A deep learning framework for gridding daily climate variables from a sparse station network

Alexandru Dumitrescu[1]

[1]Department of Climatology, Meteo Romania (National Meteorological Administration), Bucharest, 013686, Romania

Correspondence to: Alexandru Dumitrescu (dumitrescu@meteoromania.ro)

Abstract. High-resolution gridded climate datasets are essential for Earth system modelling and impact assessments, yet generating them from sparse, irregularly distributed station networks remains a significant challenge, particularly in regions with complex topography. This study evaluates the Spatial Multi-Attention Conditional Neural Process (SMACNP), a probabilistic deep learning framework, for the daily spatial interpolation of air temperature and precipitation, marking the first application of its localized encoder variant to the challenge of gridding climate data from a sparse station network. We investigate two distinct encoder configurations—Global and Localized—to determine the optimal structural prior for capturing spatial dependencies in data-scarce regimes. The models were developed and evaluated using data from a sparse network of meteorological stations in Romania from 2020 to 2023. To ensure applicability for long-term historical reconstruction, the input features were restricted to static topographic predictors derived from a Digital Elevation Model (DEM). Performance was benchmarked against Regression Kriging (RK), a standard geostatistical baseline that incorporates these same topographic covariates. Results demonstrate that the SMACNP architectures substantially outperform the RK baseline for both variables. The SMACNP (Localized) configuration, which utilizes an attention mechanism, emerged as the most robust model, achieving the lowest Mean Absolute Error (MAE) and the highest correlation across the majority of seasons. The performance gains were particularly pronounced for precipitation, where the deep learning models effectively captured fine-scale spatial heterogeneity and non-linearities that traditional methods tended to over-smooth. Furthermore, the SMACNP framework demonstrated superior uncertainty quantification; while RK exhibited significant overconfidence in precipitation estimates, the SMACNP (Localized) model produced well-calibrated probabilistic predictions with near-ideal empirical coverage. These findings indicate that localized neural process-based models offer a powerful, scalable, and physically plausible alternative to geostatistical methods for generating high-quality gridded climate datasets in complex, data-sparse environments.

1 Introduction

Daily gridded climate datasets are essential inputs for a wide range of models and impact assessments. Producing these fields from sparse, irregularly distributed station networks remains challenging, especially over complex terrain (Hijmans et al.,

2005). Air temperature tends to vary smoothly with controls such as elevation and latitude, whereas precipitation is intermittent, highly skewed, and spatially and temporally discontinuous (Daly et al., 1994). This problem is particularly critical in areas characterized by complex topography, where local factors like elevation and wind exposure introduce nonstationary relationships and heterogeneous spatial dependencies that challenge traditional geostatistical assumptions (Diggle and Ribeiro, 2007). Classical interpolators such as inverse distance weighting and kriging—including regression kriging (RK)—are widely used baselines (Hengl et al., 2007; Li and Heap, 2014). Their performance can degrade where covariance structure is nonstationary or relationships with covariates are nonlinear, conditions common in topographic-complex regions. While machine-learning approaches, particularly tree ensembles such as random forest and gradient boosting, often improve accuracy by leveraging rich covariates (e.g., satellite products, elevation, and reanalysis fields), they can still be limited in modeling the most complex spatial dependencies (Appelhans et al., 2015; Iwase and Takenawa, 2024; Sekulić et al., 2021).

Deep learning (DL) provides a powerful alternative, offering flexible, data-driven representations capable of modelling the complex spatial and temporal non-linear relationships inherent in climate data (Reichstein et al., 2019). Conditional Neural Processes (CNPs), for instance, learn distributions over functions from context points while providing uncertainty estimates (Garnelo et al., 2018). The convolutional variant, the Convolutional Conditional Neural Process (ConvCNP), uses translation-equivariant, continuous convolutions to capture local spatial dependencies and is well suited to gridded queries (Gordon et al., 2019; Vaughan et al., 2022). Attentive Neural Processes build on this by introducing cross-attention from target queries to context points, improving the handling of heterogeneity among observations (Kim et al., 2019). In parallel, Transformers—originating in natural language processing and later adapted to vision—provide a direct solution for leveraging rich covariate information from irregularly spaced stations. They use attention so each element can weight all others when forming its representation (Dosovitskiy et al., 2020; Vaswani et al., 2017). In spatial problems, this means that when geographic coordinates and relevant covariates (e.g., elevation) are encoded, a query at a target location can up-weight both nearby and distant stations that are most informative (e.g., those at similar elevation), directly addressing a key weakness of traditional distance-based interpolators. The Spatial Multi Attention Conditional Neural Process (SMACNP) exemplifies this approach by combining cross-attention from target queries to context stations with specialized spatial and feature-wise attention to capture distance-dependent structure and predictor correlations for interpolating point measurements (Bao et al., 2024b). Building on this, Gridded Transformer Neural Processes (GriddedTNP) were recently introduced to tackle the scalability issues of Transformer Neural Processes (TNPs) for large, unstructured spatio-temporal datasets like weather data, by employing specialized gridded pseudo-tokens for efficient attention (Ashman et al., 2024). Similarly, Graph Neural Processes (GNPs) combine Graph Neural Networks (GNNs) with Neural Processes (NPs) to explicitly model relationships and dependencies defined by a graph structure (Carr and Wingate, 2019). Taking a different structural approach, (Bao et al., 2024a) introduced a two-stage framework, the Location Embedded Graph Neural Networks–Residual Neural Processes (LEGNN-RNP), in which a Graph Neural Network is enhanced by a self-attention-

based location embedding generates an initial prediction, and a subsequent Residual Neural Process models the spatial patterns in the resulting errors, refining the final prediction and quantifying uncertainty in a manner similar to RK.

Despite these architectural advancements, the application of such probabilistic deep learning frameworks to extremely sparse observation networks in regions with complex topography remains under-explored. Our preliminary experiments in this specific context revealed that alternative architectures, such as the ConvCNP and GriddedTNP, struggled to generalize effectively when relying solely on static topographic predictors. We present a case study evaluating the SMACNP deep learning architecture for gridding daily temperature and precipitation over Romania from a sparse national station network (156 stations). Romania provides a challenging testbed combining complex Carpathian topography, sparse and irregular station coverage, and mixed precipitation regimes. While the methodology is general and transferable to other regions with similar data availability, the evaluation presented here is specific to Romania and should be interpreted accordingly. To the best of our knowledge, this represents the first application of the SMACNP architecture, specifically its localized encoder variant, to the challenge of gridding daily climate data from a sparse observation network. Performance is compared with Regression Kriging (RK), which serves as the geostatistical baseline, for the period 2020–2023. The contributions of this study are: (i) a systematic evaluation of the SMACNP architecture for daily climate gridding over complex terrain, demonstrating its advantage over RK for both temperature and precipitation; (ii) a comparison of global and localized attention configurations, showing that restricting attention to spatial neighbours improves interpolation from sparse, irregular networks; and (iii) a comprehensive uncertainty assessment including probabilistic calibration, precipitation occurrence skill, and climate-relevant extreme indices.

