https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

A Terrain-Aware Residual U-Net Framework for Hourly Kilometre-Scale 2 m Temperature Downscaling over Austria

Kelsey E. Ennis[1], Sebastian Scher[1,2]

[1]Wegener Centre for Climate and Global Change, University of Graz, Graz, 8010, Austria
[2]Department of Geography and Regional Sciences, University of Graz, Graz, 8010, Austria

Correspondence to: Kelsey E. Ennis (kelsey.ennis@uni-graz.at)

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

Abstract. High-resolution hourly 2 m air temperature fields are essential for weather, climate, and impact applications in complex terrain, but coarse atmospheric products cannot directly resolve the local thermal structure produced by Alpine topography. This study develops and evaluates a terrain-aware residual U-Net for kilometre-scale hourly 2 m temperature downscaling over Austria and the surrounding Alpine region. The model learns the correction between bilinearly interpolated ERA5 reanalysis and the Integrated Nowcasting through Comprehensive Analysis (INCA) high-resolution temperature analysis using dynamic atmospheric predictors, digital elevation model-derived terrain descriptors, and cyclical time features. Performance is evaluated over an independent 2023–2025 test period against INCA and compared with two reference approaches: 1) interpolated ERA5 and 2) a physically interpretable lapse rate baseline with monthly orographic and bias corrections. In addition to conventional deterministic skill metrics, the evaluation examines whether the downscaled fields reproduce physically meaningful Alpine temperature structure, including spatial error patterns, hourly lapse rate variability, elevation-dependent diurnal cycles, high-frequency spatial variability, ridge-valley temperature contrasts, and heat- and cold-event biases. The residual U-Net substantially improves predictive skill relative to both reference methods. It reduces root-mean-square error from 2.39 °C for interpolated ERA5 and 1.97 °C for the lapse rate baseline to 1.22 °C and reduces mean absolute error from 1.69 °C and 1.37 °C for ERA5 and the lapse rate baseline, respectively, to 0.89 °C. These improvements correspond to root-mean-square error reductions of approximately 49% relative to ERA5 and 38% relative to the baseline. The model also weakens terrain-locked error structures, produces smaller and less spatially coherent biases, and better reproduces observed lapse rate variability, high-elevation diurnal cycles, small-scale temperature variability, nocturnal ridge-valley contrasts, and event-scale warm and cold biases. A source-transfer experiment using the NASA MERRA-2 reanalysis predictors without retraining shows that the learned correction remains beneficial relative to raw MERRA-2, but with reduced skill compared to the ERA5-driven model, indicating sensitivity to the driving reanalysis distribution. Overall, the results demonstrate that residual deep-learning downscaling can generate physically realistic INCA-scale hourly temperature fields in Alpine terrain, while also highlighting the need to treat such products as high-resolution emulations of the reference analysis rather than direct observations.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

1 Introduction

High-resolution near-surface air temperature fields are needed for weather and climate applications at the spatial scales where impacts are experienced. Temperature influences human health, ecosystems, agriculture, hydrology, infrastructure, and energy demand, and its local variability is especially important during heatwaves, cold spells, and other high-impact conditions (e.g., McGregor et al., 2015). In regions of complex terrain, this variability can be strong over short distances because near-surface temperature is shaped by the interaction of large-scale atmospheric conditions with elevation, land-surface properties, and terrain-controlled boundary-layer processes. Coarse atmospheric products therefore often cannot represent the local temperature structure needed for impact assessment and regional climate analysis.

This challenge is especially relevant in Austria and the surrounding Alpine region, where steep relief and strong topographic heterogeneity produce pronounced terrain-dependent temperature patterns. These patterns are expressed not only in mean temperature gradients with elevation, but also in hourly lapse rate variability, elevation-dependent diurnal cycles, ridge-valley contrasts, and localized error structures that are difficult to recover from coarse reanalysis fields alone. The Integrated Nowcasting through Comprehensive Analysis (INCA) system was developed for high-resolution analysis and nowcasting in the eastern Alpine region and provides gridded meteorological analyses suitable for evaluating terrain-scale temperature variability over Austria (Haiden et al., 2011). In this study, INCA is used as the high-resolution reference analysis for developing and evaluating a Machine Learning (ML) based kilometre-scale 2 m temperature downscaling over Austria and the surrounding Alpine domain. Although INCA provides the high-resolution reference, the purpose of this study is not to replace INCA where such analyses are already available. Instead, INCA is used as the target for learning and evaluating a mapping from widely available coarse atmospheric fields to kilometre-scale terrain dependent temperature structure. This distinction is important because many intended downscaling applications do not have access to high-resolution analyses like INCA as an input. Climate model simulations, large ensembles, global reanalyses, and digital-twin workflows often provide only coarse large-scale atmospheric fields. ERA5 is therefore used here as the coarse driving dataset and this setup examines whether INCA-scale Alpine temperature structure can be reconstructed from the type of coarse atmospheric information that is more widely available.

Global reanalyses provide the large-scale atmospheric context needed for long-term weather and climate applications. ECMWF ERA5 is widely used because it provides hourly atmospheric data over multiple decades and combines numerical weather prediction with data assimilation (Hersbach et al., 2020). NASA MERRA-2 provides an additional reanalysis product that can be used to test whether a downscaling model remains useful when the driving atmospheric dataset is changed (Gelaro et al., 2017). However, these products are too coarse to directly resolve many Alpine terrain effects. ERA5, for example, is

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

available on a grid of roughly 0.25°, corresponding to about 28–30 km over central Europe, whereas the target INCA grid is approximately 1 km. This scale difference motivates the use of downscaling methods that can connect large-scale atmospheric information to local terrain-driven temperature structure.

Downscaling approaches can be divided into dynamical and statistical methods. Dynamical downscaling uses regional atmospheric models to simulate processes at finer resolution, but it is computationally expensive for long hourly archives and large ensembles (e.g., Giorgi and Mearns, 1991; Prein et al., 2015; Rummukainen, 2016). Statistical downscaling is more computationally efficient because it learns relationships between coarse-scale predictors and high-resolution observations or analyses, making it well suited for large datasets. However, statistical methods depend on the stability and transferability of the learned relationships, and their performance varies with region, variable, temporal scale, predictor selection, and evaluation metric (Wilby and Wigley, 1997; Fowler et al., 2007; Maraun et al., 2010; Maraun and Widmann, 2018). For high-resolution temperature downscaling in complex Alpine terrain, the key challenge is not only reducing mean error but also reproducing the spatial and temporal structures produced by the topography.

A common physically interpretable approach for temperature downscaling is an elevation or lapse rate correction. Because air temperature is strongly related to elevation, coarse temperature fields can be adjusted using the difference between high-resolution terrain and the terrain effectively represented by the coarse grid. Such approaches are transparent and computationally inexpensive, and previous studies have shown that elevation-aware corrections can improve coarse gridded temperature fields in complex terrain (e.g., Daly et al., 2008; Minder et al., 2010; Dutra et al., 2020). However, lapse rate corrections are limited because the temperature-elevation relationship is not fixed. It varies seasonally, diurnally, and synoptically, and it can be strongly modified by stable boundary layers, inversions, cold-air pooling, and local terrain influences. A lapse rate baseline therefore provides a physically meaningful benchmark: it captures the first-order orographic correction expected from elevation while allowing us to test whether the residual U-Net learns additional, time-varying sub-grid temperature structure beyond a transparent terrain-based adjustment.

Deep-learning methods provide a more flexible route for statistical downscaling. Convolutional neural networks (CNNs) and encoder–decoder models can learn nonlinear spatial relationships between coarse atmospheric predictors, high-resolution static fields, and high-resolution target variables. Early super-resolution approaches such as DeepSD introduced image-based downscaling concepts to climate fields (Vandal et al., 2017), while later studies evaluated deep learning for statistical downscaling over Europe and for gridded temperature downscaling in complex terrain (Baño-Medina et al., 2020; Sha et al., 2020). The U-Net architecture was originally introduced as an encoder–decoder architecture with skip connections for dense image-to-image prediction (Ronneberger et al., 2015). This structure is well suited to gridded meteorological downscaling because it can retain broad spatial context from coarse atmospheric predictors while reconstructing local spatial detail (Ayzel et al., 2020; Sha et al., 2020; Trebing et al., 2021; Lin et al., 2023). Recent work has extended machine-learning downscaling

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

toward high-resolution near-surface meteorological datasets, pan-European applications, kilometre-scale air-temperature super-resolution, and probabilistic or generative frameworks (Bhakare et al., 2024; Fuentes-Franco et al., 2025; Hess et al., 2025; Park et al., 2026).

Despite this progress, important limitations remain for Alpine 2 m temperature downscaling. Many existing studies focus on daily or monthly temperature, precipitation, or regional grids that are coarser than the local scales relevant in complex terrain. Hourly kilometre-scale temperature downscaling is more demanding because the predicted fields must follow large-scale atmospheric variability while also representing diurnal evolution, stability-dependent lapse rate changes, cold-air-pooling effects, and small-scale topographic structure. In addition, standard skill metrics such as root-mean-square error, mean absolute error, mean bias, and correlation provide only a partial assessment of model performance. These metrics quantify average agreement with a reference field, but they do not show whether errors remain topographically organized or whether the model reproduces physically meaningful Alpine temperature features. This distinction is especially important for deep-learning downscaling because a model optimized for pointwise accuracy may reduce mean error while still producing fields that are overly smooth or physically inconsistent. Therefore, evaluating high-resolution Alpine temperature downscaling requires both conventional deterministic skill metrics and diagnostics of realistic spatial and physical structure (Harder et al., 2023; Hess et al., 2025; Mardani et al., 2025).

The broader goal of this study is to test the feasibility of deep learning to learn high-resolution, temporally continuous representations of Alpine temperature variability that retain both large-scale atmospheric consistency and local terrain-driven structure. The study is therefore framed as a proof of concept: it asks whether a residual U-Net can learn the hourly mapping from coarse reanalysis-scale predictors to INCA-scale Alpine temperature structure before similar approaches are applied to climate simulations, large ensembles, or digital-twin systems where running computationally expensive dynamical downscaling for every application may be impractical. To address this goal, we develop a residual U-Net for hourly 2 m air temperature downscaling over Austria and the surrounding Alpine region. The model learns the residual between interpolated ERA5 and INCA 2 m temperatures using dynamic atmospheric predictors, high-resolution terrain information, and cyclical time features. This residual formulation preserves the large-scale temporal evolution of the reanalysis while allowing the network to learn local corrections associated with unresolved topography, systematic reanalysis error, and sub-grid thermal structure.

The primary objective of this study is to determine whether the terrain-aware residual U-Net can produce kilometre-scale hourly 2 m temperature fields that are more accurate than bilinearly interpolated ERA5 and a physically interpretable lapse rate baseline, and more physically realistic relative to INCA-based diagnostics of Alpine temperature structure. We compare the U-Net with bilinearly interpolated ERA5 data and with a physically interpretable lapse rate baseline that applies a first-order orographic correction and monthly bias adjustment. The evaluation first quantifies domain-wide predictive skill and

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

spatial error structure, then examines physical plausibility of the downscaled fields using diagnostics targeted at Alpine temperature behaviour: hourly lapse rate agreement, elevation-dependent diurnal cycles, high-frequency spatial variability, ridge-valley temperature contrasts, and extreme event-scale warm and cold biases. Finally, we test source-transfer sensitivity by applying the ERA5-trained U-Net to MERRA-2 predictors without retraining. This design allows us to assess not only whether deep learning improves hourly temperature downscaling, but whether the learned correction reproduces physically meaningful Alpine temperature structure and remains useful when the driving reanalysis source changes.

2 Data and pre-processing

2.1 Reference data, domain, and time period

Hourly 2 m temperature analyses from GeoSphere Austria's INCA are used as the high-resolution reference field. INCA is used here as the gridded target because it provides approximately 1 km analyses over Austria and resolves much of the near-surface temperature variability associated with Alpine terrain (Haiden et al., 2011). The study period extends from 1 January 2012 to 31 December 2025, corresponding to 122,736 nominal hourly time steps. The modeling grid is the rectangular INCA domain used for all model inputs and outputs, with 313 × 547 grid cells (y × x) at approximately 1 km spacing. This domain covers Austria and adjacent Alpine areas needed for spatial context. The primary evaluation region is the Austrian border mask shown in Fig. 1. All predictors, terrain variables, model outputs, and target fields are represented on this common INCA grid before model fitting or evaluation.

Quality control identified 36 hourly time steps for which the complete INCA 2 m temperature field was missing. These time steps were omitted from training, validation, testing, and diagnostic evaluation of all models. After the time steps were removed, non-finite target values at individual grid cells were not used for model fitting or evaluation; they were masked consistently in the training loss and in all reported metrics.

2.2 ERA5 atmospheric predictors

Large-scale atmospheric predictors are taken from the ERA5 reanalysis (Hersbach et al., 2020). ERA5 provides hourly global analyses on a grid of approximately 0.25 degrees, corresponding to about 28 km over central Europe. For this study, ERA5 was subset to the INCA period and to the Austria/Alpine region before interpolation. The ERA5 subset used before horizontal interpolation contained 11 latitude by 30 longitude grid points over the Austria-centered crop.

