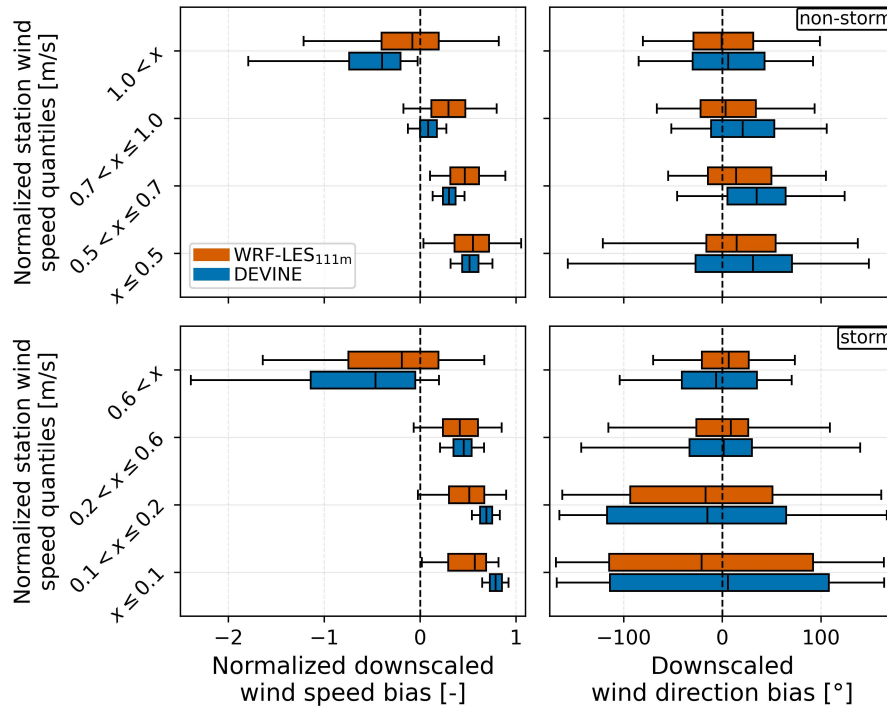


Figure 7. Normalized wind speed (left) and wind direction (right) bias between DEVINE, WRF-LES_{111m} and observations, grouped by quantiles of the normalized observed wind speed distribution, shown for the non-storm and storm period. Bias and observed wind speed is normalized by the corresponding low-resolution driving wind speed.

with DEVINE underperforming WRF-LES_{111m} on storm bulk metrics (Table 5). Considering temporal performance at the MTIC1 station location itself, DEVINE performs worse than both WRF models during the storm period, whereas during non-storm conditions DEVINE consistently outperforms WRF-LES_{111m} for *wsp* (MAE 1.50 vs. 1.71 m s⁻¹; bias -0.69 vs. -1.60 m s⁻¹; NRMSE 33 vs. 38 %; *r* 0.17 vs. 0.10) and for *wd* (MAE 30 vs. 34°; NRMSE 28 vs. 32 %). The station time series shows all three models capture the timing of the wind direction shift into the storm, but none reproduce the pronounced storm-period wind speed-up observed at the station (peaking near 28 m s⁻¹, versus 9-13 m s⁻¹ modeled). This suggests the local flow acceleration at MTIC1 reflects atmospheric-terrain flow features unresolved even by the dynamical WRF-LES model at 111 m grid spacing, a mismatch between the station's true location and its nearest grid cell, or both. Since DEVINE is trained only on terrain-forced flow, it cannot introduce additional local atmospheric processes acting in mountainous terrain. That WRF-LES itself also fails to reproduce the storm speed-up suggests this is not a limitation specific to DEVINE's terrain-forced-flow-only training, but rather points to atmospheric processes still unresolved at 111 m resolution. This is consistent with DEVINE capturing the non-storm terrain-forced flow well, while storm-period atmospheric processes it cannot represent, though whether this fully explains the storm-period gap, or additional processes are missing from WRF-LES_{111m} as well, remains an open question.