~~Consequently, in this study, we evaluate SMACNP configured with global and localized encoders for the daily spatial interpolation of very sparse station-based climate variables. To the best of our knowledge, this represents the first application of the SMACNP architecture – specifically its localized encoder variant – to the challenge of gridding daily climate data from sparse observation networks. Their performance is compared with the RK method, which serves as the baseline, for the period 2020–2023. Particular emphasis is placed on interpolation accuracy and its associated uncertainty.~~ The remainder of this paper is structured as follows. Section 2 describes the datasets employed and the preprocessing procedures. Section 3 outlines the methodological framework, including a detailed description of RK and SMACNP, and presents the experimental setup. Section 4 reports the results, with a focus on interpolation accuracy and predictive uncertainty. Finally, Section 5 and 6 discuss and summarize the main findings and provide concluding remarks. Our focus is on spatially aware probabilistic DL models (SMACNP and its variants) that natively handle irregular station sets and provide calibrated uncertainties. We therefore use RK as the geostatistical reference and do not benchmark tree-ensemble methods here.

2 Study area and data

The study is conducted over Romania, a region characterized by complex topography including the Carpathian Mountains, plateaus, and plains. We use daily meteorological station data and a topographic variable extracted from high-resolution Digital Elevation Model (DEM) to develop and evaluate the interpolation models

2.1 Meteorological and topographic datasets

The primary dataset consists of daily homogenized mean air temperature (°C) and total precipitation (mm) from a network of meteorological stations across Romania, spanning 2020–2023. The complete dataset and associated station metadata can be accessed through the Zenodo repository (Dumitrescu, 2025). While the experiments in this study are conducted on this recent period, the proposed method is ultimately intended for application to the full long-term dataset covering 1901 to the present (Dumitrescu et al., 2025). The station data is pre-partitioned into a training set (130 stations), used for model development and validation, and a held-out test set (26 stations), used exclusively for final performance evaluation for all the three models selected in the study (Figure 1~~Figure 1~~). To ensure strict evaluation integrity, the 26 test stations were selected randomly once before any model development and remained fixed throughout the study. These stations are never included in the context or target sets during training, and play no role in hyperparameter optimization or model selection.

Topographic and geographic predictors are derived from a high-resolution DEM (Jarvis et al., 2008). These predictors serve as covariates in the RK model and as static input features for the DL models. We restricted our input predictors to those derived from topographic data because the primary objective of this study is to identify an interpolation model configuration suitable for application to long-term daily climate time series (1901–present). For this extended historical period, dynamic covariates such as satellite products and land cover datasets are unavailable or inconsistent. Limiting the input to static, topographically derived predictors ensures temporal consistency and model applicability across the full time span.

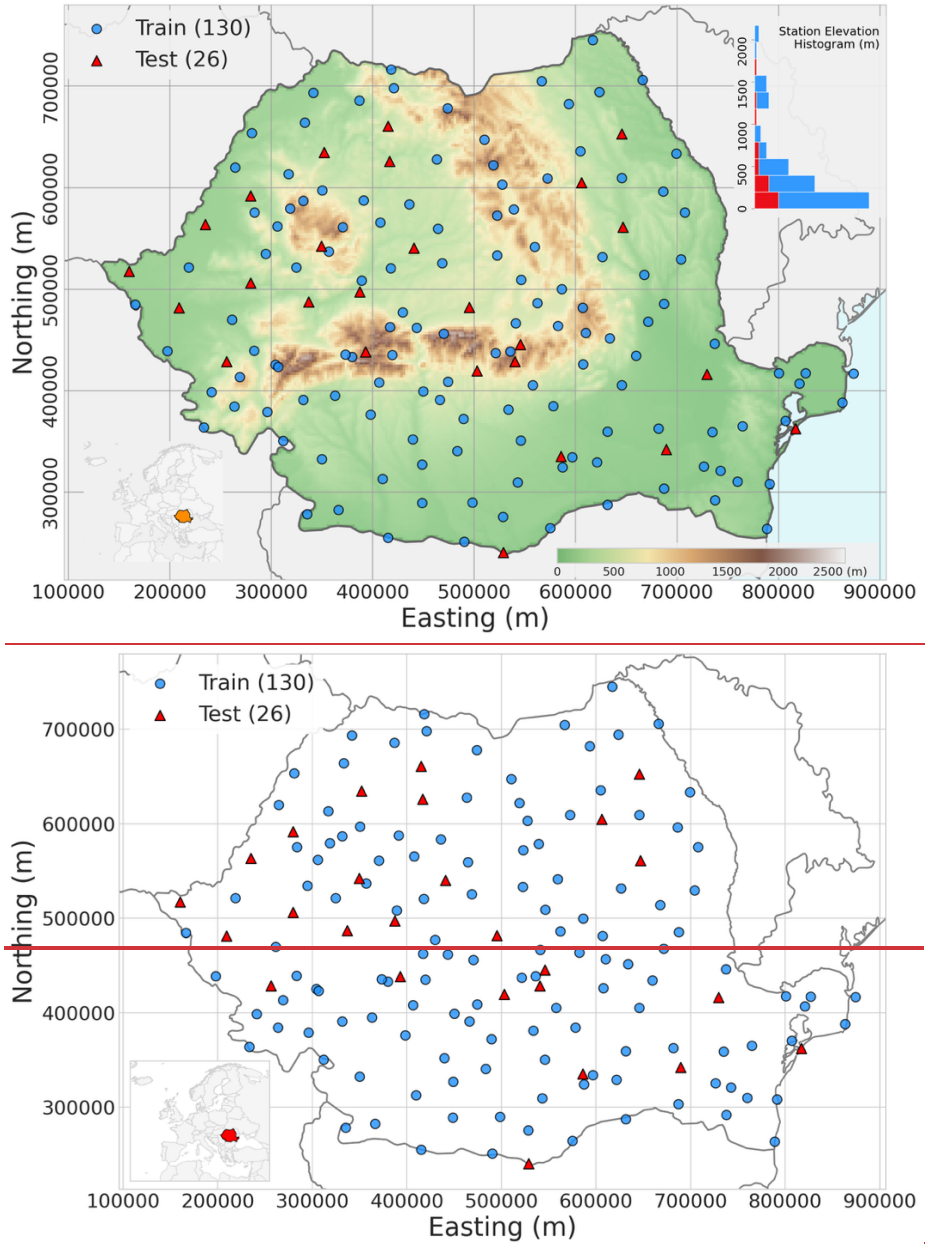

Figure 1 ~~Spatial distribution of the 156 meteorological stations used in this study, overlaid on the SRTM digital elevation model of Romania. Training stations ($n = 130$, blue circles) and held-out test stations ($n = 26$, red triangles) were assigned by random spatial split. The inset histogram shows the elevation distribution of training and test stations, confirming that both subsets span the full altitudinal range of the network (0–2,544 m). Coordinates are in the Dealul Piscului 1970 projection (EPSG:31700). Inset map shows Romania's location within Europe.~~

~~Spatial distribution of training and testing weather stations.~~

2.2 Spatio-temporal covariates

A set of predictor variables was engineered to capture the primary drivers of spatial and temporal climate variability. These are grouped into static spatial predictors and dynamic temporal predictors.

Derived directly from the DEM for every station and grid point location, spatial predictors include smoothed elevation (*smooth_elev*), a topographic position index (*topo_position*), distance to the Black Sea (*dist_to_coast*), and geographic coordinates (*latwgs*, *lonwgs*) (Figure 2~~Figure 2~~) (Daly et al., 2008). Calculated for each day to capture seasonal and long-term patterns, temporal predictors represent cyclical patterns without discontinuities, the day of the year and the month are transformed into sine and cosine components (*day_sin*, *day_cos*, *month_sin*, *month_cos*). A *year_scaled* variable, representing the year linearly scaled to the [0, 1] range over the analysis period, is included to account for inter-annual trends.