The dynamic predictor set consists of nine ERA5 variables: 2 m temperature, 10 m zonal wind, 10 m meridional wind, upper-layer volumetric soil water, surface pressure, hourly precipitation accumulation, 500 hPa geopotential height, 850 hPa temperature, and 300 hPa zonal wind. The ERA5 short names, physical quantities, vertical or surface levels, and roles of these predictors are listed in Table A1. These variables were selected to represent the near-surface thermodynamic state, low-level

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

flow, land-surface moisture, precipitation forcing, and larger-scale free-tropospheric temperature and circulation. They are also the typical variables analyzed in Earth System Model scenarios. Temperature variables were converted to °C where required, wind components to m s^{-1}, surface pressure to Pa, and elevation/geopotential height to metres.

2.3 Terrain data and derived topographic predictors

Static terrain information is derived from the Copernicus Digital Elevation Model (DEM) GLO-30, which provides global elevation data at 30 m spatial resolution. The DEM was aggregated/reprojected to the 1 km INCA grid using conservative resampling so that each INCA grid cell represents the mean high-resolution elevation within that cell. Elevation is used directly as a static predictor because Alpine 2 m temperature is strongly conditioned by terrain height, slope exposure, valley geometry, and cold-air pooling.

After this aggregation, the resulting 1 km elevation field on the INCA grid is denoted as $z(x, y)$ and is referred to here as the INCA-grid DEM. Four additional terrain descriptors are computed from this INCA-grid DEM: the east-west terrain gradient, the north-south terrain gradient, slope magnitude, and curvature. Using z(x,y) for elevation and x and y for the horizontal grid coordinates, these variables are defined in Eq. (1):

$$slope_x = \frac{\partial z}{\partial x}, \quad slope_y = \frac{\partial z}{\partial y}, \quad slope_{mag} = \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}, \quad curvature = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}. \tag{1}$$

The east-west and north-south terrain gradients are computed from the INCA-grid DEM using finite differences on the 1 km grid, with centered differences for interior grid cells and one-sided differences at the domain boundaries. Slope magnitude is then calculated as $\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}$ and curvature is approximated by the finite-difference Laplacian, $\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}$. All terrain derivatives are computed after aggregation to the INCA grid and are used only as static predictors. Together with elevation, these terrain descriptors provide the neural network with information about local relief, valley/ridge structure, and terrain gradients that are not represented explicitly by the coarse ERA5 grid.

2.4 Preparation of common model input fields

All ERA5 predictor fields were bilinearly interpolated from the cropped ERA5 grid to the common INCA grid. Static terrain variables were placed on the same grid using the DEM resampling described above. Four deterministic temporal predictors were also added for the neural-network input: sine and cosine encodings of hour of day and sine and cosine encodings of day of year. These variables provide the model with information about the diurnal and seasonal cycle without introducing artificial discontinuities at midnight or year boundaries. The resulting common input fields contain the INCA target temperature, nine interpolated ERA5 dynamic predictors, five static topographic predictors, and four cyclical time predictors on a common

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

hourly time axis and 1 km, 313 × 547 spatial grid. These fields provide the common basis for the lapse rate baseline, U-Net training, U-Net inference, and diagnostic evaluation.

The study domain and topography are shown in Fig. 1. The figure displays the Austrian evaluation mask within the larger INCA domain; the model itself is applied on the rectangular INCA grid, while Austria-masked diagnostics restrict the calculation to the Austrian region.

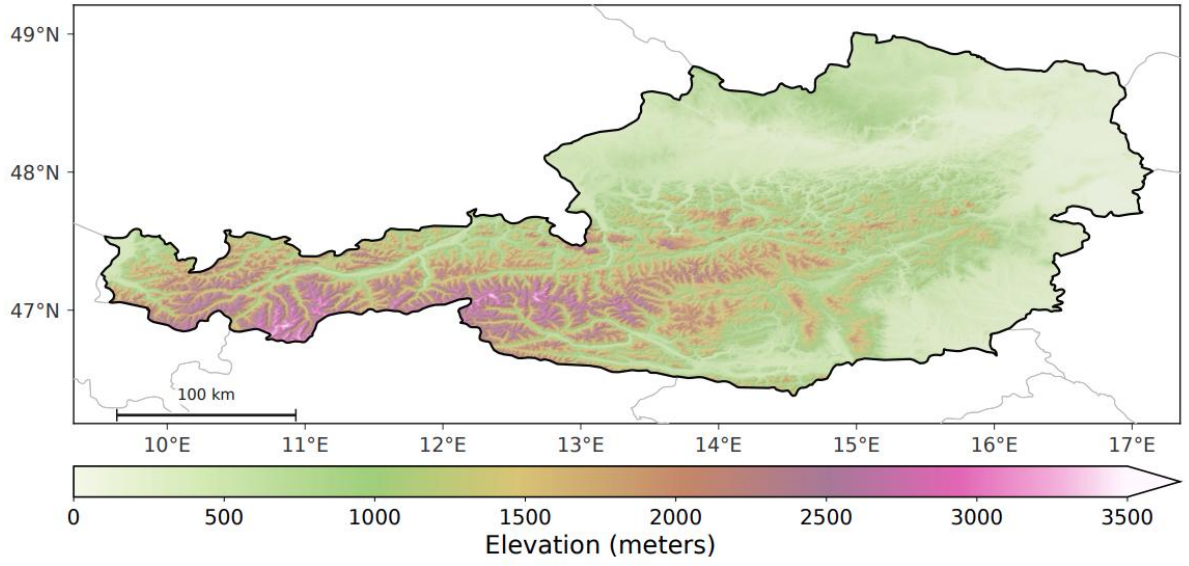

Figure 1: Study domain and topography used for hourly 2 m temperature downscaling. Shading denotes elevation from the Copernicus GLO-30 DEM after regridding to the INCA grid. The Austrian border (black outline) indicates the primary evaluation region.

2.5 Alternative global reanalysis dataset

In addition to ERA5, a MERRA-2 predictor dataset was prepared for a source-sensitivity experiment over 2023–2025. MERRA-2 was not used to train a new model. Instead, MERRA-2 fields were converted to the same variable interface and units as the ERA5 predictors and were substituted at inference time for the already trained ERA5-based U-Net. This experiment tests whether the learned residual downscaling remains robust when the driving reanalysis is changed. The mapped predictor list is: t2m, u10, v10, swvl1, sp, precip, z500, t850, and the jet-level zonal-wind input u300 (Table A1). For the jet-level wind channel, the model-facing u300 predictor is derived from the available MERRA-2 upper-level wind fields by interpolation to 300 hPa. In the processed MERRA-2 dataset used here, this is done by log-pressure interpolation between u250 and u500. The resulting field is retained under the name u300 only to match the predictor schema expected by the ERA5-trained U-Net; u250 and u500 are source variables and are not used as separate model input channels.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

3 Methods

3.1 Overview

The central modelling approach in this study is a residual convolutional deep neural network for hourly 2 m temperature downscaling. Two reference methods are evaluated alongside it: bilinearly interpolated ERA5, which represents the uncorrected coarse reanalysis field on the INCA grid, and a physically interpretable lapse rate baseline, which applies a first-order orographic correction and monthly bias adjustment.

The experiments have four objectives. First, they quantify whether any downscaling improves hourly 2 m temperature relative to interpolated ERA5. Second, they test whether the residual U-Net adds skill beyond a simple monthly lapse rate and bias correction. Third, they evaluate whether the downscaled fields reproduce physically meaningful Alpine temperature structure, including lapse rate behaviour, elevation-dependent diurnal cycles, high-frequency spatial variability, event-based extremes, and valley-ridge contrasts. Fourth, the MERRA-2 sensitivity experiment tests whether the trained U-Net model remains useful when the input reanalysis source is changed.

The data is split into training, validation and testing data using blocks of consecutive years to avoid data leakage due to temporal autocorrelation. The years 2012–2021 are the training data, which are used for model fitting and for estimating baseline correction parameters. The year 2022 is used for validation, including hyperparameter selection and training monitoring for the deep neural network. The years 2023–2025 are held out as an independent test period. A separate validation period is necessary because model choices such as learning rate, architecture, and stopping epoch must be selected without using the final test years; this preserves the independence of the 2023–2025 evaluation

3.2 Physically motivated lapse rate baseline

The lapse rate baseline provides a transparent benchmark for the U-net model. The baseline model starts from ERA5 2 m temperature bilinearly interpolated to the INCA grid and applies two corrections: a local topographic adjustment based on the difference between high-resolution terrain and coarse-scale effective terrain, and a month-dependent additive bias correction. For INCA grid coordinates (x,y), time t, and calendar month m(t), the baseline prediction is shown in Eq. (2):

$$T_{base}(x,y,t) = T^{I}_{ERA5}(x,y,t) + \gamma_{m(t)}[z(x,y) - z_c(x,y)] + \alpha_{m(t)}, \quad (2)$$

where T_{base} is the baseline downscaled temperature, T^{I}_{ERA5} is ERA5 2 m temperature after bilinear interpolation to the INCA grid, z is the INCA-grid DEM, z_c is the coarse-scale effective orography, $\gamma_{m(t)}$ is the month-dependent environmental lapse rate coefficient, and $\alpha_{m(t)}$ is the month-dependent additive bias correction. The term $z(x,y) - z_c(x,y)$ is the local elevation anomaly relative to the coarse terrain represented by the driving reanalysis.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

The coarse-scale orography z_c is obtained by applying a Gaussian smoothing filter to the INCA-grid DEM with **σ** = 35 grid cells, corresponding to an approximate 35 km smoothing scale on the 1 km INCA grid. This is a spatial smoothing operation: the output remains on the INCA grid but represents terrain variability at a scale comparable to the ERA5 grid spacing. It is therefore distinct from a domain mean or from coarsening to a single elevation value. Native ERA5 orography was not used directly because the baseline correction requires high- and coarse-scale terrain fields that are spatially collocated on the INCA grid. Smoothing the INCA-grid DEM provides a coarse-scale elevation reference at approximately 35 km while keeping the terrain field collocated with the INCA grid. The smoothed DEM is therefore used only to define the local elevation anomaly $z(x,y) - z_c(x,y)$, not as an independent terrain dataset.

Monthly lapse rate coefficients are estimated from the ERA5 lower-tropospheric temperature profiles over the 2012–2021 training period. The coefficient is a physically motivated environmental lapse rate estimate based on the vertical temperature gradient, $\frac{\Delta T}{\Delta z}$, rather than a parameter fitted to the test period. Similar elevation-correction approaches use reanalysis temperature profiles to estimate lapse rate information for downscaling temperature in complex terrain, where lapse rates vary seasonally, diurnally, and geographically (Minder et al., 2010). Using this logic, we obtain a first-order lower-tropospheric environmental lapse rate estimate through Eq (3):

$$\gamma_m = clip\left\{\frac{mean(T_{850,m}) - mean(T_{2m,m})}{z_{850} - mean(z_c)}, -0.012, -0.003\right\}\ K\ m^{-1}, \quad (3)$$

where overbars denote averages over the 2012–2021 training period and the study domain for calendar month. The value z_{850} is set to 1500 m as an approximate representative height of the 850 hPa level, and $mean(z_c)$. The clipping interval corresponds to -12 to -3 $K\ m^{-1}$ and is imposed as a physical constraint to prevent unrealistically weak or strong monthly lapse rate estimates. The clipping range was fixed before evaluation and was not tuned on the 2023–2025 test period. After applying the lapse rate term, the additive monthly bias correction is estimated from the training-period mean residual (Eq. (4)):

$$\alpha_m = mean_{train,m}\left\{T_{INCA}(x,y,t) - \left[T^I_{ERA5}(x,y,t) + \gamma_m\left(z(x,y) - z_c(x,y)\right)\right]\right\} \quad (4)$$

The baseline therefore preserves the large-scale time variability of ERA5 while adding a reproducible first-order terrain correction. Because it uses only a monthly lapse rate coefficient and a monthly additive bias term, it provides a strong but interpretable benchmark for determining whether the U-Net learns additional sub-grid temperature structure.

3.3 Residual U-Net downscaling model

The deep neural-network downscaling model is a convolutional U-Net implemented with Pytorch's segmentation models module. The U-Net architecture uses an encoder-decoder structure with skip connections, originally introduced for dense image-to-image prediction (Ronneberger et al., 2015). Related U-Net-based architectures have since been widely used for gridded Earth-system tasks such as precipitation nowcasting and climate downscaling because they combine multi-scale spatial context with reconstruction of local spatial detail (e.g., Ayzel et al., 2020; Trebing et al., 2021; Lin et al., 2023). This property

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

is well suited to statistical downscaling, where the output must retain the synoptic-scale structure of the coarse predictors while reconstructing local terrain-dependent temperature variations.