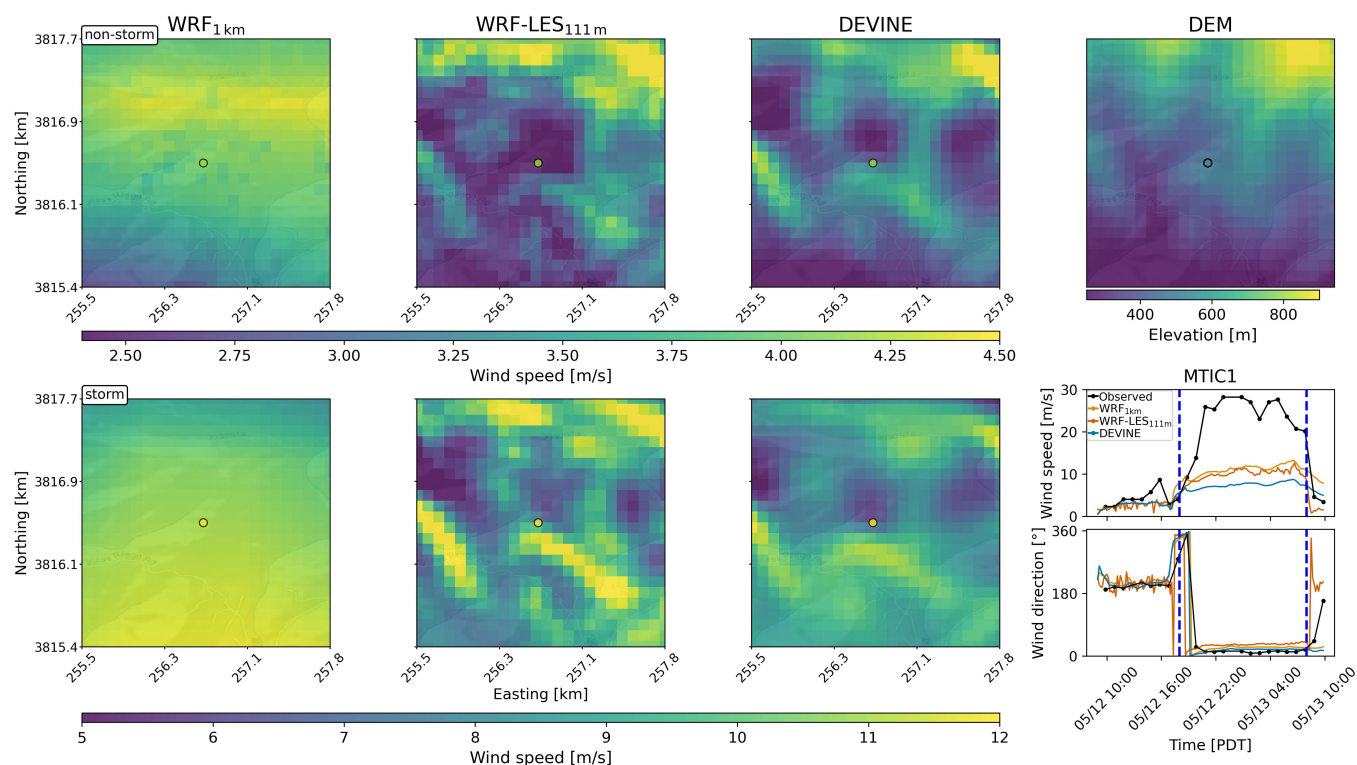


Figure 8. The first three columns show period-mean spatial wind speed from WRF_{1km}, WRF-LES_{111m}, and DEVINE for the coastal environment, for non-storm (top row) and storm (bottom row) periods, zoomed in around ground station MTIC1 (Fig. 1) in the foothills of the Santa Ynez Mountains (2.3 km x 2.3 km). The dot marks the station location, colored by the corresponding period-mean observed wind speed. The fourth column shows the DEM for the same region (top, station marked as an open circle) and time series of observed and modeled wind speed and wind direction for the full evaluation period (12–13 May), spanning both periods (bottom); blue vertical dashed lines in the time series indicate the approximate start and end of the storm period. Basemap: Esri National Geographic World Map | Powered by Esri (Esri, 2019).

Moving from spatial to point evaluation confirms patterns from the spatial evaluations at station locations: during non-storm conditions, DEVINE improves on WRF_{1km} and matches or outperforms WRF-LES_{111m} for *wsp*, consistent with DEVINE’s reliable transfer of terrain-forced flow across environments. This reverses during storm, when compound atmospheric-terrain flow produces features that neither DEVINE nor, in part, WRF-LES_{111m} fully resolve. While point-scale representativeness likely contributes to larger storm-period errors, the MTIC1 case study shows that even a dynamical model evaluated at its nearest grid cell fails to capture the local storm speed-up, indicating that missing atmospheric-terrain interaction processes, not representativeness alone, are the dominant driver of DEVINE’s storm-period station errors. This compound impact is also visible in the bias sign change of *wsp* itself as all models overestimate station-observed *wsp*, in contrast to DEVINE’s underestimation found against the spatial WRF-LES_{111m} reference.



610 5.4 Cross-environment comparison with atmospheric model wind

To derive a final cross-environment comparison, we examine DEVINE’s performance with atmospheric model wind as a function of the high-resolution reference wind speed magnitude (ARPS or WRF, depending on environment) across all evaluated settings (Fig. 9). We obtain a consistent qualitative pattern with DEVINE overestimating the lowest reference wind speeds *wsp* and underestimating the highest, across synthetic, alpine, and both coastal periods alike. The magnitude of this pattern, however, differs substantially between environments, while its underlying structure does not. For the synthetic environment, the maximum IQR of normalized *wsp* bias is roughly 0.12 (wind speed) and 6° for wind direction *wd*. Moving to the alpine environment widens this range substantially (wind speed: 0.27; wind direction: 15°), without changing the qualitative bias pattern, the alpine transition amplifies existing bias structure rather than introducing new patterns. Coastal maximum non-storm *wsp* IQR (0.29) is comparable to alpine, while coastal storm *wsp* IQR (0.59) is the widest of all four groups. For *wd*, the pattern differs: coastal non-storm (49°) is the widest of all four groups, slightly exceeding even coastal storm (44°), both far wider than alpine or synthetic. Overall, *wsp* bias scatters largest for the alpine environment at low-to-moderate *wsp*, exceeding even storm-coastal scatter, while at the highest *wsp* storm-coastal shows the largest scatter of all environments. Wind direction bias is less consistent across environments than *wsp* bias. Alpine shows a counterclockwise shift that grows with *wsp*, consistent with the counterclockwise shift found over steep, exposed alpine terrain, while coastal shows opposite-signed shifts between storm (counterclockwise) and non-storm (clockwise), without a comparably clear *wsp* dependence but overall the largest bias scatter. Synthetic only shows a marginal counterclockwise shift for largest *wsp*, consistent with the slight counterclockwise shift observed over the steepest slopes (Fig. S3). Together, this indicates that DEVINE’s core *wsp* bias behavior, overestimating the lowest *wsp*, underestimating the highest, transfers across mountain terrain, atmospheric driving model, and flow regime, while *wd* bias is more sensitive to the specific flow regime or atmospheric driving model.

630 6 Discussion

6.1 Strengths and limitations

DEVINE reliably reproduces terrain-forced wind structure across independent topographies and atmospheric driving models (Fig. 9), confirming DEVINE’s shown transferability in the Alps through station evaluations (Le Toumelin et al., 2023, 2024). We attribute this largely to the wide range of terrain characteristics covered by the synthetic GRF training dataset (Helbig et al., 2017). Le Toumelin et al. (2023) showed that terrain characteristics derived from a 30 m DEM covering the French Alps are well represented within this synthetic dataset, except for individual grid cells in extremely complex, steep, exposed terrain. The transferability holds for unseen topographies, including a real alpine and coastal mountainous domain, and for a change in the atmospheric driving model itself, from the ARPS simulations used during training to WRF simulations it never encountered. Bulk error metrics and spatial wind speed gradient comparisons both support this. Transitioning from synthetic to alpine terrain amplifies only the magnitude of existing errors, rather than altering their underlying structure or introducing new systematic errors (Table 3 vs. 4; Fig. 7). Moving to coastal mountains decreases overall performance, however, DEVINE continues to