These predictors are concatenated to form a feature vector for each observation point (*context*) and prediction grid point (*target*).

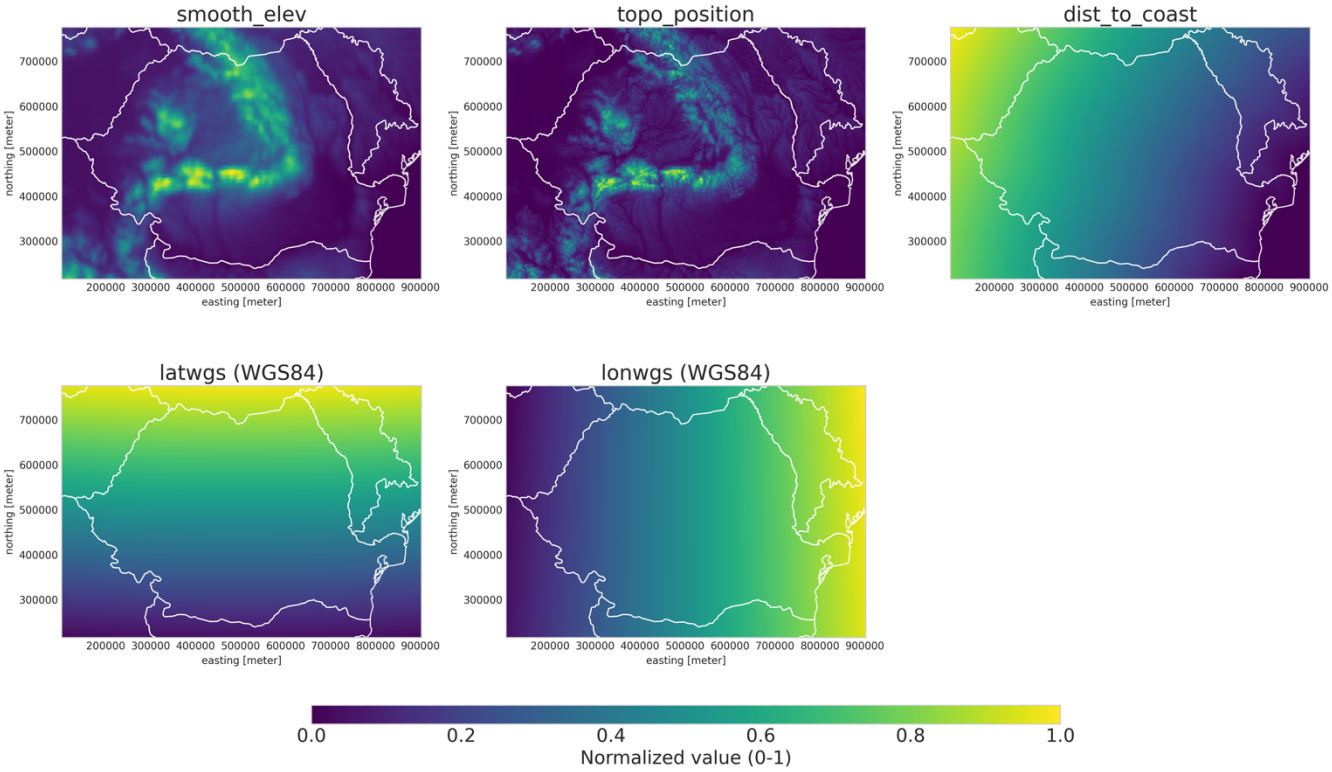

Figure 2 Static predictors derived from DEM.

To assess whether the station network adequately represents the prediction domain, Figure 3 compares the distributions of the three key static predictors — smoothed elevation, topographic position index, and distance to the Black Sea coast —

across training stations, test stations, and the prediction grid (restricted to cells within Romania's national territory). The training and test station distributions are closely aligned across all predictors, confirming that the random 130/26 split is representative. Crucially, the prediction grid falls almost entirely within the training data envelope: only 0.4% of grid cells exceed the maximum training-station elevation, 0.1% exceed the maximum topographic position, and 0.5% exceed the maximum distance to coast. These results indicate that the model operates predominantly within its training distribution during grid-scale prediction, with extrapolation limited to a small number of extreme-topography cells.