Here, "residual" refers to the model target rather than to the architectural skip connections within the U-Net. The network is trained to predict the $INCA - ERA5$ temperature correction rather than absolute 2 m temperature. The target residual is given by Eq. (5):

$$\Delta T(x,y,t) = T_{INCA}(x,y,t) - T^{I}_{ERA5}(x,y,t) \quad (5)$$

The neural network predicts a standardized version of this residual. After inference, the predicted residual is transformed back to degrees Celsius and added to interpolated ERA5 temperature as shown in Eq. (6) and Eq. (7):

$$\widehat{\Delta T}(x,y,t) = \mu_{\Delta} + \sigma_{\Delta} f_{\theta}\big(X'(x,y,t)\big) \quad (6)$$

$$T_{UNet}(x,y,t) = T^{I}_{ERA5}(x,y,t) + \widehat{\Delta T}(x,y,t) \quad (7)$$

In Eq. (6), f_{θ} is the trained U-Net, X' is the standardized input tensor, and μ_{Δ} and σ_{Δ} are the training-period mean and standard deviation of the residual target. This residual formulation stabilises optimisation and focuses the model on learning corrections to ERA5, rather than relearning the full large-scale temperature field.

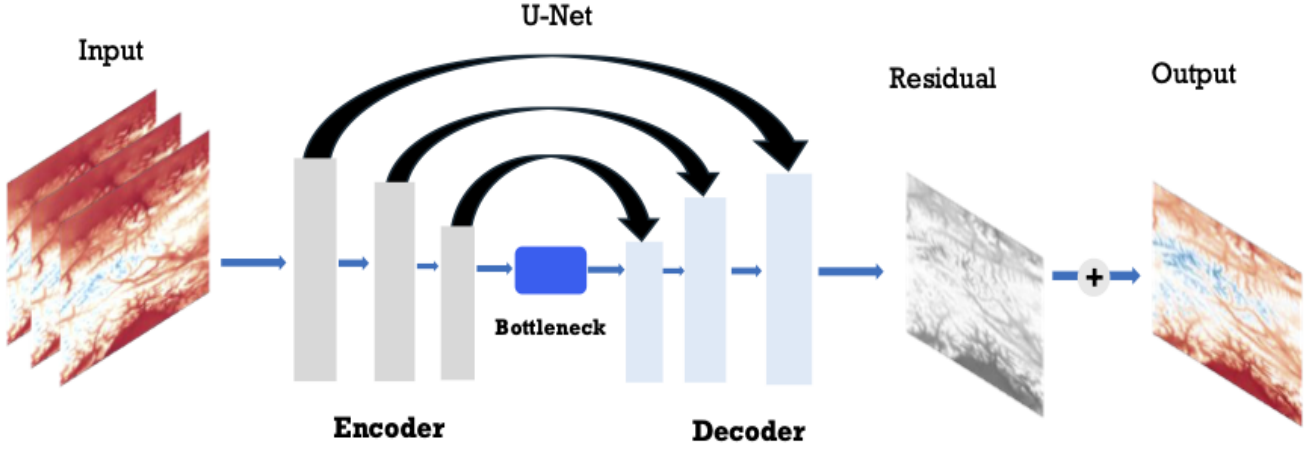

Figure 2: Residual U-Net configuration for hourly 2 m temperature downscaling. The input tensor combines interpolated reanalysis predictors, DEM-based terrain descriptors, and cyclical time features. The network predicts the local INCA minus ERA5 temperature residual, which is added back to interpolated ERA5 2 m temperature to produce the final downscaled field.

The U-Net input contains 18 channels: the nine dynamic atmospheric predictors listed in Table A1, five static topographic predictors derived from the INCA-grid DEM, and four cyclical time predictors representing hour of day and day of year as shown by Eq. (8) and Eq. (9), respectively:

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

$$hour_{sin} = sin\ sin\ \left(\frac{2\pi h}{24}\right),\ hour_{cos} = cos\ cos\ \left(\frac{2\pi h}{24}\right) \quad (8)$$

$$doy_{sin} = sin\ sin\ \left(\frac{2\pi d}{365.25}\right),\ doy_{cos} = cos\ cos\ \left(\frac{2\pi d}{365.25}\right) \quad (9)$$

All non-cyclical predictor channels are standardized using means and standard deviations computed from the 2012–2021 training period only. The residual target is standardized using the same training-period-only principle in Eq (10):

$$X_c' = \frac{X_c - \mu_c}{\sigma_c}, \quad \Delta T' = \frac{\Delta T - \mu_\Delta}{\sigma_\Delta} \quad (10)$$

Using training-period-only normalization prevents information from the validation or test periods from influencing model fitting. The same normalization statistics are reused for validation, 2023–2025 testing, diagnostic analysis, and MERRA-2 source-transfer inference.

The selected U-Net configuration is taken from an Optuna hyperparameter search (Akiba et al., 2019). Hyperparameters are modeling choices that are fixed before gradient-based training, such as the learning rate, weight decay, network width, dropout, and SmoothL1 transition parameter. They are tuned on the validation year rather than the test years because poor choices can strongly affect neural-network performance even when the training data and architecture class are unchanged.

3.4 Model training and loss function

The U-Net is trained on 64 × 64 grid-cell patches sampled from the INCA-domain dataset. The patch-sampling strategy is study-specific and was designed to prevent training from being dominated by common low-error conditions. This follows the general motivation of hard-example and imbalance-aware training (Shrivastava et al., 2016; Lin et al., 2017), in which rare, high-error, or otherwise informative samples are given increased representation during optimization. Spatially, patches are sampled across the domain with additional representation from low- and high-elevation terrain. Temporally, sampling weights are increased for time steps with large absolute INCA–ERA5 residuals, with additional emphasis on warm-season daytime cases and weaker emphasis on cold-season nighttime cases. This increases exposure to conditions in which terrain-dependent sub-grid corrections are expected to be most important. The sampling scheme was fixed before final testing and was not tuned for the 2023–2025 evaluation period.

The final model is trained with the Adam optimizer, validation-based learning-rate reduction, gradient clipping, and mixed-precision training where available. Model selection uses the validation period, and the final reported performance is evaluated only on the independent 2023–2025 test period.

The training objective combines standard robust residual fitting with study-specific spatial and terrain-aware regularization. The primary pixel-wise term is a weighted SmoothL1 loss applied to the standardized INCA-minus-ERA5 residual. SmoothL1 is closely related to the Huber loss and is less sensitive to large residual outliers than mean-squared error (Huber, 1964; Paszke

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

et al., 2019). The additional gradient and Laplacian terms are motivated by image-to-image prediction losses that penalize overly smooth outputs and encourage sharper spatial structure (Mathieu et al., 2015). The ridge-valley and valley-heating terms are study-specific physical regularizers introduced here to encourage realistic terrain-dependent temperature contrasts in Alpine terrain.

For pixel i in a training patch, let $e_i = \widehat{\Delta T_i'} - \Delta T_i'$ be the error in standardized residual space and $\Delta T_i'$ is the predicted standardized residual. The weighted SmoothL1 residual loss is given by Eq. (11) and Eq. (12):

$$L_{pixel} = \frac{\sum_i w_i \rho_\beta(e_i)}{\sum_i w_i} \quad (11)$$

$$\rho_\beta(e) = \{0.5e^2/\beta, |e| < \beta, |e| - 0.5\beta, otherwise \quad (12)$$

The weight map w_i increases the contribution of larger ERA5-INCA residuals and selected extreme or terrain-sensitive conditions, as shown by Eq. (13):

$$w_i = I_i \left[1 + 0.75 \left(\frac{|\Delta T_i|}{5}, 1\right) + 0.10 I_{night} + 0.18 I_{warm,day}\right], \quad (13)$$

where is I_i is 1 for valid target pixels and 0 otherwise, $\Delta T_i = T_{INCA,i} - T^I_{ERA5,i}$ identifies nighttime samples, and $I_{warm,day}$ identifies warm-season daytime samples.

Auxiliary losses are computed on reconstructed absolute temperature fields rather than directly on standardized residuals. The predicted and target absolute temperature fields are given by Eq. (14) and Eq. (15), respectively:

$$\widehat{T_i} = T^I_{ERA5,i} + \sigma_{\Delta T} \widehat{\Delta T_i'} + \mu_{\Delta T} \quad (14)$$

$$T_i = T^I_{ERA5,i} + \sigma_{\Delta T} \Delta T_i' + \mu_{\Delta T}, \quad (15)$$

where $\mu_{\Delta T}$ and $\sigma_{\Delta T}$ are the training-period mean and standard deviation of the INCA-minus-ERA5 residual.

Spatial structure is encouraged using gradient (Eq. (16)) and Laplacian (Eq. (17)) losses:

$$L_{grad} = \langle |\delta_x \hat{T} - \delta_x T| \rangle + \langle |\delta_y \hat{T} - \delta_y T| \quad (16)$$

$$L_{lapl} = \langle |\nabla_h^2 \hat{T} - \nabla_h^2 T|, \quad (17)$$

where δ_x and δ_y denote finite differences in the two horizontal grid directions and ∇_h^2 is the five-point discrete Laplacian.

For the terrain-aware training losses, valley and ridge masks are defined separately within each 64 × 64 training patch using the patch DEM. For patch p let $q_{25,p}(z)$ and $q_{75,p}(z)$ be the 25th and 75th percentiles of elevation within that patch. The patch-specific valley and ridge masks are shown by Eq. (18) and Eq. (19), respectively:

$$V_p = \{i \in p: z_i \leq q_{25,p}(z)\} \quad (18)$$

$$R_p = \{i \in p: z_i \geq q_{75,p}(z)\} \quad (19)$$

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

These masks are recalculated for each training patch and are used only in the training loss. They are distinct from the fixed evaluation masks used in Sect. 3.5, where valley and ridge regions are defined from the lowest and highest 20% of DEM elevations over the Austrian evaluation domain.

For any reconstructed temperature field A, the patch-level valley-minus-ridge contrast is defined by Eq. (20):

$$C_{\mathrm{vr}}(A,p) = \frac{1}{|V_p|}\sum_{i\in V_p} A_i - \frac{1}{|R_p|}\sum_{i\in R_p} A_i \tag{20}$$

The warm-season daytime (Eq. (21)) and cold-season nighttime (Eq. (22)) ridge-valley losses compare this contrast between the predicted and target absolute temperature fields using SmoothL1 loss:

$$L_{\mathrm{rv,day}} = \langle \rho_{0.5}\left[C_{\mathrm{vr}}(\hat{T},p) - C_{\mathrm{vr}}(T,p)\right]\rangle p \in \mathcal{D}_{\mathrm{warm,day}} \tag{21}$$

$$L_{\mathrm{rv,night}} = \langle \rho_{0.5}\left[C_{\mathrm{vr}}(\hat{T},p) - C_{\mathrm{vr}}(T,p)\right]\rangle p \in \mathcal{D}_{\mathrm{cold,night}} \tag{22}$$

A separate warm-season daytime valley-heating term compares the valley mean to the full-patch mean using Eq. (23) and Eq. (24):

$$C_{\mathrm{vh}}(A,p) = \frac{1}{|V_p|}\sum_{i\in V_p} A_i - \frac{1}{|p|}\sum_{i\in p} A_i \tag{23}$$

$$L_{\mathrm{heat,day}} = \langle \rho_{0.5}\left[C_{\mathrm{vh}}(\hat{T},p) - C_{\mathrm{vh}}(T,p)\right]\rangle p \in \mathcal{D}\mathrm{wr,dy} \tag{24}$$

Warm-season daytime is defined as May–September and 11:00–17:00 UTC, to emphasize periods when solar forcing, terrain-dependent daytime heating, and thermally driven slope and valley circulations are expected to be strongest. This is a fixed training design choice rather than a universal meteorological definition. This is consistent with previous studies examining how solar radiation drives mountain–valley circulation. For example, Schmid et al. (2020) identified May–September as the warm season and report strong warm season afternoon up-valley winds, supporting the physical motivation for this training emphasis.

Cold season nighttime samples are defined here as November–March and hours ≤ 06:00 or ≥ 20:00 UTC. This is also a training design choice rather than a universal meteorological definition. The November–March season follows Alpine studies that define winter as November–March (Bacer et al., 2024), while the nighttime hours target radiative cooling, stable stratification, drainage flows, and cold-air pooling in valleys (Vosper and Brown, 2008; McCaffrey et al., 2019).

The total training objective is given by Eq. (25):

$$L_{total} = L_{pix} + 0.12L_{grad} + 0.04L_{lapl} + 0.08L_{rv,day} + 0.18L_{heat,day} + 0.025L_{rv,night} \tag{25}$$

The coefficients in L_{total} are fixed empirical training design weights that were selected during preliminary model development intended to keep the auxiliary physical-structure losses secondary to the primary pixel-wise residual loss while still penalizing

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

overly smooth fields and unrealistic terrain-dependent temperature contrasts. The broad motivation for these terms is supported by previous studies on robust regression, hard-example training, derivative-based image losses, and mountain-boundary-layer processes, but the exact numerical values are study-specific and were held fixed before final test evaluation.

Although training is patch-based, final inference is performed over the full INCA grid. The trained U-Net predicts standardized residuals for overlapping tiles, the residuals are transformed back to degrees Celsius using the training-period residual statistics, and the reconstructed residual field is added to interpolated ERA5 2 m temperature. The same checkpoint and normalization statistics are used for validation, test inference, physical diagnostics, and the MERRA-2 source-sensitivity experiment.