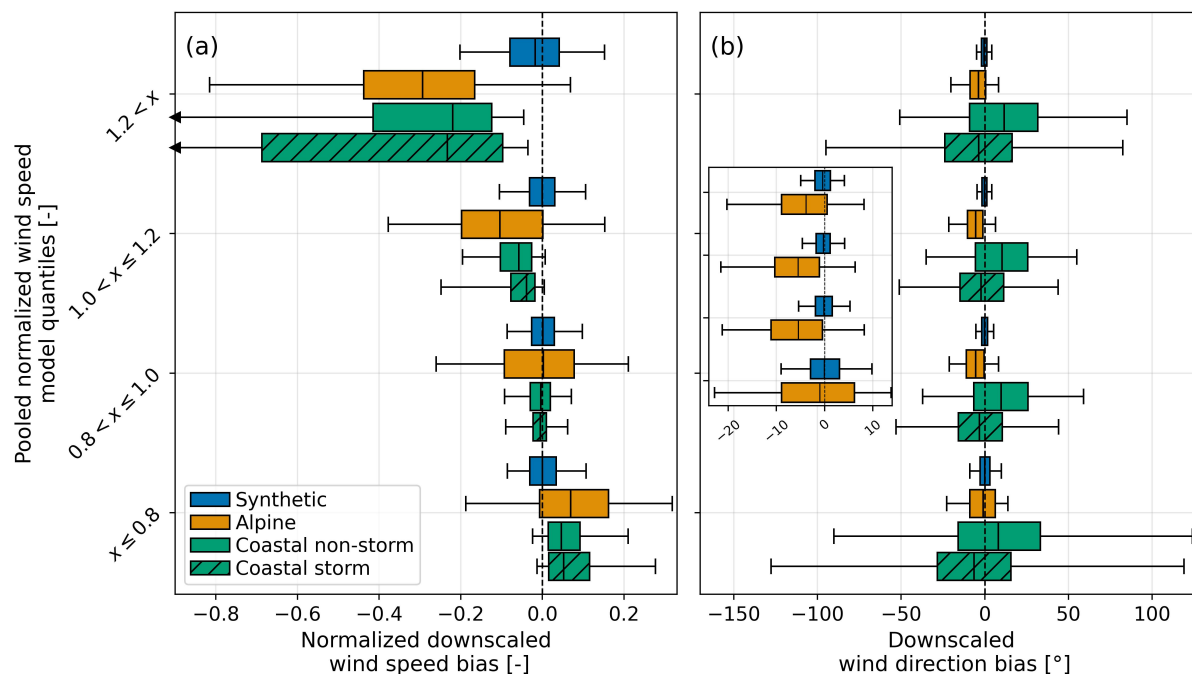


Figure 9. Normalized wind speed (a) and wind direction (b) bias between DEVINE and the high-resolution model, grouped by quantiles of the pooled normalized wind speed distribution, shown for the synthetic, alpine, and coastal environments. Bias and modeled wind speed is normalized by the corresponding low-resolution driving wind speed. Quantile thresholds are computed from the pooled normalized high-resolution wind speed distribution across all three environments, corresponding to an overall raw wind speed range of approximately $0.5\text{--}23\text{ m s}^{-1}$. The wind speed bias axis is truncated at -0.9 ; left-pointing arrows indicate that the 5th percentile whisker extends beyond the plotting range, to -3 and -1.8 for the coastal storm and non-storm periods, respectively.

reproduce terrain-induced spatial wind speed gradients reasonably well, even under a different driving model and mixed flow regimes (Section 5.3). For non-storm conditions, a mix of sea-breeze and terrain-forced flow, DEVINE even predicts station wind speed similar to or better than the dynamic model WRF-LES_{111m} in the coastal domain.

Thanks to DEVINE's well-defined training data, performance degradation points mostly to specific, identifiable causes rather than a general failure when applied out-of-sample. An example is the coastal environment's storm period, where DEVINE's relative error growth against station observations closely tracks that of the low-resolution driving model (WRF_{1km}) rather than exceeding it, and where DEVINE performs worse than WRF-LES_{111m} on nearly every metric, as WRF-LES_{111m} directly resolves atmospheric dynamics beyond terrain forcing (Table 5). This suggests DEVINE's storm-period degradation is inherited substantially from limitations already present in its driving data, rather than introduced independently by the downscaling step itself. In contrast to WRF-LES_{111m}, DEVINE cannot represent atmosphere-terrain interaction processes such as flow blocking, separation, or hydraulic jumps, since its training data represents only neutral, quasi-steady terrain-forced flow, without these processes present to learn from.



The relative complexity of the alpine and coastal topographies, judged against DEVINE’s GRF training coverage, offers a plausible explanation for where their error patterns diverge. The alpine domain has a substantially larger mean slope (37°) and wider local slope range ($0\text{--}78^\circ$) than the coastal domain (mean 18° , range $0\text{--}46^\circ$); while both fall broadly within the synthetic training distribution, alpine terrain sits toward the steep, complex GRF end where synthetic errors are also largest (Fig. S3). The coastal domain is less steep and more centrally situated within GRF characteristics, though it has a typical topographic feature height ($\sigma = 654$ m) that falls outside the range covered by the training data. This asymmetry may help explain why alpine shows stronger *wsp* scatter dependence across the modeled wind speed range than coastal non-storm (Fig. 9), and may also contribute to coastal’s overall elevated error level relative to alpine (e.g. roughly four- to sevenfold overall non-storm MAE increase from alpine to coastal; Table 4). The concentration of larger MAE of the coastal domain specifically in flattest, sheltered terrain, however, is better explained by wind-regime, as this is where sea-breeze and downslope-storm forcing dominate over terrain forcing, flow characteristics DEVINE cannot represent on its own (Fig. 6).

Despite these wind regime-dependent limitations, DEVINE’s behavior relative to atmospheric model wind remains stable across all three environments in one respect, a consistent bias direction, overestimating the lowest wind speeds and underestimating the highest, regardless of atmospheric driving model or flow regime (Fig. 9). Wind direction bias is comparatively less stable, consistent with DEVINE’s larger discrepancies in spatial gradients of the horizontal wind components (mean combined component gradient magnitude difference of 50–60 %; Figs. S7, S11), which are largest in steep and convex terrain. This instability is most strikingly illustrated by a sign reversal in wind direction bias: the alpine environment shows a counterclockwise shift, coastal non-storm instead shows a clockwise shift, reverting to counterclockwise during storm (Fig. S6 vs. S10). Why DEVINE introduces a clockwise shift specifically under coastal non-storm conditions, opposite to the alpine pattern, remains unclear and is not resolved by our current analysis. A related pattern appears temporally rather than spatially, which is that bias and correlation counterintuitively improve during the coastal storm period relative to non-storm (except for u), despite MAE, NRMSE, and σ_{error} all worsening (Table 4). This may parallel the alpine finding, where bias also grew disproportionately relative to MAE, NRMSE, and σ_{error} under the same quasi-steady, terrain-forced flow regime. While terrain-conditioned bias remains clearly present during storm (Fig. S10), bulk mean bias reduces relative to non-storm, which may be because increased flow variability during storm drives compensating biases across different terrain classes and times, while MAE and NRMSE, continue to grow.