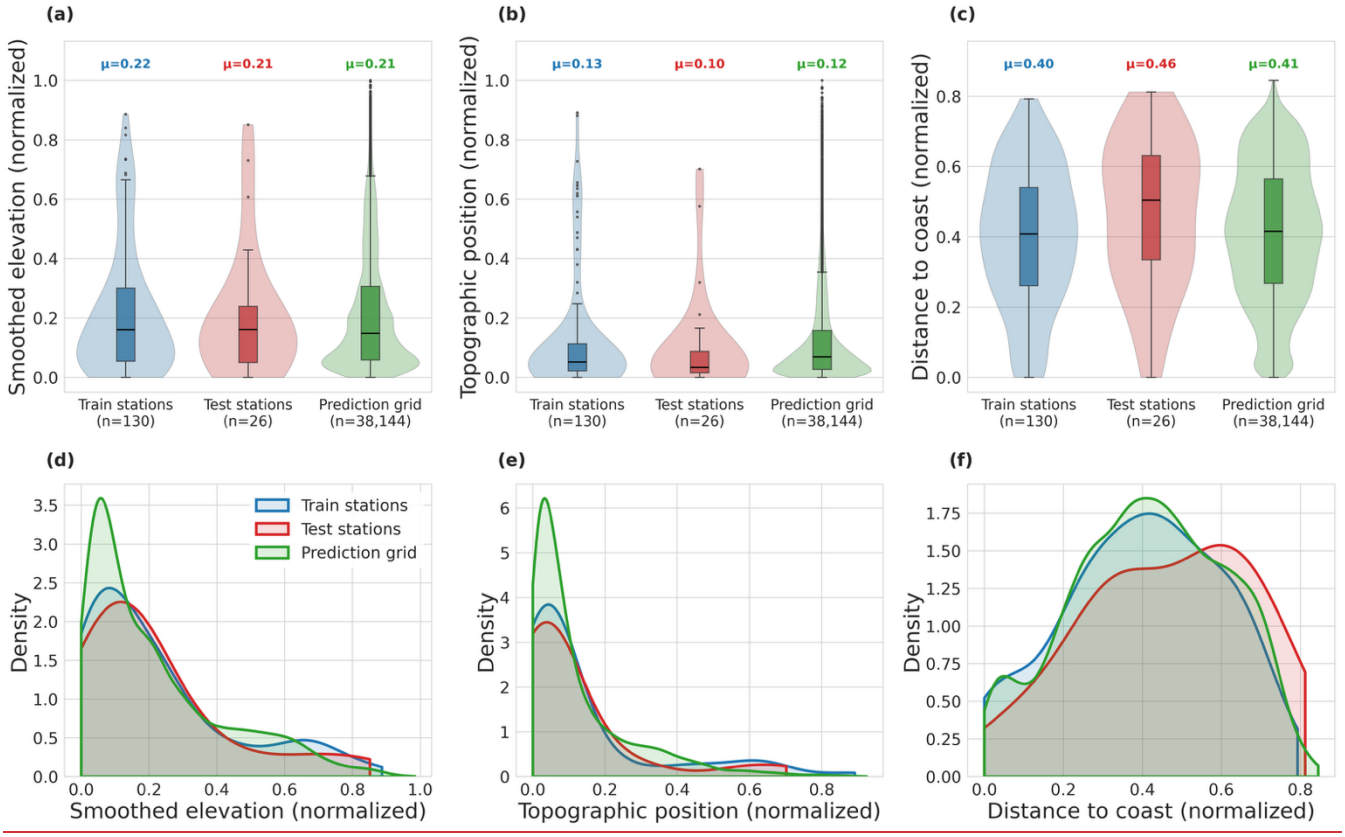

Figure 3 Distribution of key static predictors across training stations ($n = 130$), test stations ($n = 26$), and prediction grid cells within Romania ($n = 38,144$). (a–c) Violin plots with overlaid box plots for smoothed elevation, topographic position index, and distance to the Black Sea coast, respectively; the mean value (μ) for each distribution is indicated at the top of the panels. (d–f) Kernel density estimates for the same three predictors. Predictor values are shown in their normalized $[0, 1]$ form.

2.3 Data preprocessing

The raw data is processed into a format suitable for training the deep learning models. This involves normalization of the target variables and structuring the data into daily context and target sets.

The two climate variables, temperature and precipitation, are normalized independently. The daily mean temperature (*tavg)* is linearly scaled to a [0, 1] range using a *Min-Max scaler* fitted on the training data (Pedregosa et al., 2011). Raw

precipitation (*prec*) values p are transformed via $\log(1 + p)$ to handle zero values and reduce skewness, and the transformed values are then scaled to a [0, 1].

The parameters of these scalers are saved and used to de-normalize the model's predictions back to their original physical units (°C and mm) during validation and inference.

3 Methods

The core of our study is the comparison of a standard approach — the geostatistical method Regression Kriging (RK) — with deep learning architectures, namely the Spatial Multi-Attention Conditional Neural Process (SMACNP). Unlike RK, the deep learning models are used to jointly predict both air temperature and precipitation in a single realisation. The modelling approach differs fundamentally for the two variables due to their distinct statistical characteristics: air temperature, a continuous variable, is modelled directly, whereas daily precipitation, which is intermittent and non-negative, is handled using a two-part hurdle modelling framework in all tested configurations.

3.1 Regression Kriging (RK)

For air temperature, RK is applied directly to observed values. For precipitation, we employ a hurdle framework where occurrence and amount are modelled independently. In both cases, the field is decomposed into a deterministic trend estimated via multiple linear regression on all static and temporal predictor variables (smoothed elevation, topographic position index, distance to coast, geographic coordinates, and cyclic day-of-year/month/year encodings), and a stochastic residual interpolated using Ordinary Kriging with a spherical semi-variogram model. The variogram parameters (partial sill, range, and nugget) are estimated automatically for each daily field via weighted least-squares fitting. The detailed mathematical formulation is provided in Appendix A.

~~For air temperature, RK is applied directly to observed values. For precipitation, we employ a hurdle framework where occurrence and amount are modelled independently. The temperature and precipitation components are modelled as the sum of a deterministic trend, estimated via linear regression on topographic predictors, and a stochastic residual interpolated using Ordinary Kriging. As this is a well-established methodology, the detailed mathematical derivations for both variables are provided in Appendix A.~~

3.2 SMACNP

The Spatial Multi-Attention Conditional Neural Process (SMACNP) is a probabilistic deep learning model designed for interpolating sparse and irregularly distributed spatio-temporal data. It employs an encoder-decoder architecture that maps a set of observed context points to predictive distributions at unobserved target locations.

The model is conditioned on a set of N context points representing observed data. Each context point consists of spatial coordinates ($\mathbf{s}$), a vector of auxiliary attributes ($\mathbf{x}$), and the corresponding climate variable measurement (y). Predictions are made at a set of M target locations, each defined by its spatial coordinates ($\mathbf{s}^*$) and attributes ($\mathbf{x}^*$) (Figure 4).

The encoder is composed of three parallel pathways that process the input data into latent representations (Figure 4):

(i) *Spatial Pathway (Mean Location Encoder).* This pathway captures the spatial structure of the observations. It projects the concatenation of context coordinates and observations through a Multi-Layer Perceptron (MLP — a feedforward neural network applying successive linear transformations with nonlinear activations), then applies Laplace Attention — a distance-based weighting kernel using a configurable L_p norm ($p = 1$ for Manhattan distance, $p = 2$ for Euclidean distance):

$$w(\mathbf{s}^*, \mathbf{s}_i) \propto \exp\left(-\tau \cdot \| \mathbf{s}^* - \mathbf{s}_i \|_p\right) \quad (1)$$

where τ is a learnable temperature parameter. This produces a location-based representation $\mathbf{w}^*$ that encodes the spatial structure of the data.

(ii) *Attribute Pathway (Mean Attribute Encoder).* This pathway computes an attribute-based representation $\mathbf{r}^*$ using Multi-Head Attention. The mechanism operates through three vectors - Query (Q), Key (K), and Value (V) - that can be understood intuitively in the context of spatial interpolation. The Query encodes *what the target location needs*: "find stations with similar geographic and seasonal characteristics to mine" (e.g., similar elevation, topographic position, and time of year). The Key encodes *what each context station offers*: its own geographic and seasonal characteristics, which are compared against the Query to determine relevance. The Value contains *the actual information to transfer*: the context station's characteristics together with its observed measurements (temperature, precipitation). The attention output is a weighted combination of V, where stations whose Key is most similar to the target's Query receive the highest weights — effectively retrieving weather observations from the most geographically and seasonally relevant stations. The two encoder configurations differ in how Q, K, and V are constructed:

Component	Global Configuration	Localized Configuration
Query (Q)	Target attributes ($\mathbf{x}^*$)	Target coordinates + attributes ($\mathbf{s}^*$, $\mathbf{x}^*$)
Key (K)	Context attributes ($\mathbf{x}$)	Context coordinates + attributes of k-nearest neighbors ($\mathbf{s}$, $\mathbf{x}$)
Value (V)	Context attributes + observations ($\mathbf{x}$, y)	Context coordinates + attributes + observations of k-nearest neighbors ($\mathbf{s}$, $\mathbf{x}$, y)
Attention scope	All N context points ($N \times M$)	k-nearest spatial neighbors ($k \times M$)
Key property	Attribute-only similarity	Spatial locality + attribute similarity

The Localized configuration restricts attention to the k nearest spatial neighbors, introducing an inductive bias towards spatial locality that is well-suited for meteorological interpolation.

(iii) *Variance Pathway.* This pathway produces a representation $\mathbf{v}^*$ for estimating predictive uncertainty. It uses the same attention mechanism as the attribute pathway, but crucially excludes the observations (y) from the Value vectors. This prevents the model from memorizing input-output relationships and enables well-calibrated uncertainty estimates that reflect the information available at each target location.

The latent representations are routed to specialized decoder heads, each implemented as an MLP. The spatial and attribute representations ($\mathbf{w}^*$, $\mathbf{r}^*$) are concatenated with target information and fed to the heads predicting mean values and probabilities. The variance representation ($\mathbf{v}^*$) is similarly combined with target information and fed to the variance heads.