3.5 Evaluation and physical diagnostics

Evaluation is performed over the independent 2023–2025 test period after excluding the identified missing INCA target times. Interpolated ERA5, the lapse rate baseline, and the residual U-Net are compared with INCA on the common high-resolution grid and common valid-time mask. For model M, the temperature error is given by Eq. (26):

$$e_M(x,y,t) = T_M(x,y,t) - T_{INCA}(x,y,t) \tag{26}$$

The reported scalar metrics are root-mean-square error (RMSE; Eq. (27)), mean absolute error (MAE; Eq. (28)), mean bias (Eq. (29)), and coefficient of determination (R^2; Eq. (30)):

$$RMSE = \sqrt{mean(e_M^2)} \tag{27}$$

$$MAE = mean(|e_M|) \tag{28}$$

$$Bias = mean(e_M) \tag{29}$$

$$R^2 = 1 - \frac{\sum(T_M - T_{\mathrm{INCA}})^2}{\sum(T_{\mathrm{INCA}} - \overline{T}_{\mathrm{INCA}})^2} \tag{30}$$

In addition to domain-wide scalar metrics, spatial error structure is evaluated using grid-cell RMSE and mean bias maps. For each grid cell, RMSE and bias are computed from the full 2023–2025 test-period time series using the same valid-time mask as the domain-wide metrics. Seasonal maps are computed analogously for DJF, MAM, JJA, and SON. These maps are used to assess whether errors are spatially uniform or concentrated in specific terrain regions, and to compare model skill between high-relief Alpine areas and flatter lowland regions.

Anomaly skill is evaluated relative to the same INCA-based monthly climatology. For each grid cell and calendar month, the INCA climatological mean over the 2023–2025 test period is computed. INCA, ERA5, baseline, and U-Net anomalies are then defined by subtracting this same INCA monthly climatology using Eq. (31):

$$A_D(x,y,t) = T_D(x,y,t) - mean_{test,month}[T_{INCA}(x,y,t)], \tag{31}$$

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

where D denotes INCA or one of the model products. Agreement between model anomalies and INCA anomalies is summarized using least-squares slope and intercept, Pearson correlation, and the coefficient of determination R^2.

Physical realism is assessed using diagnostics targeted at Alpine temperature structure. Lapse rate realism is evaluated by fitting, for each hourly field, a linear relationship between temperature and elevation as shown in Eq. (32):

$$T(x, y, t) = a_t + b_t z(x, y) \quad (32)$$

The slope b_t is converted to $K\ m^{-1}$ and compared among INCA, ERA5, the baseline, and the U-Net. This tests whether each product reproduces the observed coupling between near-surface temperature and terrain height.

Terrain-dependent diurnal cycles are evaluated by grouping grid cells into elevation bands and computing mean temperature as a function of hour of day. These diagnostics test whether the models reproduce the observed dependence of mean temperature and diurnal amplitude on elevation. Small-scale spatial variability is evaluated using the high-pass temperature field in Eq. (33):

$$H(T) = T - G_\sigma(T), \quad (33)$$

where G_σ is a Gaussian smoothing operator with σ = 8 grid cells corresponding to approximately 8 km on the 1 km grid. Because a Gaussian filter does not impose a sharp spectral cutoff, this diagnostic should be interpreted as emphasizing temperature variability at wavelengths of order tens of kilometres and smaller, rather than as power within a single discrete frequency band. The scale was chosen to isolate spatial variability below the approximate ERA5 grid scale while avoiding sensitivity to single-grid-cell noise.

High-frequency power is computed as Eq. (34):

$$P_{HF}(x, y, t) = H_T(x, y, t)^2 \quad (34)$$

For the KDE diagnostic, P_{HF} is first averaged over the sampled evaluation times at each grid cell, and kernel-density estimates are then computed from the spatial distribution of this time-mean high-frequency power. The same sampled times, spatial mask, KDE bandwidth, and axis limits are used for INCA, ERA5, the baseline, and the U-Net.

Topographic consistency is assessed using valley-ridge contrasts. Valley and ridge masks are derived from the lowest and highest 20% of DEM elevations within the evaluation domain. The valley-minus-ridge temperature contrast is computed separately for daytime and nighttime conditions to test whether each model captures terrain-dependent heating and nocturnal cold-pool structure.

Event-based extremes are evaluated using heat and cold events (Murray et al., 2021) identified from the INCA daily domain-mean 2 m temperature series. Heat events are defined using the upper-tail 95th percentile threshold and cold events using the corresponding lower-tail 5th percentile threshold, together with a minimum duration of at least two consecutive days. This

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

threshold-and-duration approach follows common heatwave/cold-spell practice in which events are defined relative to local climatology and persist for multiple days (Murray et al., 2021). In the 2023–2025 test period this procedure identifies 10 heat-event days and 11 cold-event days, with event durations of approximately 2–3 days. The same INCA-defined event windows are then used for all model products, ensuring that ERA5, the baseline, and the U-Net are evaluated under identical extreme-temperature conditions. Composite maps of model error during these windows are used to diagnose whether extreme-event biases are spatially organized by topography, elevation, valleys, or regional gradients.

3.6 MERRA-2 sensitivity experiment

A separate source-sensitivity experiment tests whether the trained ERA5-based U-Net is robust to a different driving reanalysis. For this experiment, a MERRA-2 inference dataset is constructed for the 2023–2025 test period using the same INCA target fields, INCA grid, DEM processing, and temporal evaluation mask as the ERA5 workflow. The MERRA-2 predictors are mapped to the same final INCA-grid input interface: the same 313 × 547 spatial grid, same variable order, same units, and same valid-time mask used for the ERA5-based inference. All MERRA-2 predictors were horizontally interpolated to the INCA grid using the same bilinear-interpolation convention applied to ERA5, so that differences in the source-transfer experiment arise from the reanalysis source rather than from a different model input grid.

No separate MERRA-2 model is trained. The ERA5-derived normalization statistics, U-Net architecture, Optuna-selected hyperparameters, trained checkpoint, and inference procedure are held fixed. Only the driving reanalysis predictors are replaced. The resulting comparison isolates the sensitivity of the learned residual correction to the reanalysis source and indicates whether the U-Net has learned corrections that transfer beyond the specific ERA5 predictor distribution used for training.

4 Results

This section evaluates the residual U-Net against INCA over the independent 2023–2025 test period and compares it with two reference methods: bilinearly interpolated ERA5 and the physically interpretable lapse rate baseline. The analysis first examines anomaly agreement and domain-wide deterministic skill, then assesses whether the models reproduce physically meaningful Alpine temperature structure through lapse rate behaviour, elevation-dependent diurnal cycles, high-frequency spatial variability, extreme-event composites, and ridge-valley contrasts. Finally, a MERRA-2 source transfer experiment tests whether the ERA5-trained U-Net remains useful when the driving reanalysis is changed without retraining.

4.1 Domain-wide predictive skill

The anomaly diagnostic first tests whether the downscaled fields reproduce departures from the local monthly mean temperature state. Figure 3 compares U-Net and baseline anomalies against INCA anomalies, where all anomalies are defined relative to the same INCA monthly climatology at each grid cell. This common reference removes the local monthly mean

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

temperature state but does not remove the diurnal cycle. The anomaly diagnostic therefore evaluates whether each model captures hourly departures from the local monthly mean, including both synoptic variability and diurnal variability during the test period.

As shown in Fig. 3a, the U-Net anomaly scatter is more tightly clustered around the one-to-one line – which would be perfect predictions – than the baseline scatter. The U-Net explains a larger fraction of the INCA anomaly variance, with $R^2 = 0.93$ (Fig. 3a) compared with $R^2 = 0.83$ for the baseline (Fig. 3b). The fitted slopes are similar for both models, indicating that both reproduce much of the anomaly amplitude, but the U-Net substantially reduces dispersion around the one-to-one line. The slightly negative U-Net intercept in Fig. 3a indicates a small cool offset in the anomaly fit; the corresponding domain-wide mean bias is quantified separately in Fig. 4c.

The domain-wide deterministic skill metrics in Fig. 4a–d show that both downscaling approaches improve substantially upon interpolated ERA5, with the residual U-Net achieving the strongest overall performance. Relative to ERA5, the lapse rate baseline reduces RMSE from 2.39 to 1.97 °C (Fig. 4a) and MAE from 1.69 to 1.37 °C (Fig. 4b), while also increasing R^2 from approximately 0.93 to 0.95 (Fig. 4d). The U-Net further improves performance, reducing RMSE to 1.22 °C and MAE to 0.89 °C (Fig. 4a,b), with R^2 increasing to approximately 0.98 (Fig. 4d). This corresponds to an RMSE reduction of approximately 49% relative to interpolated ERA5 and approximately 38% relative to the baseline. Bias remains small for all three products (Fig. 4c), although both downscaling approaches exhibit small cool biases. The U-Net has a somewhat larger mean cool bias than the baseline, but this bias is small relative to the reduction in RMSE and MAE. Together, these results show that the residual U-Net adds substantial predictive skill beyond both the interpolated ERA5 field and the physically interpretable lapse rate baseline.

4.2 Spatial distribution of errors

The domain-wide metrics summarize overall predictive skill, but they do not show where errors occur within the Alpine domain. To assess the spatial structure of model performance, Fig. 5 shows annual grid-cell RMSE against INCA over the 2023–2025 evaluation period for interpolated ERA5, the lapse rate baseline, and the residual U-Net. The RMSE maps show that errors are strongly structured by topography. ERA5 has the largest errors across the high-relief Alpine part of the domain, especially in western and central mountainous regions where unresolved elevation gradients, valley systems, and ridge-scale thermal contrasts are most important (Fig. 5a). The baseline reduces RMSE relative to ERA5 over much of Austria, but larger errors remain spatially organized along Alpine terrain features (Fig. 5b). The U-Net produces the lowest RMSE overall and strongly weakens the terrain-locked error pattern, with particularly clear improvements across the Alpine region while maintaining low errors over flatter lowland areas (Fig. 5c).

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

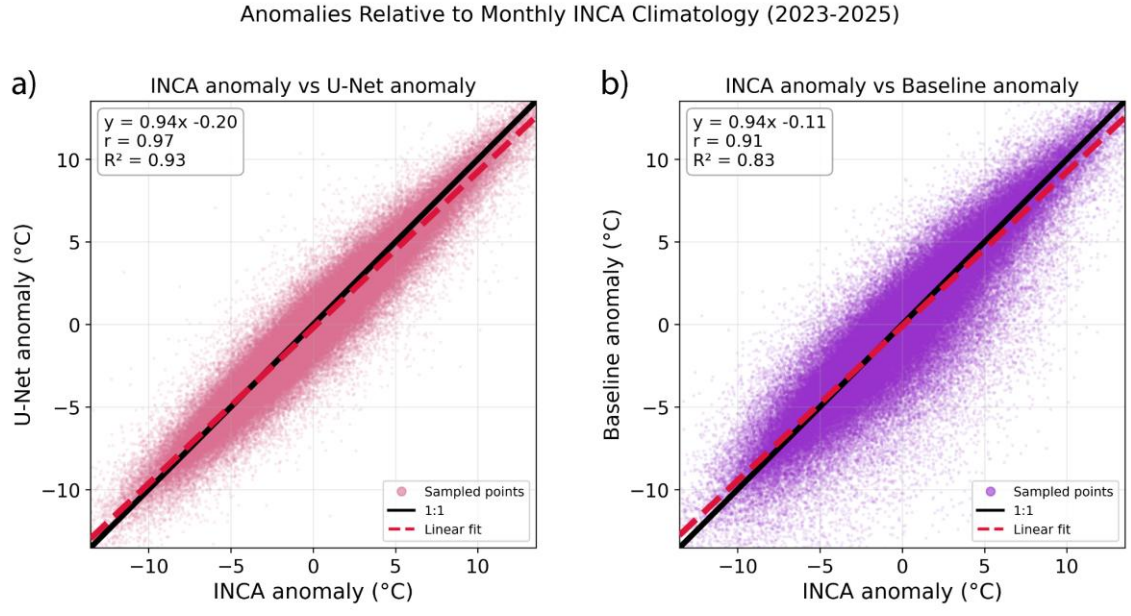

Figure 3: Monthly temperature anomalies relative to the local INCA climatology for the 2023–2025 evaluation period. Anomalies for the (a) U-Net and (b) baseline are defined relative to the same monthly INCA climatology at each grid cell, allowing direct comparison of departures from the local seasonal cycle.