Overall, our results indicate that DEVINE’s strength, robust terrain-forced flow generalization, and its central limitation, an inability to represent atmospheric-terrain interaction processes, follow directly from its design and training data. This clarifies where DEVINE can be applied with confidence (unseen mountainous terrain dominated by terrain-forced flow) and where results should be interpreted cautiously (regimes where downslope storms, sea-breeze, or synoptic forcing are expected to dominate).

Finally, as a result of its parsimony in input features and new wind direction-dependent features that replace the computationally costly rotations in the original DEVINE, the enhanced DEVINE achieves faster inference, a characteristic highly valued for operational applications. The runtime estimates indicate a substantial reduction in computational time compared with Le Toumelin et al. (2023), although the comparison is only indicative due to differences in hardware and prediction work-



loads. Beyond the reduction in inference time, eliminating this preprocessing step simplifies the workflow and facilitates the
 use and operational deployment of DEVINE. As a result, the enhanced DEVINE is planned for operational deployment at
 Météo-France in the near future within the upcoming EDELWEISS distributed snow prediction system (Haddjeri et al., 2024).

6.2 Representativity of evaluation approach

Station-referenced evaluation reveals systematically different error behavior in the coastal environment than spatial gridded
 evaluation with atmospheric model wind, for example, in bias sign (Fig. S10 vs. S12) and error magnitude (e.g. NRMSE 55 %
 to 130 % larger; Table 4 vs. 5), highlighting that evaluation choice itself shapes apparent model performance. Since statis-
 tically trained models are known to degrade out-of-sample, we tested this rigorously by evaluating DEVINE across distinct
 environments, wind regimes, and driving models. Spatial evaluation against high-resolution atmospheric model wind proved
 essential for identifying where and why performance degrades. It allowed systematic terrain-conditioned analysis across far
 greater variability than achievable with station observations alone. For instance, the coastal domain spans a substantially wider
 range of terrain characteristics than the station network samples (e.g. Laplacian: -0.30 to 0.28 vs. -0.18 to 0.11 slope: 0-46°
 vs. 1-21°), consistent with station networks typically being sited for specific operational purposes rather than representative
 model evaluation. That point evaluation alone can be misleading is illustrated by a single coastal station (MTIC1; Fig. 8).
 While its time series shows good non-storm agreement, none of the three models captures the observed storm-period speed-up.
 Spatial non-storm-mean maps around the station, however, show DEVINE and WRF-LES_{111m} both resolve substantial terrain-
 forced structure absent from the smooth WRF_{1km} input. WRF-LES_{111m}'s storm mean map even shows a large speed-up in the
 1.1 km station vicinity (up to 18 m s⁻¹), but still misses the station's own 28 m s⁻¹ peak by roughly 19 m s⁻¹. A dynamical
 LES model capturing strong local speed-up nearby, yet missing it at the exact station location, points towards how sharply
 point representativity can diverge from surrounding spatial performance. This means station networks alone, however carefully
 sited, cannot reliably distinguish a genuine model limitation from a local representativity artifact, whereas high-resolution spa-
 tial reference modeling can make this distinction. Our evaluation strategy therefore relied on both high-resolution reference
 atmospheric model wind and station observations. By additionally isolating terrain-forced flow first (synthetic, alpine), then
 introducing mixed atmospheric-terrain regimes (coastal non-storm, storm), and finally comparing spatial and point representa-
 tivity directly, we systematically disentangled DEVINE's strengths from its limitations.

6.3 Contribution to downscaling models

We stratified DEVINE's performance by terrain characteristics, wind regime, and atmospheric driving model. While previ-
 ous studies developing statistical downscaling models of similar complexity used different evaluation strategies and study
 regions, setting our results in this context helps interpret their significance. Dujardin and Lehning (2022) report, for 44 Swiss
 Alps stations over one year of hourly data, wsp bias/MAE/ r = -0.07 m s⁻¹/1.21 m s⁻¹/0.55 and wd MAE = 50.1° (wsp >
 1 m s⁻¹). Le Toumelin et al. (2023) evaluated the original DEVINE against 61 French Alps stations over three years: wsp
 bias/MAE/ r = -0.24 m s⁻¹/1.37 m s⁻¹/0.58, wd MAE = 57°. Le Toumelin et al. (2024) evaluated the original DEVINE against
 39 Pyrenees/Corsica stations over three years (wsp bias/MAE/ r = -0.13 m s⁻¹/1.51 m s⁻¹/0.73, wd MAE = 40°) and 55 Alps



stations over one year (wsp bias/MAE/ r = $-0.02 \text{ m s}^{-1}/1.29 \text{ m s}^{-1}/0.66$, wd MAE = 43°). Saigger et al. (2026) report wsp MAE of $1.6\text{--}4 \text{ m s}^{-1}$ and wd MAE of $70\text{--}100^\circ$ against three glacier stations in the Austrian Alps over 24 hours. DEVINE's station performance in coastal Southern California outside the storm period (wsp bias/MAE/ r = $0.49 \text{ m s}^{-1}/1.13 \text{ m s}^{-1}/0.50$, wd MAE = 52° ; Table 5) falls within the same range as these European mountain studies, despite being applied here to an unseen topography, wind regime, and driving model simultaneously. Storm-period station performance (wsp bias/MAE/ r = $3.91 \text{ m s}^{-1}/5.20 \text{ m s}^{-1}/-0.12$, wd MAE = 72°) is worse than all reported values; a direct comparison is limited, however, since the driving WRF_{1km} itself shows comparably low storm-period performance (wsp bias/MAE/ r = $4.98 \text{ m s}^{-1}/6.23 \text{ m s}^{-1}/-0.14$; wd MAE = 70°), which DEVINE notably consistently improves except for wd . Relative to spatial reference simulations rather than stations (ARPS/WRF-LES_{111m}; Table 4), DEVINE's non-storm coastal and alpine wsp MAE's are smaller and r larger and its alpine wsp MAE (0.23 m s^{-1}) is up to an order of magnitude smaller than any station-referenced value above, reflecting the grid-cell-versus-point representativity gap noted earlier. All DEVINE wsp MAE values against spatial references, from synthetic to the most extreme coastal storm ($0.15\text{--}1.60 \text{ m s}^{-1}$), are also lower than Saigger et al. (2026)'s comparison against 48 m-WRF-LES output (3.9 m s^{-1} wsp MAE), though the latter uses only three point locations over 24 hours rather than a full spatial grid and is therefore not directly comparable in scope.