The decoder produces parameters for two predictive distributions. A Gaussian distribution for air temperature, parameterized by its mean (μ) and variance (σ^2). The mean uses a linear activation (allowing negative temperatures), while the variance uses a Softplus activation to ensure positivity. A hurdle model for precipitation separating occurrence from amount. A Bernoulli component models the probability of precipitation occurrence (p) via logistic activation; a log-normal component models the amount conditional on occurrence through its mean (μ) and variance (σ^2), both ensured positive via Softplus activation.

The model is trained end-to-end by minimizing the combined negative log-likelihood (NLL) loss — a weighted sum of the Gaussian NLL for temperature and the precipitation hurdle loss (Binary Cross-Entropy for occurrence + log-normal NLL for amount). To address class imbalance between rainy and dry days, the BCE loss incorporates a positive class weight. The full loss equations are provided in Appendix C.

~~The SMACNP is a probabilistic deep learning model designed for interpolating sparse and irregularly distributed spatio-temporal data, and employs an encoder-decoder architecture.~~

~~The model is conditioned on a set of n context points representing observed data. Each context point consists of spatial coordinates ($\mathbf{s}_c$), a vector of auxiliary attributes ($\mathbf{a}_c$), and the corresponding climate variable measurement ($\mathbf{y}_c$). Predictions are made at a set of m target locations, each defined by its spatial coordinates ($\mathbf{s}_t$) and attributes ($\mathbf{a}_t$) (Figure 3).~~

~~The encoder is composed of three parallel pathways that process the input data to generate a set of latent representations:~~

~~(1) Mean Location Encoder (Spatial Pathway) — This pathway is shared across all encoder configurations. It takes the concatenation of context spatial coordinates and observations $[\mathbf{s}_c, \mathbf{y}_c]$ as input. The pathway first projects this input through an MLP and then applies Laplace Attention, which computes distance based weights (α_{ij}) using a configurable L_p norm:~~

$$\alpha_{ij} = \frac{\exp\left(-\frac{\| \mathbf{s}_t^i - \mathbf{s}_c^j \|_p}{\tau}\right)}{\sum_{k=1}^{n} \exp\left(-\frac{\| \mathbf{s}_t^i - \mathbf{s}_c^k \|_p}{\tau}\right)} \quad (1)$$

where τ is a learnable temperature parameter and $p \in \{1,2\}$ selects between L1 (Manhattan) or L2 (Euclidean) distance. This produces a location-based representation $\mathbf{w}^*$ that captures the spatial structure of the data.

(2) Mean Attribute Encoder (Attribute Pathway) — This pathway processes context information to produce an attribute based representation $\mathbf{r}^*$. It projects the inputs into query (Q), key (K), and value (V) vectors to compute a weighted representation using standard Multi-Head Attention. For a single head, the operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where d_k is the dimension of the keys. The distinction between the two configurations lies in the definition of the input sets used to construct these vectors:

- *Global Configuration*: The query Q is derived from target attributes $\mathbf{a}_t$. Keys K are derived from context attributes $\mathbf{a}_c$ only (excluding observations), while values V are derived from both context attributes and observations $[\mathbf{a}_c, \mathbf{y}_c]$. This separation ensures that attention weights are computed purely based on attribute similarity, while the retrieved values incorporate the observed climate measurements. The mechanism computes attention over all context points, producing a full $n \times m$ attention matrix. Spatial coordinates are intentionally excluded from this pathway, as spatial information is captured separately by the location encoder.
- *Localized Configuration:* The query Q incorporates target spatial and attribute features $[\mathbf{s}_t, \mathbf{a}_t]$ (with optional positional encoding). Crucially, keys K are derived from context spatial and attribute features $[\mathbf{s}_c, \mathbf{a}_c]$, while values V additionally include observations $[\mathbf{s}_c, \mathbf{a}_c, \mathbf{y}_c]$. Both K and V are restricted to the k nearest spatial neighbors in the context set. This restricts attention to a local neighborhood k, reducing computational complexity from $\mathcal{O}(nm)$ to $\mathcal{O}(mk)$. This localized attention mechanism introduces a strong inductive bias towards spatial locality, a critical feature for meteorological interpolation that has not been previously explored within the SMACNP framework for real-world climate data.

(3) Variance Encoder (Variance Pathway) — This pathway produces a representation $\mathbf{v}^*$ for estimating predictive uncertainty. Crucially, observations ($\mathbf{y}_c$) are excluded from the value (V) in this pathway in both configurations to enable proper predictive uncertainty estimation without memorizing input-output relationships:

- *Global Configuration*: Takes context attributes only $[\mathbf{a}_c]$ to form K and V, and target attributes $[\mathbf{a}_t]$ for Q, using full $n \times m$ multi-head cross-attention.
- *Localized Configuration*: Takes context spatial coordinates and attributes $[\mathbf{s}_c, \mathbf{a}_c]$ (with optional positional encoding) to form K and V, and target features $[\mathbf{s}_t, \mathbf{a}_t]$ for Q, using k nearest neighbor attention.

The latent representations are routed to specialized decoder heads to generate the parameters for the predictive distributions. The location and attribute representations ($\mathbf{w}^*, \mathbf{r}^*$) are concatenated with target point information ($\mathbf{s}_t, \mathbf{a}_t$) and fed to the heads predicting mean values and probabilities. The variance representation ($\mathbf{v}^*$) is similarly combined with target information and fed to the heads predicting variance. Each decoder head is implemented as an MLP.

In our implementation, the decoder is equipped with two specialized sets of output heads corresponding to the target variables:

(1) Temperature outputs — The temperature head parameterizes a Gaussian distribution through its mean (μ) and variance (σ^2):

$$p(y_t|\phi_t) = \mathcal{N}(y_t; \mu_t, \sigma_t^2) \tag{3}$$

The mean decoder uses a linear output activation (allowing negative temperatures), while the variance decoder applies a Softplus activation to ensure $\sigma^2 > 0$.

(2) Precipitation outputs — The precipitation head employs a hurdle model to handle the mixed discrete-continuous nature of precipitation data. The model produces three outputs: the probability of occurrence (p), and the mean (μ) and variance (σ^2) of the precipitation amount conditional on precipitation being nonzero. The decoder implements this through two distinct components:

a) Occurrence Model (Bernoulli): Models the probability of precipitation occurrence:

$$p\left(y_t > 0|\phi_t^{\text{prob}}\right) = \text{Bernoulli}(p_t) \tag{4}$$

b) Amount Model (Log-Normal): Models the precipitation amount conditional on occurrence:

$$p(y_t|y_t > 0, \phi_t^{\text{amount}}) = \text{LogNormal}(\mu_t, \sigma_t^2) \tag{5}$$

The probability head outputs logits that are passed through a sigmoid function, while the mean and variance heads use *Softplus* activations to ensure positive outputs.