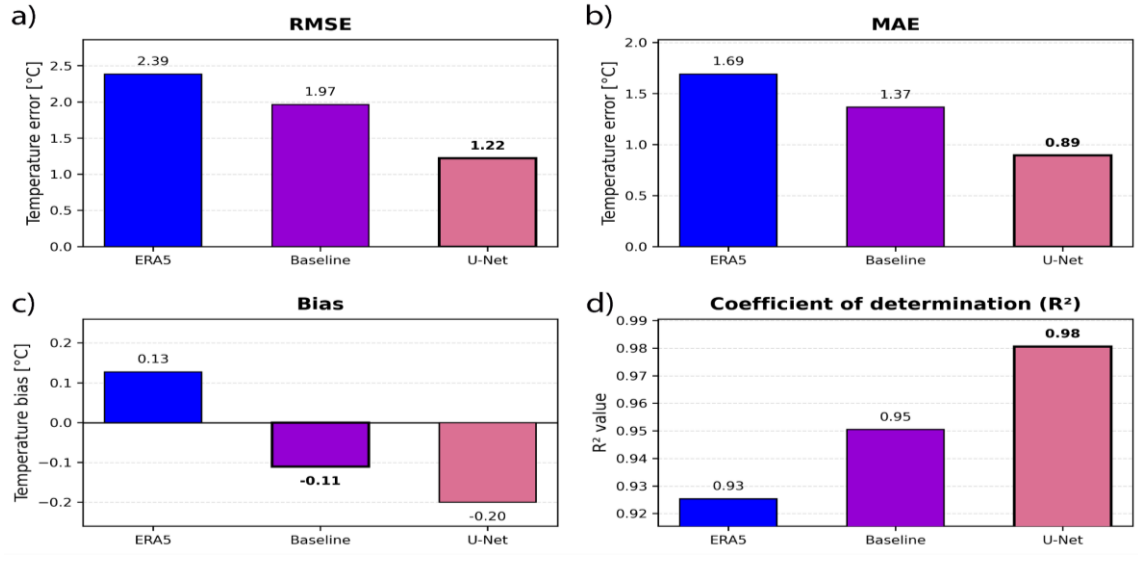

Figure 4: Domain-wide predictive skill against INCA for 2023–2025: (a) RMSE, (b) MAE, (c) mean bias, and (d) coefficient of determination (R^2). Metrics are shown for interpolated ERA5, the lapse rate baseline, and the residual U-Net using the same valid time–pixel sample. The best model per metric is indicated with a bold black frame.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

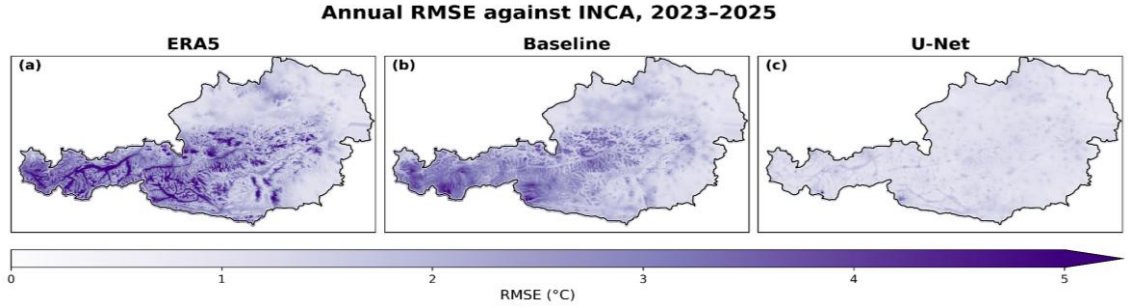

Figure 5: Annual grid-cell (1 x 1 km) RMSE against INCA over the 2023–2025 evaluation period for (a) interpolated ERA5, (b) the lapse rate baseline, and (c) the residual U-Net. RMSE is computed independently at each grid cell using the same valid evaluation times as the domain-wide metrics.

The annual bias maps in Fig. 6 provide a complementary view of spatial error structure. ERA5 exhibits strong, spatially coherent warm and cool bias patterns that follow terrain features, indicating that many of the ERA5 errors are linked to unresolved topographic contrasts rather than being spatially random (Fig. 6a). The baseline reduces some of this fine-scale terrain bias but has broader regional bias structures (Fig. 6b). In contrast, the U-Net bias field is much closer to zero across most of the domain, with weaker and less coherent residual biases than either ERA5 or the baseline (Fig. 6c). Some localized biases remain, but their magnitude and spatial organization are substantially reduced. Together, Figs. 5 and 6 show that the U-Net improvement is not only a domain-mean effect: the model also improves the spatial distribution of errors, with the largest practical gains in high-relief Alpine terrain.

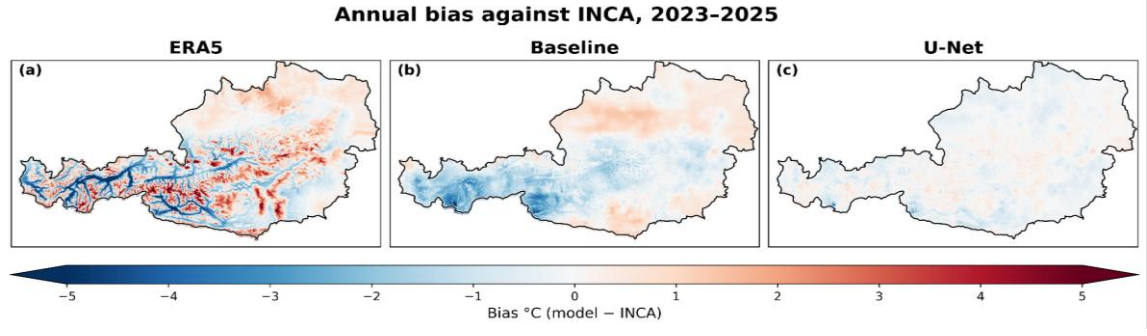

Figure 6: Annual grid-cell mean bias against INCA over the 2023–2025 evaluation period for (a) interpolated ERA5, (b) the lapse rate baseline, and (c) the residual U-Net. Bias is defined as a model minus INCA and is computed independently at each grid cell using the same valid evaluation times as the domain-wide metrics.

Seasonal RMSE maps in Fig. S1 shows that the topographic structure of error persists throughout the year, but with seasonal differences. ERA5 has large terrain-organized RMSE in all seasons, with particularly strong Alpine error patterns in DJF and

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

SON and sharper ridge- and valley-linked errors during MAM and JJA (Fig. S1a–d). The baseline reduces RMSE in every season, but residual errors remain concentrated over the western and central Alpine terrain, especially during DJF and SON (Fig. S1e–h). The U-Net maintains the lowest and least topographically organized RMSE in all seasons, with only weak residual Alpine structures compared with ERA5 and the baseline (Fig. S1i–l). This seasonal behavior indicates that the U-Net improves spatial skill year-round rather than only during a single season.

The seasonal bias maps in Fig. S2 show that the sign and structure of model error also vary through the year. ERA5 exhibits persistent warm and cool bias patterns, with alternating small-scale biases over Alpine terrain during MAM and JJA and broader cool Alpine biases during DJF and SON (Fig. S2a–d). The baseline reduces some of the ERA5 fine-scale bias structure but retains broader seasonal bias patterns, including cool biases across large parts of the Alpine region and warm biases in some northern and lower elevation areas, most clearly during DJF and SON (Fig. S2e–h). The U-Net bias fields remain much closer to zero in every season, with only localized residual cool and warm patches (Fig. S2i–l). Overall, the U-Net reduces both RMSE and spatial bias across the full evaluation period, with especially strong improvements in mountainous terrain while preserving comparatively weak errors over flatter regions.

4.3 Physical plausibility of downscaled temperature fields

The domain wide scalar metrics demonstrate that the residual U-Net shows skill relative to ERA5 and the baseline, but the physical diagnostics test whether those improvements correspond to more realistic Alpine temperature structure. The following diagnostics evaluate terrain temperature coupling, elevation dependent diurnal variability, and fine-scale spatial variability, following the diagnostic framework described in Sect. 3.5.

4.3.1 Hourly lapse rate agreement

The hourly lapse rate diagnostic in Fig. 7a,b reveals a clear difference in how the two downscaling approaches represent terrain–temperature coupling. The U-Net closely follows the one-to-one line in Fig. 7a, with a fitted slope of 0.99 and $R^2 = 0.97$, indicating that it reproduces both the mean lapse rate relationship and much of its hourly variability. In contrast, the baseline shows a flatter fitted slope of 0.57 and $R^2 = 0.63$ (Fig. 7b). Although the baseline captures the broad cooling-with-height structure imposed by the lapse rate correction, it compresses the observed range of hourly lapse rate variability. This is expected from a model with monthly lapse rate coefficients and a fixed topographic correction: it can represent the mean elevation dependence, but it cannot fully adjust to hourly changes in stability, inversion strength, or synoptic conditions. The residual U-Net therefore improves not only the average temperature error, but also the temporal variability of the terrain temperature relationship.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

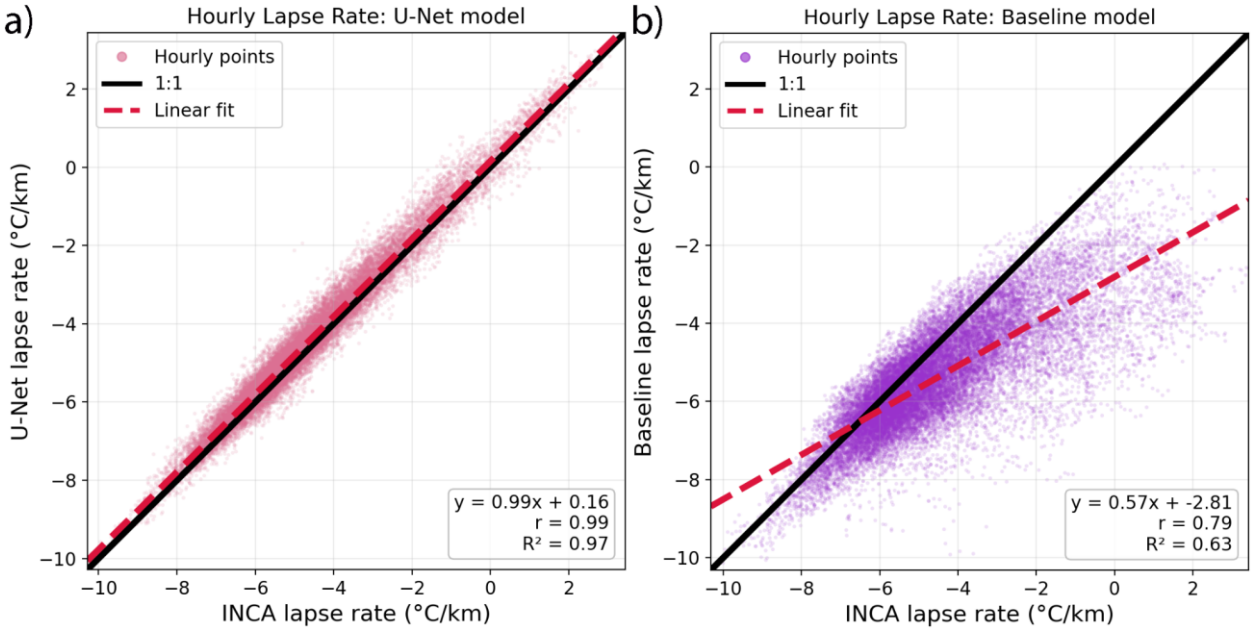

Figure 7: Hourly lapse rate performance of the residual (a) U-Net and the (b) baseline for 2023–2025. Each point represents one hourly domain wide temperature elevation slope, with the model lapse rate plotted against the corresponding INCA lapse rate. The solid black line denotes one-to-one agreement, and the dashed red line denotes the fitted linear regression.

4.3.2 Elevation-dependent diurnal variability

Elevation binned diurnal cycles provide a complementary test of whether the models reproduce the dependence of daily temperature variability at varying altitudes. Figure 8 shows that at low elevations, both models and ERA5 are consistent with INCA, but their behavior diverges with increasing altitude. Above 1500 m, the baseline (Fig. 8b) shows progressively larger errors in both mean level and diurnal amplitude, with the largest deviations occurring in the highest elevation bands. In contrast, the U-Net (Fig. 8a) remains much closer to the observed INCA cycles across the elevation range, indicating that it captures the terrain dependent daily temperature variability more accurately.

4.3.3 High-frequency spatial variability

The high-frequency power diagnostic evaluates whether the downscaled fields recover small-scale spatial variability that is absent from the coarse reanalysis. The diagnostic should be interpreted as emphasizing spatial variability at wavelengths of order 30–40 km and smaller, rather than power above a single discrete frequency threshold. In Fig. 9, ERA5 has the lowest high-frequency power, as expected for a coarse-resolution product interpolated to the INCA grid. Both downscaling approaches shift the distribution toward higher values, indicating that they introduce additional sub-grid spatial structure. The baseline increases high-frequency power relative to ERA5, but its distribution remains shifted toward lower values than INCA,

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

especially through the main part of the distribution. This indicates that the lapse rate correction restores some terrain-related spatial variability but remains too smooth overall. The U-Net distribution is shifted closer to INCA than the baseline, with less probability mass at very low high-frequency power and more mass at intermediate-to-high values. This shows that the residual model recovers substantially more small-scale temperature structure than either ERA5 or the baseline. However, the U-Net still does not fully match INCA; its distribution remains displaced toward lower high-frequency power for part of the range, indicating a residual smoothing of the finest resolved temperature variability. Overall, Fig. 9 shows that the U-Net provides the most realistic representation of high-frequency spatial variability among the tested model products, while still retaining a modest smoothness bias relative to INCA.