Statistical downscaling models are often found to produce overly smooth spatial wind fields. Saigger et al. (2026), for instance, report smoother predictions particularly where sharp gradients occur, such as ridges and lee slopes. Our wind speed gradient magnitude comparisons (Fig. 4) quantify this, DEVINE underestimates wind speed gradient magnitudes by approximately 9 % on average in the alpine and 11 % in the coastal non-storm environment. This underestimation increases toward the most extreme terrain (steepest and flattest slopes, exposed and convex terrain) in the alpine environment, though only toward flatter terrain in the coastal non-storm case. The identification of such limitations in statistical downscaling models, supported by metrics such as the error in horizontal wind speed gradient magnitude, provides a basis for developing strategies to reduce this bias. One potential approach for future ML downscaling methods would be to incorporate spatial gradient errors into the training loss, which could help better preserve sharp spatial wind gradients.

While Le Toumelin et al. (2023, 2024) also applied DEVINE under a different driving model than used in training (AROME, rather than ARPS) in their station-based evaluations, our study is, to our knowledge, the first to systematically isolate driving-model change alongside wind regime and spatial terrain conditioning, and to complement point-based station evaluation with full spatial, grid-cell comparison against high-resolution atmospheric simulations. Unlike downscaling approaches that directly ingest atmospheric state variables e.g., Dujardin and Lehning (2022)'s use of COSMO output beyond wind alone, DEVINE's terrain-forced flow design means driving model changes affect it only through the low-resolution wind field itself, consistent with the robustness to driving model change found throughout our evaluation. This also reflects a broader architectural distinction of DEVINE, which is that DEVINE doesn't learn site-specific corrections, but is trained exclusively on synthetic terrain-flow relationships and never sees real station data during training. This likely explains why DEVINE's driving model robustness, and its comparable station-level performance in an unseen region (Table 5), is a characteristic of its terrain-forced flow design rather than of having learned regional or instrument-specific corrections.



In summary, DEVINE performs competitively with, and in several cases exceeds, published statistical downscaling models evaluated under comparable conditions, while having been tested here across a substantially broader range of terrain, wind regimes, and driving models than prior studies.

7 Conclusions

760 This study extended the CNN-based wind downscaling model DEVINE (Le Toumelin et al., 2023) by incorporating additional wind-direction dependent input features enabling rotational data augmentation while substantially simplifying its preprocessing. More importantly, in this study we performed a systematic assessment of the model beyond its original training data by evaluating performance across previously unseen mountain environments (synthetic topographies, the Swiss Alps, and coastal Southern California), contrasting wind regimes (terrain-forced flow, sea-breeze, and downslope storm conditions), different
 765 driving atmospheric models, and multiple reference datasets, including high-resolution atmospheric wind and ground station data, at both spatial and point scale.

The results demonstrated that DEVINE generalizes well to unseen mountain regions when the flow is primarily terrain-forced, confirming that the synthetic-topography training strategy captures transferable terrain-flow relationships. At the spatial, grid-cell scale, DEVINE achieved lower wind speed MAE and higher correlation across the alpine and non-storm coastal
 770 environments than any previous model's station-based evaluation, reflecting its performance once point-scale representativity effects are removed. At the point scale, its station-based performance under these unseen conditions remained comparable to, or improved on, that of previous ML near-surface wind downscaling studies, even though the transfer considered here simultaneously included unseen topography, wind regime, and driving model data, and, for non-storm wind speed, it matched the dynamical WRF-LES_{111m} model in NRMSE and outperformed it in MAE, while consistently improving on its own low-
 775 resolution driving wind speed data (WRF_{1km}) across all coastal wind regimes. Performance degraded, however, when shifts in wind regime introduced complex atmosphere-terrain interactions, such as the downslope windstorm event studied here. We found that station-based and spatial-pattern evaluations deviated substantially (e.g., even in *wsp* bias sign), indicating that evaluation choice itself shapes apparent performance, and that spatial, grid-cell comparison is the more reliable way to separate genuine model limitations from local representativity effects. Together, our findings showed that DEVINE's potential extends
 780 well beyond its training data, while also revealing where and why that generalization breaks down, highlighting the importance of systematic, across scale evaluation for identifying the limits of ML-based downscaling under unseen training conditions.

Because DEVINE's terrain-forced flow design separates terrain representation from atmospheric forcing, computationally efficient and transferable models built this way may serve not only as practical downscaling tools but also, more broadly, as a terrain-aware framework for assessing wind biases in low-resolution climate or operational forecast fields. Such a framework
 785 could complement dynamical downscaling and enable more terrain-disentangled model evaluation at station locations, since failing to account for local terrain effects in low-resolution wind may obscure errors that are in fact due to unresolved terrain, rather than genuine deficiencies in the low-resolution model.



Code and data availability. The DEVINE model is available from <https://doi.org/10.5281/zenodo.22962195> (Helbig and Hammer, 2026). The associated GitHub repository for DEVINE (<https://github.com/weid-ost/devine.git>) also provides scripts for the pre- and post-processing steps described in Section 4.1.4 and example data. Further measurement and model data are available from the corresponding author upon reasonable request.

Author contributions. Funding acquisition: NH, LC, CJ, SB; Conceptualization: NH, IG, SB; Methodology: NH, FH, GJD; Software: NH, FH, GJD, HM; Investigation: NH, FH, GJD; Data curation: NH, FH, GJD; Formal analysis: NH, FH, GJD; Validation: NH, FH, GJD, LC, CJ, HM, IG; Visualization: NH; Resources: LC, CJ; Supervision: NH; Writing - review & editing: All authors

795 *Competing interests.* The authors declare that they have no conflict of interest.

Acknowledgements. We thank Callum Thompson and Marian de Orla-Barile for providing access to and support with the station network data, and Callum Thompson for valuable discussions and data insights. We also thank Hannes Scherrer for adapting software code to ensure consistency between versions. Nora Helbig was funded by a grant of the Swiss National Science Foundation (SNF) (grant no. IZSEZ0_229691), as well as by the Swiss Federal Office of Energy (SFOE; grant no. SI/502636). Leila Carvalho, Charles Jones and Gert-
800 Jan Duine acknowledge the support from the US National Science Foundation NSF AGS-1921595 and NSF AGS-648 2331728. This work was supported by the National Center for Atmospheric Research, which is a major facility sponsored by the US National Science Foundation under Cooperative Agreement No. 1852977. We used AI tools to enhance language editing, clarity, and code development and reviewed all outputs.