The model is trained end-to-end by minimizing the negative log-likelihood of the observed data. The total loss is a weighted combination of the temperature and precipitation losses:

$$\mathcal{L}_{\text{total}}(\theta) = w_{\text{temp}}\mathcal{L}_{\text{temp}}(\theta) + w_{\text{precip}}\mathcal{L}_{\text{precip}}(\theta) \tag{6}$$

(1) Temperature Loss — The standard Gaussian negative log-likelihood:

$$\mathcal{L}_{\text{temp}}(\theta) = \sum_{t \in T}\left[\frac{1}{2}\log(2\pi\sigma_t^2) + \frac{(y_t - \mu_t)^2}{2\sigma_t^2}\right] \tag{7}$$

(2) Precipitation Loss — A weighted sum of the binary cross-entropy for the occurrence model and the negative log-likelihood of the Log-Normal distribution for the amount model:

$$\mathcal{L}_{\text{precip}}(\theta) = \sum_{t \in T}\left[w_{\text{prob}} \cdot \text{BCE}\left(p_t, \mathbb{I}(y_t > 0)\right) + w_{\text{amount}} \cdot \mathbb{I}(y_t > 0) \cdot \text{NLL}_{\text{LogNorm}}(y_t; \mu_t, \sigma_t^2)\right] \tag{8}$$

where $\mathbb{I}(y_t > 0)$ is an indicator function that equals 1 if precipitation is greater than zero and 0 otherwise, BCE denotes binary cross-entropy, and $w_{\text{prob}}, w_{\text{amount}}$ are weights balancing the two components.

To address class imbalance between rainy and dry days, the binary cross-entropy loss incorporates a positive class weight computed as the ratio of dry to rainy samples, capped at a maximum value to prevent instability:

$$w_{\text{pos}} = \min\left(\frac{n_{\text{dry}}}{n_{\text{rainy}}}, w_{\text{max}}\right) \tag{9}$$

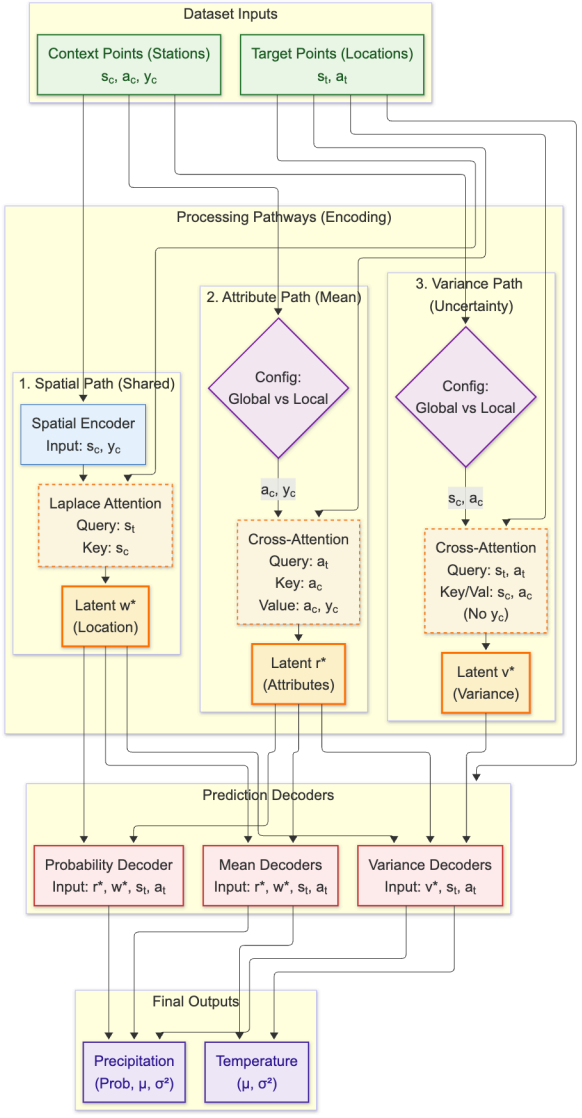

Figure 4~~3~~ –SMACNP architecture with dual encoder configurations. Three encoding pathways produce latent representations: spatial (w*) via Laplace attention, mean attributes (r*), and variance (v*). The mean and variance pathways offer interchangeable configurations: Global — full N×M attention on attributes only; Localized — k-NN attention on spatial coordinates plus attributes with optional positional encoding. Note that the variance pathway excludes observations (y_c) in both configurations to enable proper uncertainty estimation. Shared MLP decoders output temperature (μ, σ²) and precipitation (probability p, μ, σ²) predictions.

3.~~3~~2 Experimental setup

The training protocol involves two nested levels of data partitioning. First, the available days (2020–2023) are split into 80% for training and 20% for validation; the validation set is used exclusively for early stopping and model checkpoint selection. Second, within each training day, the 130 training-pool stations are randomly partitioned into a context set (~50% of stations,

used as model input) and the training loss is evaluated at all 130 stations ("on-the-grid" training). This means the model must learn both to reconstruct observations at context stations and to interpolate at the remaining non-context stations — the latter constituting the genuine test of generalization at each training step. The context partition is re-randomized at every training step, ensuring that every station regularly serves in both roles and that the model learns to interpolate from arbitrary subsets. All data normalization (min–max scaling) is fitted exclusively on the training-pool stations and applied without re-fitting to the test stations at evaluation time.

An extensive hyperparameter optimization (HPO) was conducted for the SMACNP model for both *global* and *localized* encoder configurations to identify the optimal architecture and training settings. The HPO process was managed using the Optuna framework (Akiba et al., 2019). For each model variant, we executed a total of 1000 trials. The optimization was guided by a Tree-structured Parzen Estimator (TPE) sampler, initialized with a fixed random seed to ensure reproducibility. The objective for each trial was to minimize the model's validation loss. While the learning rate was included in the hyperparameter search space (ranging from 1e-5 to 5e-4), other training parameters were held constant. We used the *AdamW* optimizer with a weight decay of 1e-5. The learning rate was managed by a scheduler configured to reduce the rate upon plateauing validation loss. To efficiently explore the search space, we implemented an aggressive early-stopping strategy, training each trial for up to 1,500 epochs. A patient pruner (wrapping a median pruner) was configured to terminate unpromising trials. Trials were pruned if their validation loss exceeded the median of previously completed trials for consecutive checks (patience of 5 checks), performed every 10 epochs after an initial warmup period (150 epochs).

The search spaces for the two configurations, including embedding dimensions, attention heads, and learning rates, are detailed in Appendix B. A constraint was enforced to ensure that embedding dimensions (r_dim, v_dim) were divisible by the number of attention heads.

Following the optimization process, the configurations yielding the lowest validation loss were selected for the final training and evaluation (Table B1). The SMACNP (Global) configuration favored a balanced architecture with a representation dimension (r_dim) of 128 and a variance dimension (v_dim) of 256. It utilized a shallower decoder (2 MLP layers) and higher attention complexity (16 heads) with minimal regularization (dropout rate of 0.1 and no layer normalization).

The SMACNP (Localized) configuration achieved optimal performance with a neighborhood size $k = 15$ and, like the global variant, utilized the Manhattan distance metric ($p = 1$) for the Laplace attention. The architecture favored a highly asymmetric embedding structure, allocating significant capacity to the variance ($v_{\text{dim}} = 512$) and location ($w_{\text{dim}} = 256$) pathways while keeping the representation dimension compact ($r_{\text{dim}} = 64$). The decoder was deeper than the global variant (3 vs. 2 MLP layers), but the optimization selected the same lightweight regularization - dropout rate of 0.1 and no Layer Normalization - suggesting that the inductive bias of localized attention provides sufficient regularization on its own. Both models converged to similar learning rates (4.93×10^{-4} for the global variant, 4.19×10^{-4} for the localized one).