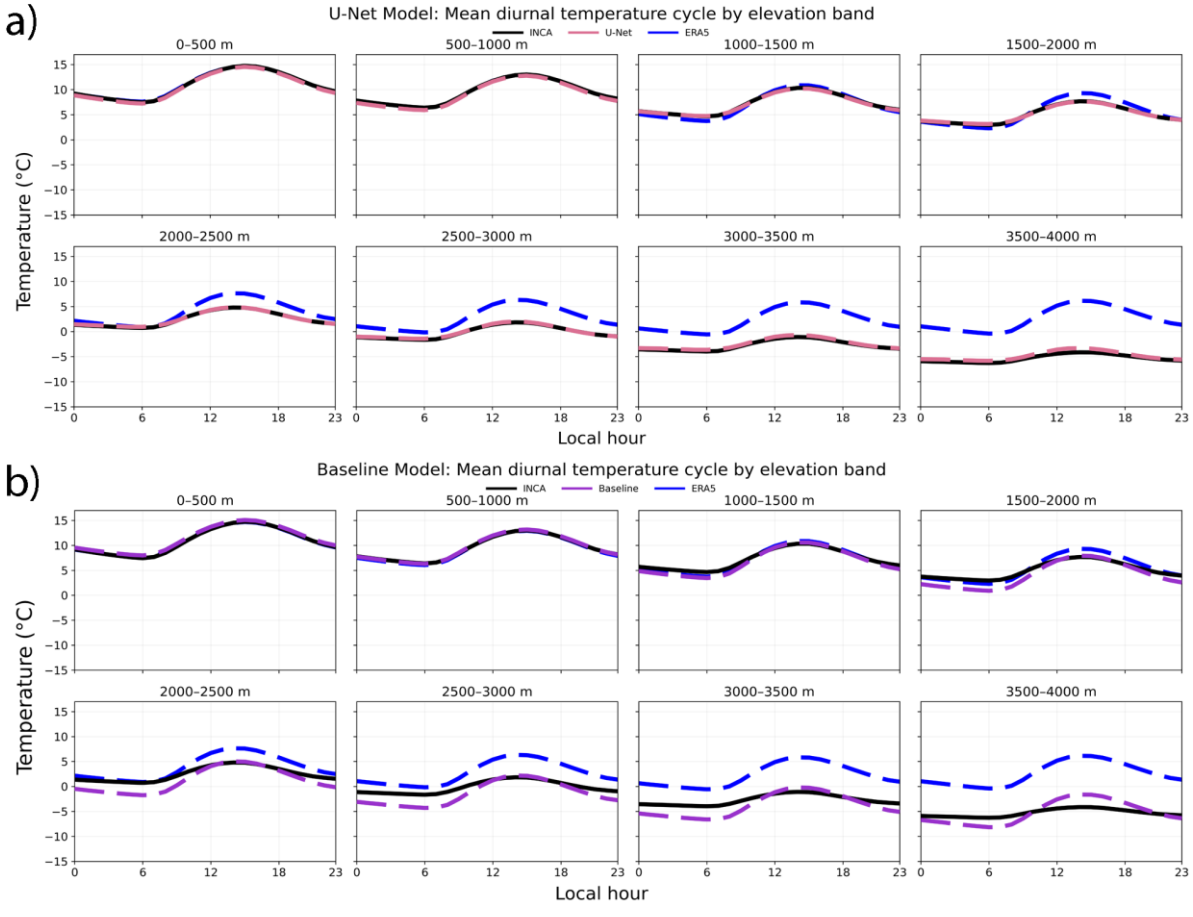

Figure 8: Mean diurnal cycle of 2 m temperature by elevation band for 2023–2025: (a) residual U-Net and (b) lapse rate baseline. Each panel compares the downscaled product with INCA and interpolated ERA5 for the same elevation bands and valid evaluation times. Black lines show INCA, pink lines show the U-Net in panel (a), purple lines show the baseline in panel (b), and blue dashed lines show ERA5. The x-axis gives local hour, and the y-axis gives temperature in °C.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

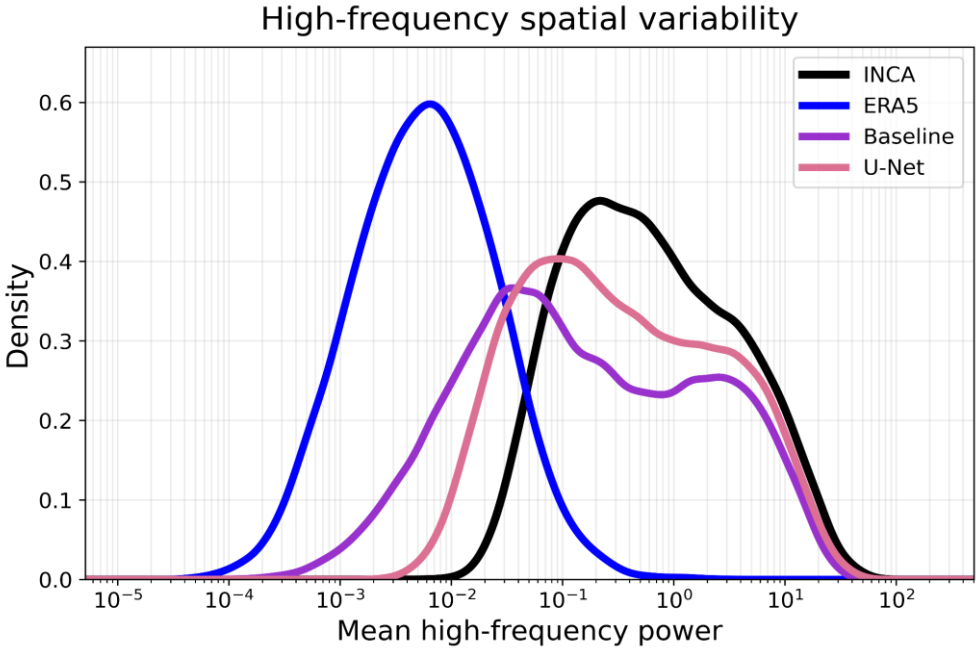

Figure 9: Kernel density estimates of the spatial distribution of time-mean high-frequency temperature power for INCA (black), interpolated ERA5 (blue), the residual U-Net (pink), and the lapse rate baseline (purple) over the 2023–2025 evaluation period. High-frequency power is computed from the high-pass temperature component after subtracting a Gaussian-smoothed field with σ=8 grid cells, corresponding to approximately 8 km on the INCA grid. Larger values indicate stronger small-scale spatial variability.

4.3.4 Topographic consistency evaluation

The ridge-valley contrast diagnostic in Fig. 10 shows us whether the models reproduce terrain mediated temperature differences that are especially important in complex Alpine terrain. During daytime (Fig. 10a,c), both downscaling approaches improve upon ERA5 and reproduce the seasonal evolution of the INCA valley-ridge contrast skillfully. This suggests that both the baseline and the U-Net capture part of the daytime thermal contrast between lower valleys and higher terrain.

Under nocturnal conditions (Fig 10b,d), when stable boundary layers and cold air pooling are more likely to structure the surface temperature field, the baseline performs much less skillfully. In particular, it tends to overestimate the magnitude of the valley-ridge contrast through much of the year. The U-Net remains closer to the INCA seasonal cycle, indicating that it better captures terrain dependent nocturnal thermal structure. This result is consistent with the model design: the U-Net is not limited to a fixed monthly lapse rate adjustment but can use dynamic atmospheric predictors and local terrain descriptors to produce spatially and temporally varying residual corrections.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

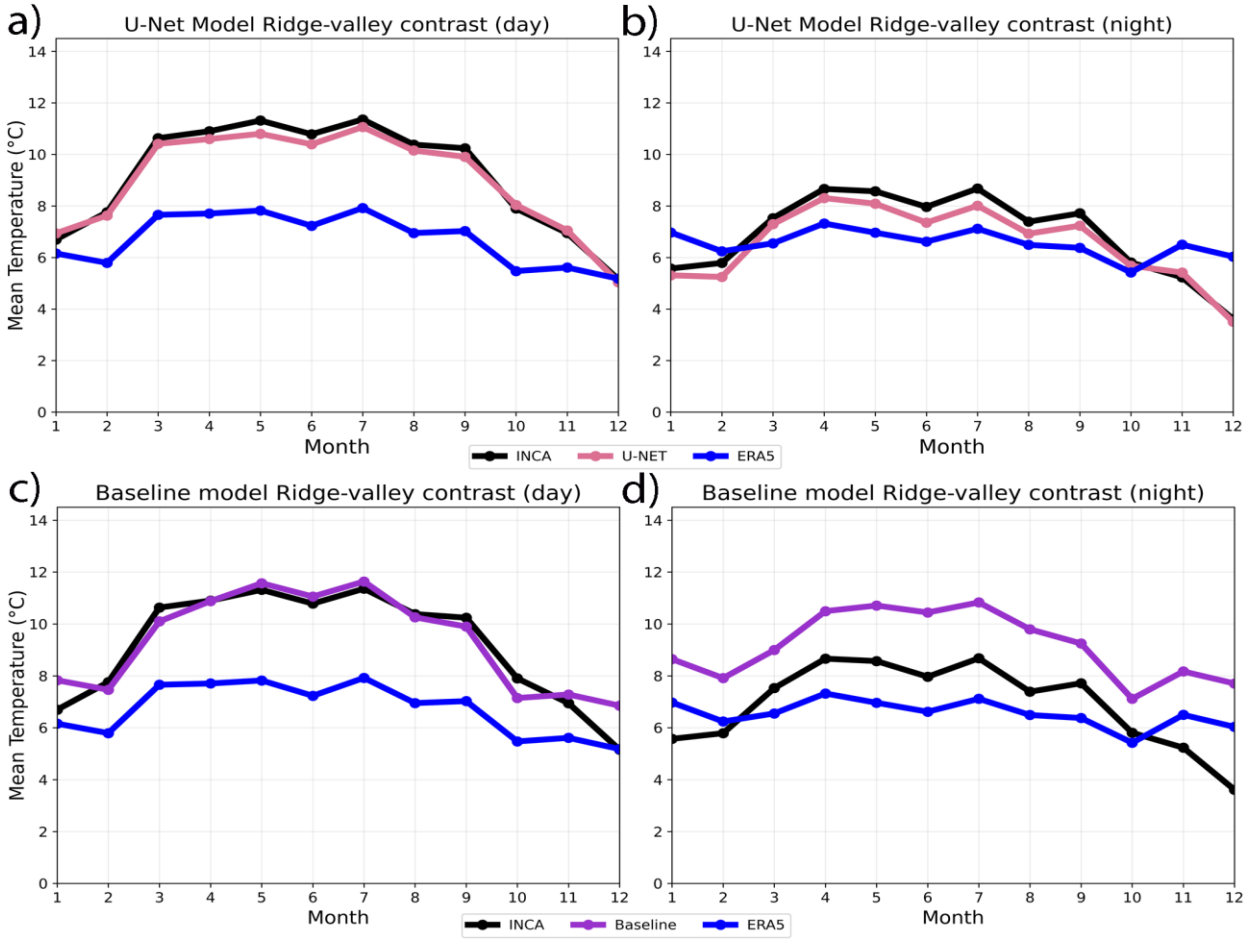

Figure 10: Monthly mean ridge-valley temperature contrast for the (a,b) U-Net and (c,d) baseline model, shown separately for (left) daytime and (right) nighttime. The curves show the seasonal evolution of valley-ridge thermal differences.

4.4 Extreme event evaluation

To evaluate model performance during extreme temperature conditions, a short event-based analysis was performed using multi-day heat-wave and cold-spell events identified from the INCA daily domain-mean 2 m temperature series. Heat events are defined from the upper-tail threshold (95th percentile) and cold events from the lower-tail (5th percentile) threshold. This procedure identifies 10 heat-event days and 11 cold-event days during the 2023–2025 test period. The same INCA defined event windows are then applied to ERA5, the baseline, and the U-Net, ensuring that all models are evaluated under identical extreme temperature events.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

The composite error maps in Fig. 11 reveal that the U-Net produces smaller and less spatially organised event-scale errors than the baseline. During heat events, Fig. 11b shows the baseline retains pronounced regional and terrain-locked error structures, including larger residual biases across Alpine terrain and adjacent lowland areas. In Fig. 11a, the U-Net reduces the magnitude and spatial coherence of these errors, leaving weaker and more localized residuals. During cold events, the contrast is even clearer; in Fig. 11d the baseline shows stronger topographically organised biases, while in Fig. 11c the U-Net more effectively suppresses these patterns. These results indicate that the U-Net improves not only the overall skill, but also the spatial behaviour of errors during extreme hot and cold periods. The remaining U-Net residuals show that event intensity is not reproduced perfectly, but the learned residual correction substantially reduces the terrain locked bias structures that persist in the lapse rate baseline.