References

- 805 Arthur, R. S., Lundquist, K. A., Wiersema, D. J., Bao, J., and Chow, F. K.: Evaluating Implementations of the Immersed Boundary Method in the Weather Research and Forecasting Model, *Monthly Weather Review*, 148, 2087 – 2109, <https://doi.org/10.1175/MWR-D-19-0219.1>, 2020.
- Blier, W.: The Sundowner Winds of Santa Barbara, California, *Weather and Forecasting*, 13, 702 – 716, [https://doi.org/10.1175/1520-0434\(1998\)013<0702:TSWOSB>2.0.CO;2](https://doi.org/10.1175/1520-0434(1998)013<0702:TSWOSB>2.0.CO;2), 1998.
- 810 Bühler, Y., Stoffel, A., and Salzmann, C. M.: Photogrammetric Drone Data Latschuelfurgga, <https://doi.org/http://dx.doi.org/10.16904/envdatat.374>, 2022.
- Cannon, F., Carvalho, L. M., Jones, C., Hall, T., Gomberg, D., Dumas, J., and Jackson, M.: WRF simulation of downslope wind events in coastal Santa Barbara County, *Atmospheric Research*, 191, 57–73, <https://doi.org/https://doi.org/10.1016/j.atmosres.2017.03.010>, 2017.
- Carvalho, L. M., Duine, G.-J., Jones, C., Zigner, K., Kane, H., Gore, C., Bell, G., Gamelin, B., Gomberg, D., Hall, T., Johnson, M., Dumas, J., Boldt, E., Hazard, R., and Enos, W.: The Sundowner Winds Experiment (SWEX) Pilot Study: Understanding Downslope windstorms in the Santa Ynez Mountains, Santa Barbara, California, *Mon. Wea. Rev.*, 148, 1519–1539, <https://doi.org/10.1175/MWR-D-25-0015.1>, 2020.
- 815 Carvalho, L. M. V., Duine, G.-J., Clements, C., Wekker, S. F. J. D., Fernando, H. J. S., Fitzjarrald, D. R., Fovell, R. G., Jones, C., Wang, Z., White, L., Bucholtz, A., Brewer, M. J., Brown, W., Burkhart, M., Creegan, E., Deng, M., de Orla-Barile, M., Emmitt, D., Greco, S., Hock, T., Kasic, J., Malarkey, K., Modjeski, G., Oncley, S., Rockwell, A., Seto, D., Thompson, C., and Vömel, H.: The Sundowner Winds Experiment (SWEX) in Santa Barbara, California: Advancing Understanding and Predictability of Downslope Windstorms in Coastal Environments, *Bulletin of the American Meteorological Society*, 105, E532 – E558, <https://doi.org/10.1175/BAMS-D-22-0171.1>, 2024.
- 820 Cullen, N. J., Steffen, K., and Blanken, P. D.: Nonstationarity of turbulent heat fluxes at Summit, Greenland., *Boundary-Layer Meteorol.*, 122, 439–455, <https://doi.org/10.1007/s10546-006-9112-2>, 2007.
- de Orla-Barile, M., Carvalho, L. M. V., Duine, G.-J., Jones, C., Wekker, S. F. D., Fernando, H. J. S., Modjeski, G., Clements, C., Seto, D., Brown, W., Gomberg, D., Kittell, R., and Murray, A. T.: Insights on the Spatiotemporal Variability of Downslope Winds in Coastal Santa Barbara: A Case Study from the Sundowner Winds Experiment (SWEX), *Monthly Weather Review*, 153, 2635 – 2649, <https://doi.org/10.1175/MWR-D-25-0015.1>, 2025.
- 825 Duine, G.-J., Jones, C., Carvalho, L. M., and Fovell, R. G.: Simulating Sundowner Winds in Coastal Santa Barbara: Model Validation and Sensitivity, *Atmosphere*, 10, <https://doi.org/10.3390/atmos10030155>, 2019.
- Duine, G.-J., Carvalho, L. M., Jones, C., and Zigner, K.: The Effect of Upstream Orography on the Onset of Sundowner Winds in Coastal Santa Barbara, CA, *J. Geophys. Res.-Atmos.*, 126, e2020JD033791, <https://doi.org/10.1029/2020JD033791>, 2021.
- Duine, G.-J., Carvalho, L. M., and Jones, C.: Mesoscale patterns associated with two distinct heatwave events in coastal Santa Barbara, California, and their impact on local fire risk conditions, *Weather and Climate Extremes*, 37, 100482, <https://doi.org/10.1016/j.wace.2022.100482>, 2022.
- 835 Duine, G.-J., Carvalho, L. M., Jones, C., Thompson, C. F., Brown, W. O., Dumas, J., Gomberg, D., Hall, T., and Kittell, R.: The Relevance of Mean-State Critical Levels for the Intensification of Downslope Winds in a Coastal Mountainous Environment, *Weather Forecast.*, 40, 307–318, <https://doi.org/10.1175/WAF-D-24-0084.1>, 2025.
- Dujardin, J. and Lehning, M.: Wind-Topo: Downscaling near-surface wind fields to high-resolution topography in highly complex terrain with deep learning, *Quarterly Journal of the Royal Meteorological Society*, 148, 1368–1388, <https://doi.org/10.1002/qj.4265>, 2022.
- 840