~~The SMACNP (Localized) configuration achieved optimal performance with a neighborhood size (k) of 15 and, like the global variant, utilized the Manhattan distance metric ($p = 1.0$) for the Laplace attention. This architecture optimized~~

~~towards a highly asymmetric embedding structure, allocating significant capacity to the variance ($v\_dim = 512$) and location ($w\_dim = 256$) pathways relative to a compact representation dimension ($r\_dim = 64$). While it favored a deeper decoder structure (3 MLP layers), the optimization process ultimately selected the same lightweight regularization settings as the global model—a dropout rate of 0.1 and no Layer Normalization—suggesting that the architectural bias of the localized encoder provided sufficient regularization without requiring heavy stochastic noise or normalization. Both models converged to similar learning rates, with the global variant at 4.93×10^{-4} and the localized variant at 4.19×10^{-4}.~~

Results

We evaluated the results both qualitatively, by plotting multiannual-mean maps to verify that the predictions follow the expected spatial distribution of the analysed variables, and quantitatively, by computing accuracy metrics and corresponding plots using independent station data withheld from the training and prediction processes.

Visually, the outputs for air temperature from the three models are remarkably consistent (Figure 5~~Figure 4~~ a), showing nearly identical spatial distributions, successfully capturing the expected climatological patterns driven by the topography. All methods correctly identify the same locations for the coldest areas in the mountain ranges and the warmest areas in the lowlands, displaying comparable temperature gradients. This suggests that the deep learning models are fully capable of reproducing the dominant geographical factors influencing temperature distribution as effectively as the geostatistical baseline. Similar to the temperature results, all three models successfully represent the fundamental influence of topography on precipitation (Figure 5~~Figure 4~~ b). However, the visual differences in spatial detail are much more pronounced here.

The SMACNP Localized precipitation fields exhibit finer spatial detail than RK, particularly over the Carpathian arc where orographic gradients are steepest. Whether this additional detail reflects genuine skill in resolving local precipitation patterns or spatial artefacts would require validation against independent high-resolution observations (e.g., radar composites or dense temporary networks), which is beyond the scope of this study.

~~The RK model produces a slightly smoothed and generalized map; while it correctly identifies the broad zones of high precipitation in the Carpathians, the gradients are overly gradual, lacking fine-scale variability. In contrast, both SMACNP configurations generate significantly more detailed and textured precipitation fields. They resolve specific, concentrated areas of intense precipitation along mountain ridges and exhibit sharper transitions between wet and dry zones. These outputs appear physically more realistic, capturing the high spatial variability typical of precipitation in complex terrain, which the smoother RK baseline fails to resolve.~~

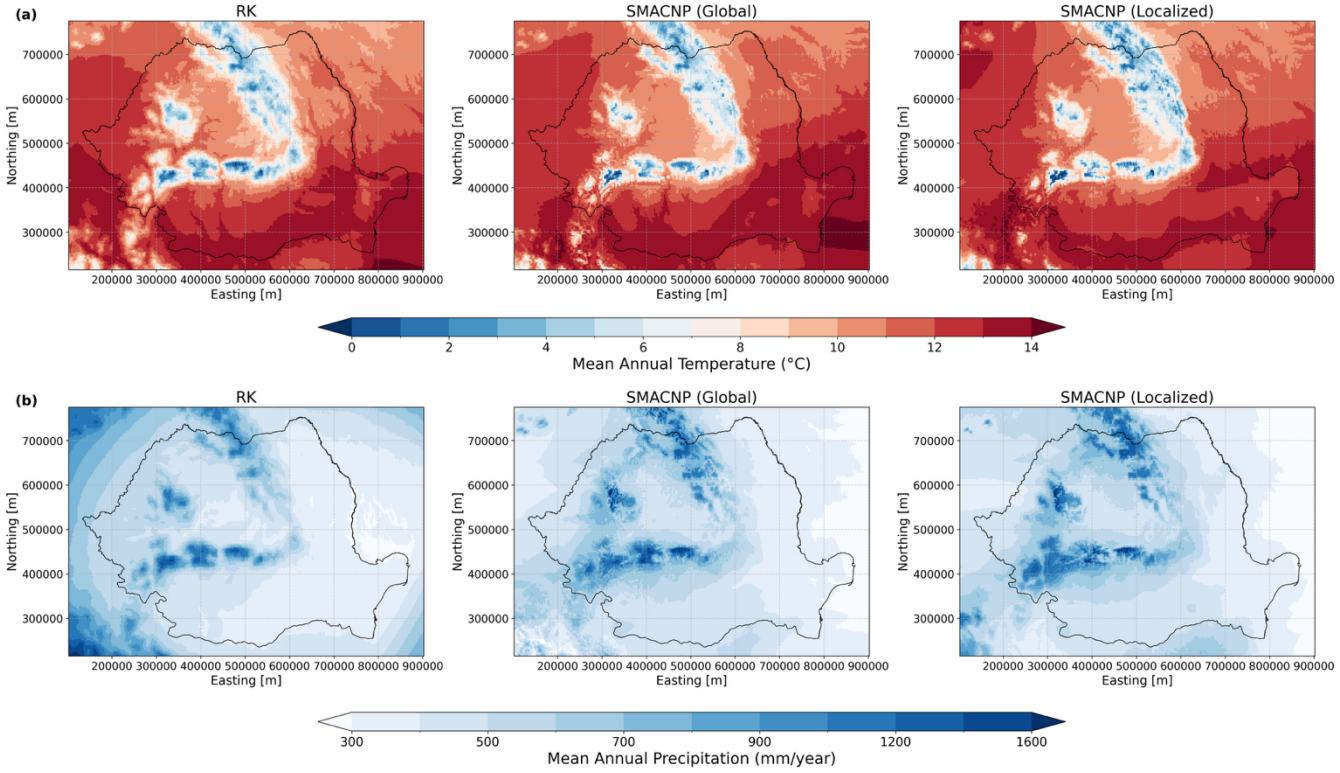

Figure 54 Comparison of predicted mean annual temperature (a) and precipitation (b) from three models: RK, SMACNP (Global), and SMACNP (Localized).

The density scatter plots in Figure 6~~Figure 5~~ demonstrate that all three models provide highly accurate predictions for daily mean temperature (panel a), with points clustering tightly along the 1:1 line of perfect agreement. The RK baseline achieves a correlation of 0.989 but exhibits slightly larger dispersion around the 1:1 line compared to the deep learning models, resulting in a higher MAE (0.938 °C). Both deep learning configurations outperform the baseline, with the SMACNP (Localized) variant achieving the best overall performance (MAE = 0.803 °C, RMSE = 1.108 °C), surpassing the SMACNP (Global) configuration (MAE = 0.838 °C). Across all models, prediction errors are slightly more pronounced in the tails of the distribution, yet the high density of points near the diagonal indicates strong agreement across the majority of the temperature range.

In contrast, the performance gap between the methods is more distinct for daily precipitation (panel b). Both SMACNP configurations substantially outperform the RK baseline, which suffers from larger errors (MAE = 1.181 mm) and a lower correlation (0.692). The SMACNP (Localized) model again delivers the strongest performance, achieving the lowest MAE (1.050 mm) and a correlation of 0.749, representing a notable improvement over the geostatistical baseline. While all models struggle with extreme precipitation events—a common challenge in climatological modeling—the deep learning approaches demonstrate a superior ability to capture the variability of moderate-to-high rainfall events compared to RK.