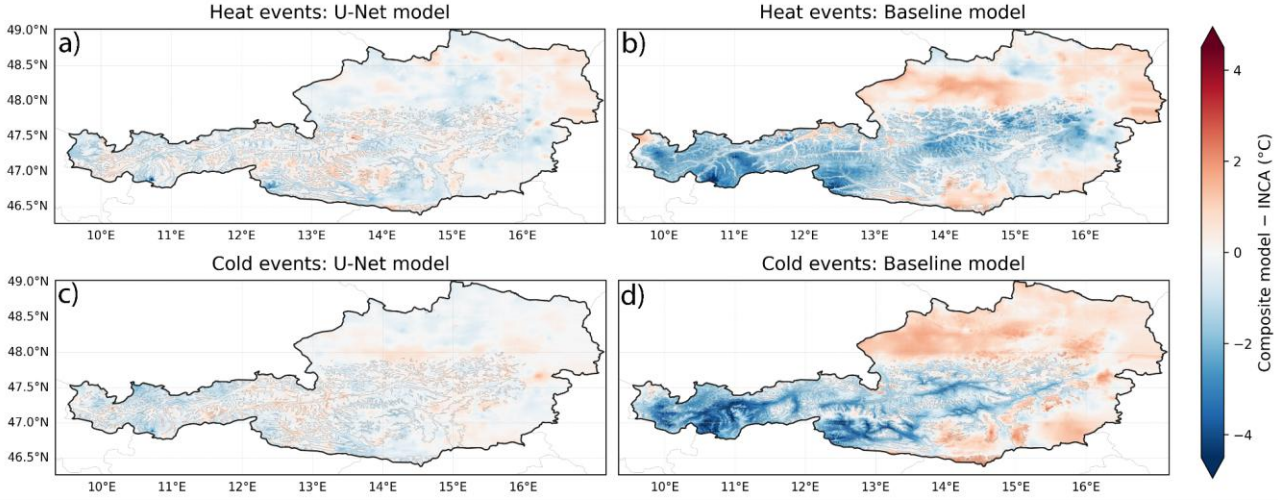

Figure 11: Composite spatial fields of 2 m temperature error relative to INCA during identified heat and cold events. Panels show mean model minus INCA error for the (left) residual U-Net and (right) lapse rate baseline during (a,b) heat events and (c,d) cold events. Positive values indicate a warm bias and negative values a cold bias.

4.5 MERRA-2 source transfer experiment

The MERRA-2 sensitivity experiment tests whether the residual U-Net remains useful when the driving reanalysis is changed without retraining. The resulting comparison therefore evaluates source transfer rather than a newly optimized MERRA-2-specific downscaling model.

The scalar metrics in Fig. 12a–d show that the learned residual correction transfers partially across reanalysis products. MERRA-2 interpolated to the INCA grid without statistical downscaling has higher errors than the downscaled ERA5-based products. Applying the ERA5-trained U-Net to the MERRA-2 predictors improves upon interpolated MERRA-2, reducing

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

RMSE by approximately 0.79 °C (Fig. 12a) and MAE by approximately 0.59 °C (Fig. 12b). The MERRA-2-driven U-Net also increases R^2 relative to interpolated MERRA-2 (Fig. 12d), indicating that the learned correction adds useful skill even when the input reanalysis source is changed. However, the MERRA-2-driven U-Net remains less skillful than the standard ERA5-driven U-Net, which achieves lower RMSE and MAE and higher R^2 (Fig. 12a,b,d). It is also slightly weaker than the lapse rate baseline in the domain-wide metrics, particularly for MAE (Fig. 12b). These results indicate a clear source-shift penalty: the residual correction learned from ERA5 remains beneficial when applied to MERRA-2, but it transfers with reduced skill because the inference predictors come from a different reanalysis distribution.

The spatial RMSE and bias maps in Figs. S3–S6 shows that this source-transfer penalty is also spatially structured. Interpolated MERRA-2 has large RMSE and strong terrain-linked warm and cool bias patterns across the Alpine region, similar in structure to the errors seen for interpolated ERA5 but generally larger in magnitude (Figs. S3a and S4a). The MERRA-2-driven U-Net reduces both RMSE and coherent bias patterns over much of the domain (Figs. S3b and S4b), but it remains less spatially neutral than the ERA5-driven U-Net, which shows the weakest RMSE and bias structure overall (Figs. S3c and S4c). Seasonal maps show the same pattern across DJF, MAM, JJA, and SON: the MERRA-2-driven U-Net improves upon interpolated MERRA-2 in every season, but residual terrain-organized errors remain larger than in the ERA5-driven U-Net (Figs. S5 and S6).

Figures S7–S10 show the corresponding MERRA-2 source-transfer diagnostics for lapse rate agreement, elevation-binned diurnal cycles, high-frequency spatial variability, and ridge-valley contrast and support the same interpretation. The MERRA-2 driven U-Net improves on raw MERRA-2 in several physically relevant aspects, including elevation dependent diurnal structure, high-frequency spatial variability, and ridge-valley contrast. The hourly lapse rate in Fig. S7 shows a compressed response relative to INCA, so the MERRA-2 driven U-Net captures the broad terrain–temperature relationship but underrepresents the full range of hourly lapse rate variability. Diurnal-cycle diagnostics in Fig. S8 reveals that the MERRA-2-driven U-Net corrects much of the raw MERRA-2 high-elevation bias, while the high frequency power KDE in Fig. S9 shifts closer to INCA-like spatial variability. Ridge-valley contrasts in Fig. S10a–b are also improved relative to raw MERRA-2, especially seasonally, but remaining departures indicate that the source-transfer model does not reproduce terrain dependent thermal structure as skillfully. Overall, the sensitivity experiment reveals that the MERRA-2 driven U-Net has learned transferable downscaling structure, but also that the strongest performance requires consistency between the training and inference reanalysis sources.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

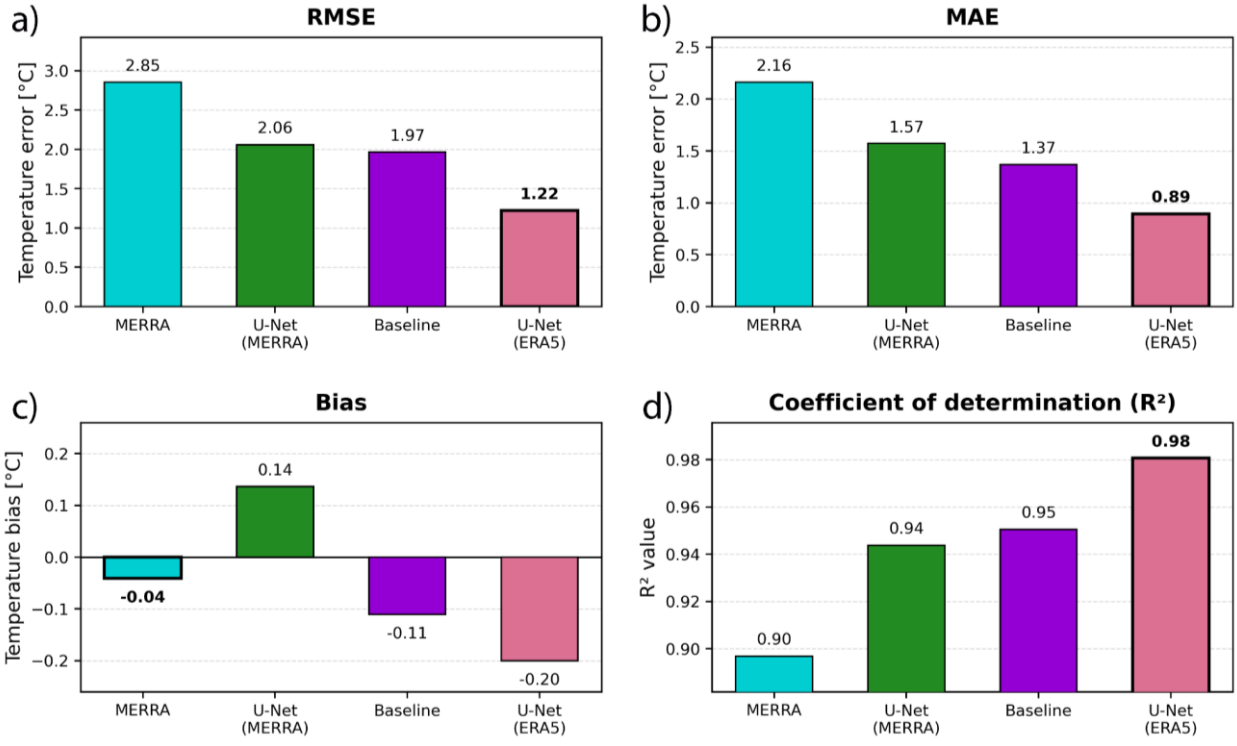

Figure 12: MERRA-2 source-transfer skill metrics against INCA for 2023–2025: (a) RMSE, (b) MAE, (c) mean bias, and (d) R^2. Metrics are shown for MERRA-2 interpolated to the INCA grid without statistical downscaling (cyan), the ERA5-trained residual U-Net driven by MERRA-2 predictors (green), the lapse rate baseline (purple), and the standard ERA5-driven residual U-Net (pink). All metrics are computed using the same valid time-pixel sample. RMSE and MAE measure absolute error magnitude, bias is defined as model minus INCA, and R^2 summarizes agreement with the INCA reference field. The best model per metric is indicated with a bold black frame.

5 Discussion

This study demonstrates that residual deep-learning downscaling can substantially improve hourly 2 m temperature fields over complex Alpine terrain. The residual U-Net reduced domain-wide errors relative to both interpolated ERA5 and the baseline, while also producing more realistic terrain-dependent structure. This is important because improved scalar error metrics alone do not guarantee physically plausible downscaled fields. Here, the U-Net showed consistent skill across anomaly agreement, spatial error structure, lapse rate variability, elevation-dependent diurnal cycles, high-frequency spatial variability, ridge-valley contrasts, and extreme-event conditions. The results therefore suggest that the model is not only correcting mean errors, but is also learning spatially and temporally varying temperature structure linked to topography and atmospheric state.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

The lapse rate baseline provides a strong physically interpretable benchmark. Its improvements over ERA5 are consistent with other downscaling complex terrain studies showing that near-surface temperature variability in complex terrain is strongly governed by elevation and that elevation- or lapse rate-based corrections can improve coarse gridded temperature fields (Daly et al., 2008; Minder et al., 2010; Gao et al., 2017; Dutra et al., 2020). In this study, the baseline's gain over ERA5 therefore confirms that a substantial part of the ERA5-INCA mismatch is associated with unresolved orography. However, its limitations are also clear: The baseline compresses hourly lapse rate variability (Fig. 5b), shows larger high-elevation diurnal-cycle errors (Fig. 6b), remains too smooth in high-frequency spatial variability (Fig. 7), and overestimates nocturnal ridge-valley contrast during much of the year (Fig. 8d).

The residual U-Net performs better because it is not restricted to a fixed lapse rate correction. By learning residuals relative to the coarse reanalysis field, the model can preserve the large-scale meteorological signal while adding local corrections informed by terrain, elevation derivatives, time of day, season, and dynamic atmospheric predictors. This design is especially effective in mountainous terrain, where temperature errors are rarely spatially uniform. The U-Net's closer agreement with INCA in the lapse rate, diurnal cycle, high-frequency spatial variability, and ridge-valley diagnostics indicates that it learns a more flexible representation of sub-grid thermal structure than the baseline. The spatial RMSE and bias maps reinforce this interpretation. ERA5 and the baseline retain terrain-organized error patterns, whereas the U-Net produces lower RMSE and weaker coherent bias structures. Seasonal analysis shows that this improvement persists throughout the year, with the clearest gains in complex mountainous terrain where unresolved topographic effects are strongest.

The extreme-event composites further show the value of the learned residual correction. During both warm and cold events, the baseline retains larger and more spatially organized terrain-locked errors, whereas the U-Net reduces both the magnitude and coherence of these residual patterns. This suggests that the model improves not only average conditions but also the spatial behavior of errors during meteorologically important periods. Some residual errors remain, particularly in event intensity and high-frequency spatial variability, indicating that the U-Net does not perfectly reproduce the INCA reference. Nevertheless, its errors are weaker and less systematically organized than those of the baseline.

The MERRA-2 source-transfer experiment provides an important robustness test and can be interpreted as a predictor-source shift problem. The U-Net is trained and normalized using ERA5 predictors but is then evaluated using MERRA-2 predictors mapped onto the same input interface. This mapping includes both horizontal interpolation to the INCA grid and vertical harmonization of the jet-level zonal-wind predictor; the model-facing u300 channel is derived by log-pressure interpolation between the available MERRA-2 u250 and u500 fields and is retained under the name u300 only to match the ERA5-trained input schema. This situation is closely related to covariate shift, where the input distribution changes between training and inference while the prediction task remains nominally the same (Shimodaira, 2000; Sugiyama et al., 2007; Quionero-Candela et al., 2009). The reduced skill of the MERRA-driven U-Net in Fig. 10a,b,d therefore reflects a source-distribution penalty

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

rather than a failure of the residual downscaling concept. The model has learned useful downscaling structure, but this result highlights the importance of predictor source compatibility when applying trained neural downscaling models beyond their original training dataset.

6 Conclusions

This study developed and evaluated a deep-learning downscaling model based on a residual U-Net framework for hourly 2 m temperature downscaling over Austria and the surrounding Alpine region, using INCA as the high-resolution reference analysis. The results were compared to a physically interpretable lapse rate baseline model. Over the independent 2023–2025 evaluation period, the ERA5-driven residual U-Net achieved the strongest agreement with INCA, reducing RMSE and MAE relative to both interpolated ERA5 and the lapse rate baseline model. Beyond these scalar improvements, the U-Net more faithfully reproduced key aspects of INCA Alpine temperature behaviour, including spatial distribution, hourly lapse rate variability, elevation-dependent diurnal cycles, high-frequency spatial variability, ridge-valley contrasts, and event-scale error patterns during heat and cold events.

The lapse rate baseline remains valuable because it improves over ERA5 and provides a transparent benchmark for terrain-based correction. However, its fixed monthly structure limits its ability to represent rapidly changing thermal regimes in complex terrain. The residual U-Net overcomes many of these limitations by learning dynamic, terrain aware corrections to the coarse reanalysis field.