- Esri: National Geographic World Map, ArcGIS Online basemap, https://services.arcgisonline.com/ArcGIS/rest/services/NatGeo_World_Map/MapServer/, Sources: National Geographic, Esri, DeLorme, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC. Accessed 19 August 2026, 2019.
- Forthofer, J. M., Butler, B. W., and Wagenbrenner, N. S.: A comparison of three approaches for simulating fine-scale surface winds in support
845 of wildland fire management. Part I. Model formulation and comparison against measurements, *International Journal of Wildland Fire*,
23, 969–981, <https://doi.org/doi.org/10.1071/WF12089>, 2014.
- Gerber, F., Lehning, M., Hoch, S. W., and Mott, R.: A close-ridge small-scale atmospheric flow field and its influence on snow accumulation,
Journal of Geophysical Research: Atmospheres, 122, 7737–7754, <https://doi.org/https://doi.org/10.1002/2016JD026258>, 2017.
- Gerber, F., Besic, N., Sharma, V., Mott, R., Daniels, M., Gabella, M., Berne, A., Germann, U., and Lehning, M.:
850 Spatial variability in snow precipitation and accumulation in COSMO–WRF simulations and radar estimations
over complex terrain, *The Cryosphere*, 12, 3137–3160, <https://doi.org/10.5194/tc-12-3137-2018>, 2018.
- Gerber, F., Mott, R., and Lehning, M.: The Importance of Near-Surface Winter Precipitation Processes in Complex Alpine Terrain, *J. Hy-*
drometeor., 20, 177–196, <https://doi.org/10.1175/JHM-D-18-0055.1>, 2019.
- Göbel, M., Serafin, S., and Rotach, M. W.: Numerically consistent budgets of potential temperature, momentum, and moisture in Cartesian
855 coordinates: application to the WRF model, *Geoscientific Model Development*, 15, 669–681, <https://doi.org/10.5194/gmd-15-669-2022>,
2022.
- Goger, B., Rotach, M. W., Gohm, A., Stiperski, I., Fuhrer, O., and de Morsier, G.: A New Horizontal Length Scale for a Three-Dimensional
Turbulence Parameterization in Mesoscale Atmospheric Modeling over Highly Complex Terrain, *Journal of Applied Meteorology and*
Climatology, 58, 2087 – 2102, <https://doi.org/10.1175/JAMC-D-18-0328.1>, 2019.
- 860 Goger, B., Stiperski, I., Nicholson, L., and Sauter, T.: Large-eddy simulations of the atmospheric boundary layer over an Alpine glacier:
Impact of synoptic flow direction and governing processes, *Quarterly Journal of the Royal Meteorological Society*, 148, 1319–1343,
<https://doi.org/https://doi.org/10.1002/qj.4263>, 2022.
- Haddjeri, A., Baron, M., Lafaysse, M., Le Toumelin, L., Deschamps-Berger, C., Vionnet, V., Gascoin, S., Vernay, M., and Dumont, M.:
Analyzing the sensitivity of a blowing snow model (SnowPappus) to precipitation forcing, blowing snow, and spatial resolution, *The*
865 *Cryosphere*, 18, 3081–3116, <https://doi.org/10.5194/tc-18-3081-2024>, 2024.
- Helbig, N. and Hammer, F.: DEVINE, <https://doi.org/10.5281/zenodo.22962195>, 2026.
- Helbig, N. and Löwe, H.: Shortwave radiation parameterization scheme for subgrid topography, *J. Geophys. Res.*, 117,
<https://doi.org/10.1029/2011JD016465>, 2012.
- Helbig, N., Löwe, H., and Lehning, M.: Radiosity approach for the surface radiation balance in complex terrain, *J. Atmos. Sci.*, 66, 2900–
870 2912, <https://doi.org/10.1175/2009JAS2940.1>, 2009.
- Helbig, N., Mott, R., van Herwijnen, A., Winstral, A., and Jonas, T.: Parameterizing surface wind speed over complex topography, *J. Geophys.*
Res., 121, <https://doi.org/10.1002/2016JD025593>, 2017.
- Helbig, N., Mott, R., Bühler, Y., Le Toumelin, L., Bartelt, P., and Lehning, M.: Snowfall deposition in mountainous ter-
rain: a statistical downscaling scheme from high-resolution model data on simulated topographies, *Front. Earth Sci.*, 11,
875 <https://doi.org/10.3389/feart.2023.1308269>, 2024.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.:
The ERA5 global reanalysis, *Q. J. Roy. Meteor. Soc.*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.



- Jones, C., Carvalho, L. M., Duine, G.-J., and Zigner, K.: Climatology of Sundowner winds in coastal Santa Barbara, California, based on 30 yr high resolution WRF downscaling, *Atmospheric Research*, 249, <https://doi.org/10.1016/j.atmosres.2020.105305>, 2021.
- 880 Körner, C., Jetz, W., Paulsen, J., Beck, P. S. A., Bielczyk, N. Z., Bolliger, J., Cavender-Bares, J., Fisichelli, N. A., Grytnes, J.-A., Heegaard, E., Karger, D. N., Lenoir, J., Liss, K. N., Nascimbene, J., Price, J. T., Rixen, C., Sandvik, G., Schmid, B., Spehn, E. M., Steinbauer, M. J., and Treier, U. A.: A global inventory of mountains for bio-geographical applications, *Alpine Botany*, 127, 1–15, <https://doi.org/10.1007/s00035-016-0182-6>, 2017.
- Le Toumelin, L., Gouttevin, I., Helbig, N., Galiez, C., Roux, M., and Karbou, F.: Emulating the adaptation of wind fields to complex terrain with deep-learning, *Artificial Intelligence for the Earth Systems*, 2, 1–39, <https://doi.org/10.1175/AIES-D-22-0034.1>, 2023.
- 885 Le Toumelin, L., Gouttevin, I., Galiez, C., and Helbig, N.: A two-folds deep learning strategy to correct and downscale winds over mountains, *Nonlin. Processes Geophys.*, 31, 75–97, <https://doi.org/10.5194/npg-31-75-2024>, 2024.
- Liston, G. E. and Elder, K.: A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet), *J. Hydrometeorol.*, 7, 217–234, <https://doi.org/10.1175/JHM486.1>, 2006.
- 890 Liston, G. E. and Sturm, M.: A snow-transport model for complex terrain, *J. Glaciol.*, 44, <https://doi.org/10.3189/S0022143000002021>, 1998.
- Lundquist, K., Chow, F. . K., and Lundquist, J. K.: An Immersed Boundary Method for the Weather Research and Forecasting Model, *Monthly Weather Review*, 138, 796 – 817, <https://doi.org/https://doi.org/10.1175/2009MWR2990.1>, 2010.
- Mott, R. and Lehning, M.: Meteorological modeling of very-high resolution wind fields and snow deposition for mountains, *J. Hydrometeorol.*, 11, 934–949, <https://doi.org/10.1175/2010JHM1216.1>, 2010.
- 895 Raderschall, N., Lehning, M., and Schär, M.: Fine-scale modeling of the boundary layer wind field over steep topography, *Water Resour. Res.*, 44, <https://doi.org/10.1029/2007WR006544>, 2008.
- Reynolds, D., Haugeneder, M., Lehning, M., and Mott, R.: Intermediate complexity atmospheric modeling in complex terrain: is it right?, *Frontiers in Earth Science*, Volume 12 - 2024, <https://doi.org/10.3389/feart.2024.1388416>, 2024.
- Reynolds, D. S., Gutmann, E., Kruyt, B., Haugeneder, M., Jonas, T., Gerber, F., Lehning, M., and Mott, R.: The High-resolution Intermediate Complexity Atmospheric Research (HICAR v1.0) Model Enables Fast Dynamic Downscaling to the Hectometer Scale, *Geoscientific Model Development Discussions*, 2023, 1–30, <https://doi.org/10.5194/gmd-2023-16>, 2023.
- 900 Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Springer, Cham, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.
- 905 Ryan, B. C.: A mathematical model for diagnosis and prediction of surface winds in mountainous terrain, *J. Appl. Meteorol.*, 16, 571–584, [https://doi.org/10.1175/1520-0450\(1977\)016<0571:AMMFDA>2.0.CO;2](https://doi.org/10.1175/1520-0450(1977)016<0571:AMMFDA>2.0.CO;2), 1977.
- Saigger, M., Goger, B., and Mölg, T.: SNOWstorm (v1.0) – a deep-learning based model for near-surface winds and drifting snow in mountain environments, *EGUsphere*, 2026, 1–27, <https://doi.org/10.5194/egusphere-2025-5608>, 2026.
- Serafin, S., Schmidli, J., Goger, B., Ban, N., Westerhuis, S., Kotlarski, S., and Arpagaus, M.: TEAMx Numerical Modelling Plan, TEAMx Programme Coordination Office; https://www.teamx-programme.org/files/TEAMx_NumericalModellingPlan_final_22may2024.pdf; Accessed 2024-09-30, 2024.
- 910 Seto, D., Carvalho, L. M., Duine, G.-J., De Wekker, S., Wang, Z., Jones, C., Emmitt, D., Greco, S., Fernando, H. J., Modjeski, G., et al.: Mountain Waves, Downslope Jets, and Boundary Layer Interactions during the Sundowner Winds Experiment (SWEX)-IOP2, *Atmos. Res.*, 337, 108 920, <https://doi.org/10.1016/j.atmosres.2026.108920>, 2026.



- 915 Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., and Huang, X.-Y.: A Description of the Advanced Research WRF Model Version 4, Technical report, UCAR/NCAR, <https://doi.org/10.5065/1DFH-6P97>, 2019.
- Umek, L., Gohm, A., Haid, M., Ward, H. C., and Rotach, M. W.: Large-eddy simulation of foehn–cold pool interactions in the Inn Valley during PIANO IOP 2, *Quarterly Journal of the Royal Meteorological Society*, 147, 944–982, <https://doi.org/10.1002/qj.3954>,
920 2021.
- U.S. Geological Survey: 3D Elevation Program 1/3 arc-second Digital Elevation Model, U.S. Geological Survey Data Catalog, <https://data.usgs.gov/datacatalog/data/USGS:3a81321b-c153-416f-98b7-cc8e5f0e17c3>, accessed 2026, 2021.
- Vionnet, V., Martin, E., Masson, V., Lac, C., Naaim Bouvet, F., and Guyomarc’h, G.: High-Resolution Large Eddy Simulation of Snow Accumulation in Alpine Terrain, *Journal of Geophysical Research: Atmospheres*, 122, 11,005–11,021,
925 <https://doi.org/10.1002/2017JD026947>, 2017.
- Wagenbrenner, N. S., Forthofer, J. M., Lamb, B. K., Shannon, K. S., and Butler, B. W.: Downscaling surface wind predictions from numerical weather prediction models in complex terrain with WindNinja, *Atmospheric Chemistry and Physics*, 16, 5229–5241, <https://doi.org/10.5194/acp-16-5229-2016>, 2016.
- Whiteman, C. D.: *Mountain Meteorology: Fundamentals and Applications*, Oxford University Press, 2000.
- 930 Winstral, A., Marks, D., and Gurney, R.: An efficient method for distributing wind speeds over heterogeneous terrain, *Hydrol. Process.*, 23, 2526–2535, <https://doi.org/10.1002/hyp.7141>, 2009.
- Winstral, A., Jonas, T., and Helbig, N.: Statistical Downscaling of Gridded Wind Speed Data Using Local Topography, *J. Hydrometeorol.*, 18, 335 – 348, <https://doi.org/10.1175/JHM-D-16-0054.1>, 2017.
- Winstral, A., Magnusson, J., Schirmer, M., and Jonas, T.: The Bias-Detecting Ensemble: A New and Efficient Technique for Dynamically Incorporating Observations Into Physics-Based, Multilayer Snow Models, *Water Resources Research*, 55, 613–631,
935 <https://doi.org/10.1029/2018WR024521>, 2019.
- Wood, N. and Mason, P.: The pressure force induced by neutral, turbulent flow over hills, *Quart. J. R. Met. Soc.*, 119, 1233–1267, <https://doi.org/10.1002/qj.49711951402>, 1993.
- Xue, M., Droegemeier, K. K., and Wong, V.: The Advanced Regional Prediction System (ARPS) - A multi-scale, non-hydrostatic
940 atmospheric simulation and prediction model. Part I: Model dynamics and verification, *Meteorol. Atmos. Phys.*, 75, 161–193, <https://doi.org/10.1007/s007030070003>, 2000.
- Xue, M., Droegemeier, K. K., Wong, V., Shapiro, A., Brewster, K., Carr, F., Weber, D., Liu, Y., and Wang, D.: The Advanced Regional Prediction System (ARPS) - A multi-scale, non-hydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications, *Meteorol. Atmos. Phys.*, 76, 143–165, <https://doi.org/10.1007/s007030070003>, 2001.
- 945 Zachariassen, J., Zeller, K. F., Nikolov, N., and McClelland, T.: A Review of the Forest Service Remote Automated Weather Station (RAWS) Network, General Technical Report RMRS-GTR-119, U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, <https://doi.org/10.2737/RMRS-GTR-119>, 2003.
- Zigner, K., Carvalho, L. M., Jones, C., Benoit, J., Duine, G.-J., Roberts, D., Fujioka, F., Moritz, M., Elmquist, N., and Hazard, R.: Wildfire Risk in the Complex Terrain of the Santa Barbara Wildland–Urban Interface during Extreme Winds, *Fire*, 5, 138,
950 <https://doi.org/10.3390/fire5050138>, 2022.