The MERRA-2 experiment shows that the learned correction has some transferable value, since the MERRA-2 driven residual U-Net improves upon raw MERRA-2 reanalysis. However, the reduced skill relative to the ERA5-driven residual U-Net demonstrates that source consistency matters, and the trained model is not completely robust against changes in the input reanalysis. Overall, the results support residual U-Net downscaling as a powerful approach for producing physically realistic high-resolution temperature fields in Alpine terrain.

In this study, INCA is treated as the high-resolution reference against which both the lapse rate baseline and the residual U-Net are trained or evaluated. Therefore, the resulting downscaled products should be interpreted as high-resolution emulations of the INCA temperature analysis rather than direct observations of the true near-surface temperature field. Although INCA has much higher spatial resolution than ERA5 and resolves much of the Alpine near-surface temperature structure, it is still a data assimilation and analysis product with associated uncertainties. Future work should therefore compare the downscaled ERA5 fields against independent weather station observations to assess performance relative to measurements that were not used as the gridded target.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix A

Group	ERA5 short name	Physical quantity	Atmospheric/ land-surface level	Role in the downscaling dataset
Surface atmosphere	t2m	2 m air temperature	2 m	Coarse surface temperature field and residual reference
Surface atmosphere	u10, v10	Zonal and meridional wind	10 m	Surface wind and advection-related information
Land surface	swvl1	Volumetric soil water	Upper soil layer	Land-surface moisture state
Surface atmosphere	sp	Surface pressure	Surface	Surface pressure and synoptic state
Hydrometeorological forcing	precip	Precipitation accumulation/rate	Surface	Precipitation-related forcing and cloud/radiative proxy
Free atmosphere	z500	Geopotential height	500 hPa	Large-scale geopotential height/synoptic structure
Free atmosphere	t850	Air temperature	850 hPa	Lower-tropospheric temperature and lapse rate information
Free atmosphere	u300	Zonal wind	300 hPa	Upper-level wind information

Table A1: ERA5 predictor variables used in the downscaling dataset. Variables are grouped by physical category and listed with their ERA5 short name, physical quantity, atmospheric or land-surface level, and role in the downscaling workflow. All dynamic predictors are bilinearly interpolated to the INCA grid before model training, inference, and evaluation.

Code and data availability

The U-Net architecture and downscaling Python scripts used in this paper are publicly available at https://doi.org/10.5281/zenodo.21326767 (Ennis, 2026). INCA data are publicly available from Geosphere Austria at https://data.hub.geosphere.at/en/dataset/inca-v1-1h-1km. ERA5 data are publicly available at the ECMWF Climate Data Store at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=download. MERRA-2 data are publicly available at https://gmao.gsfc.nasa.gov/gmao-products/merra-2/data-access_merra-2/.

Supplement link

The link to the supplement will be included by Copernicus, if applicable.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

Author contributions

The study was conceptualized, and the methodology was designed by KEE and SS; KEE performed the data curation, investigation, formal analysis, visualization, validation, and writing (original draft preparation). SS was responsible for project administration, supervision, and writing (review and editing).

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

The authors gratefully acknowledge the Know Center Research GmbH for its partnership and support throughout this project. We thank Dr. Maximilian Troller for providing feedback on the paper. We thank ECMWF and NASA for making ERA5 and MERRA-2 freely available, and Geosphere Austria for doing the same with INCA data. We also are grateful to Dr. Shawn Milrad of Embry-Riddle Aeronautical University for his careful editing and constructive feedback on the manuscript. Finally, we acknowledge Colorado State University for providing the computing infrastructure that supported the model training, inference, and analysis presented in this study.

Financial support

This work is part of the project “RegDTAlp - Towards a digital twin of mountainous weather and climate for impact assessment in the Austrian Alps”, funded by the Climate and Energy Fund (Klima- und Energiefonds) and Austrian Research Promotion Agency (FFG) under the Austrian Climate Research Program (ACRP), grant Nr. FO999924969/60607023. The authors acknowledge the financial support by the University of Graz.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

Review statement

The review statement will be added by Copernicus Publications listing the handling editor as well as all contributing referees according to their status anonymous or identified.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

Ayzel, G., Scheffer, T. and Heistermann, M.: RainNet v1. 0: A convolutional neural network for radar-based precipitation nowcasting. Geosci. Model Dev., 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, 2020.

Bacer, S., Beaumet, J., Ménégoz, M., Gallée, H., Le Bouëdec, E. and Staquet, C.: Impact of climate change on persistent cold-air pools in an alpine valley during the 21st century. Weather Clim. Dyn., 5, 211–229, https://doi.org/10.5194/wcd-5-211-2024, 2024.

Baño-Medina, J., Manzanas, R. and Gutiérrez, J.M.: Configuration and intercomparison of deep learning neural models for statistical downscaling. Geosci. Model Dev., 13, 2109–2124, https://doi.org/10.5194/gmd-13-2109-2020, 2020.

Bhakare, S., Dal Gesso, S., Venturini, M., Zardi, D., Trentini, L., Matiu, M. and Petitta, M.: Intercomparison of machine learning models for spatial downscaling of daily mean temperature in complex terrain. Atmosphere, 15, 1085, https://doi.org/10.3390/atmos15091085, 2024.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J. and Pasteris, P.P.: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol., 28, 2031–2064, https://doi.org/10.1002/joc.1688, 2008.

Dutra, E., Muñoz Sabater, J., Boussetta, S., Komori, T., Hirahara, S., and Balsamo, G.: Environmental lapse rate for high-resolution land surface downscaling: An application to ERA5. Earth Space Sci., 7, e2019EA000984, https://doi.org/10.1029/2019EA000984, 2020.

Ennis, K.: A terrain-aware residual U-Net framework for temperature downscaling over Austria, v1.0, Zenodo [code]. https://doi.org/10.5281/zenodo.21326767, 2026.

Fowler, H.J., Blenkinsop, S. and Tebaldi, C.: Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int. J. Climatol., 27, 1547–1578, https://doi.org/10.1002/joc.1556, 2007.

Fuentes–Franco, R., Krus, K., Ivanov, M., Koenigk, T., Wang, F., and Aldama-Campino, A.: Pan-European high-resolution downscaling using deep learning. J. Geophys. Res. Mach. Learn. Comput., 2025JH000630, https://doi.org/10.1029/2025JH000630, 2025.

Gao, L., Bernhardt, M., Schulz, K. and Chen, X.: Elevation correction of ERA-Interim temperature data in the Tibetan Plateau. Int. J. Climatol., 37, 3540–3552. https://doi.org/10.1002/joc.4935, 2017.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., Randles, C.A., Darmenov, A., Bosilovich, M.G., Reichle, R. and Wargan, K.: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J. Clim., 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Giorgi, F. and Mearns, L.O.: Approaches to the simulation of regional climate change: a review. Rev. Geophys., 29, 191–216, https://doi.org/10.1029/90RG02636, 1991.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B. and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the Eastern Alpine region. Weather Forecast., 26, https://doi.org/10.1175/2010WAF2222451.1, 166–183, 2011.

Harder, P., Hernandez-Garcia, A., Ramesh, V., Yang, Q., Sattegeri, P., Szwarcman, D., Watson, C. and Rolnick, D.: Hard-constrained deep learning for climate downscaling. J. Mach. Learn. Res., 24, 1–40, 2023.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D. and Simmons, A.: The ERA5 global reanalysis. Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hess, P., Aich, M., Pan, B., Boers, N.: Fast, scale-adaptive and uncertainty-aware downscaling of Earth system model fields with generative machine learning. Nat. Mach. Intell. **7**, 363–373, https://doi.org/10.1038/s42256-025-00980-5, 2025.

Huber, P. J.: Robust estimation of a location parameter. Ann. Math. Stat., 35, 73–101, 1964.

Lin, H., Tang, J., Wang, S., Wang, S. and Dong, G.: Deep learning downscaled high-resolution daily near surface meteorological datasets over East Asia. Sci. Data, 10, 890, https://doi.org/10.1038/s41597-023-02805-9, 2023.

Lin, H.W., Tegmark, M. and Rolnick, D.: Why does deep and cheap learning work so well? J. Stat. Phys., 168, 1223–1247, https://doi.org/10.1007/s10955-017-1836-5, 2017.

Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H.W., Sauter, T., Themeßl, M. and Venema, V.K.: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Rev. Geophys., 48, RG3003, https://doi.org/10.1029/2009RG000314, 2010.

Maraun, D. and Widmann, M.: Statistical downscaling and bias correction for climate research, Cambridge University Press, https://doi.org/10.1017/9781107588783, 2018.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.Y., Liu, C.C., Vahdat, A., Nabian, M.A., Ge, T., Subramaniam, A. and Kashinath, K.: Residual corrective diffusion modeling for km-scale atmospheric downscaling. Commun. Earth Environ., 6, 24, https://doi.org/10.1038/s43247-025-02042-5, 2025.

Mathieu, M., Couprie, C. and LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, https://doi.org/10.48550/arXiv.1511.05440, 2015.

McCaffrey, K., Wilczak, J.M., Bianco, L., Grimit, E., Sharp, J., Banta, R., Friedrich, K., Fernando, H.J.S., Krishnamurthy, R., Leo, L.S. and Muradyan, P.: Identification and characterization of persistent cold pool events from temperature and wind profilers in the Columbia River Basin. J. Appl. Meteorol. Climatol., 58, 2533–2551, https://doi.org/10.1175/JAMC-D-19-0046.1,2019.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

McGregor, G.R., Bessemoulin, P., Ebi, K.L. and Menne, B.: Heatwaves and Health: Guidance on Warning-system Development, World Meteorological Organization, Geneva, Switzerland, 2015.

Minder, J.R., Mote, P.W., and Lundquist, J.D.: Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains. J. Geophys. Res. Atmos., 115, D14122, https://doi.org/10.1029/2009JD013493, 2010.

Murray, V., Abrahams, J., Adballah, C., Ahmed, K., Angeles, L., Benouar, D., Brenes, T.A., Hun, C.C., Cox, S., Douris, J. and Fagan, L.: Hazard Information Profiles: Supplement to: UNDRR-ISC Hazard Definition & Classification Review-Technical Report, 2021.

Park, H., Park, S., Kang, D., and Kim, J-H.: A super-resolution framework for downscaling machine learning weather prediction toward 1-km air temperature. npj Clim Atmos Sci, **9**, 56, https://doi.org/10.1038/s41612-026-01328-5, 2026.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: an imperative style, high-performance deep learning library, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 721, pp. 8026–8037, Vancouver, Canada, 8–14 December 2019, https://dl.acm.org/doi/10.5555/3454287.3455008 (last access: 5 November 2025), 2019.

Prein, A.F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O., Feser, F. and Brisson, E.: A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. Rev. Geophys., 53, 323–361, https://doi.org/10.1002/2014RG000475, 2015.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N.D.: 365 Dataset Shift in Machine Learning, MIT Press, 2009.

Ronneberger, O., Fischer, P. and Brox, T.: October. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, Berlin, Heidelberg, Germany, 234–241, 2015.

Rummukainen, M.: Added value in regional climate modeling. WIREs Clim. Change, 7, 145–159, https://doi.org/10.1002/wcc.378, 2016.

Schmid, F., Schmidli, J., Hervo, M. and Haefele, A.: Diurnal valley winds in a deep alpine valley: observations. Atmosphere, 11, 54, https://doi.org/10.3390/atmos11010054, 2020.

Sha, Y., Gagne II, D.J., West, G. and Stull, R.: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. J. Appl. Meteorol. Climatol., 59, 2075–2092, https://doi.org/10.1175/JAMC-D-20-0058.1, 2020.

Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference, 227–244, https://doi.org/10.1016/S0378-3758(00)00115-4, 2000.

Shrivastava, A., Gupta, A. and Girshick, R.: Training region-based object detectors with online hard example mining., in: Proceedings of the IEEE conference on computer vision and pattern recognition, 761–769, 2016.

https://doi.org/10.5194/egusphere-2026-4059
Preprint. Discussion started: 21 July 2026
© Author(s) 2026. CC BY 4.0 License.

Sugiyama, M., Krauledat, M. and Müller, K.R.: Covariate shift adaptation by importance weighted cross validation. J. Mach. Learn. Res., 8, 985–1005, 2007.

Trebing, K., Stańczyk, T. and Mehrkanoon, S.: SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. Pattern Recognit. Lett., 145, 178–186, https://doi.org/10.1016/j.patrec.2021.01.036, 2021.

Vandal, T., Kodra, E. and Ganguly, A.R.: Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. arXiv preprint arXiv:1702.04018, https://doi.org/10.48550/arXiv.1702.04018, 2017.

Vosper, S.B. and Brown, A.R.: Numerical simulations of sheltering in valleys: The formation of nighttime cold-air pools. Bound.-Layer Meteorol., 127, 429–448, https://doi.org/10.1007/s10546-008-9272-3, 2008.

Wilby, R.L. and Wigley, T.M.: Downscaling general circulation model output: a review of methods and limitations. Prog. Phys. Geogr., 21, 530–548, https://doi.org/10.1177/030913339702100403, 1997.