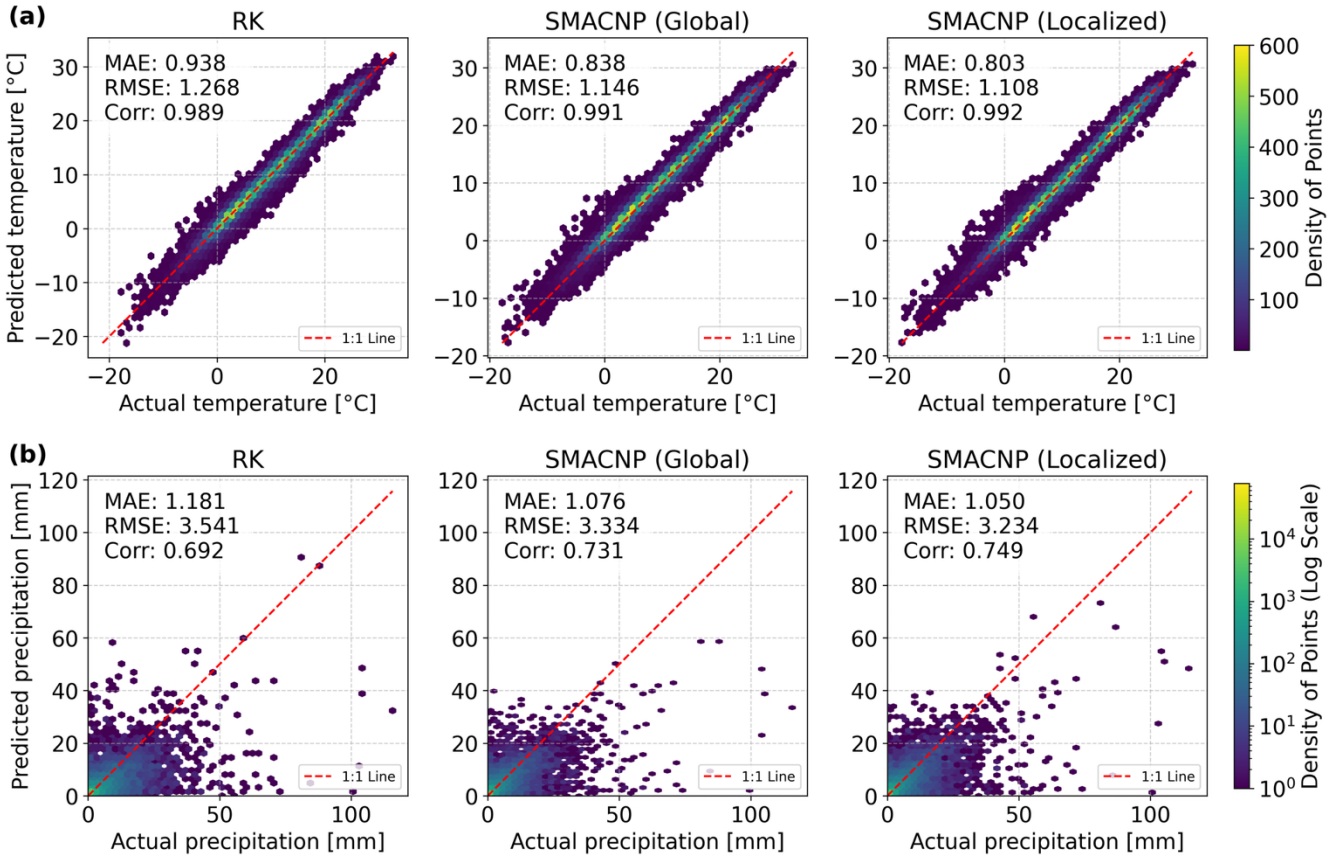

Figure ~~6~~5 Density scatter plots of predicted versus actual (a) daily mean temperature (°C) and (b) daily precipitation (mm) for the three models on the test dataset: RK, SMACNP (Global), and SMACNP (Localized). For temperature, the colour of each hexagonal bin indicates point density on a linear scale, while for precipitation (zero-inflated data), density is shown on a logarithmic scale. The dashed red line represents the 1:1 line of perfect agreement.

Across all seasons, SMACNP (Localized) and SMACNP (Global) consistently outperform the RK model for both precipitation and temperature (Table 1~~Table 1~~). For temperature, all models perform well, with correlations exceeding 0.94 in all cases. SMACNP (Localized) surpasses the others across all four seasons, consistently achieving the lowest MAE and RMSE, reflecting its superior advantage in capturing fine-scale temperature variations compared to both the Global variant and the RK baseline. For precipitation, SMACNP (Localized) demonstrates dominant performance, achieving the lowest MAE and RMSE across all seasons, as well as the highest correlations (e.g., 0.808 in DJF). While SMACNP (Global) remains competitive, SMACNP (Localized) outperforms it even in summer (JJA), yielding the lowest RMSE (4.958) and highest correlation (0.677). The RK model exhibits the weakest performance, particularly for precipitation during JJA, where RMSE reaches 5.374. Overall, the results demonstrate that neural process–based models substantially enhance predictive accuracy over the baseline RK approach, especially for precipitation, where spatial heterogeneity and nonlinearity are more pronounced. Based on its superior performance across the majority of metrics—particularly MAE—and seasons, SMACNP (Localized) emerges as the most consistently high-performing method overall.

Table 1 Seasonal evaluation metrics for temperature and precipitation across all models

Season	Method	Temperature			Precipitation		
		MAE	**RMSE**	**CORR**	**MAE**	**RMSE**	**CORR**
DJF	RK	1.038	1.445	0.947	0.862	2.289	0.758
	SMACNP (Global)	0.950	1.327	0.957	0.796	2.205	0.780
	SMACNP (Localized)	**0.902**	**1.296**	**0.957**	**0.756**	**2.066**	**0.808**
MAM	RK	0.889	1.176	0.982	1.034	2.773	0.727
	SMACNP (Global)	0.798	1.060	0.984	0.915	2.509	0.783
	SMACNP (Localized)	**0.737**	**0.985**	**0.986**	**0.903**	**2.497**	**0.784**
JJA	RK	0.856	1.131	0.969	1.912	5.374	0.606
	SMACNP (Global)	0.742	0.987	0.972	1.769	5.100	0.655
	SMACNP (Localized)	**0.733**	**0.961**	**0.973**	**1.739**	**4.958**	**0.677**
SON	RK	0.973	1.301	0.978	0.907	2.859	0.787
	SMACNP (Global)	0.866	1.185	0.981	0.817	2.670	0.815
	SMACNP (Localized)	**0.845**	**1.160**	**0.982**	**0.793**	**2.570**	**0.830**

To validate the choice of SMACNP over alternative neural process architectures, we also evaluated ConvCNP (Gordon et al., 2019) and GriddedTNP (Ashman et al., 2024), two grid-based models that discretize station observations onto a regular grid before processing with a convolutional backbone. For temperature, neither grid-based architecture matched SMACNP Localized: GriddedTNP achieved an MAE of 1.05°C — approaching Regression Kriging (0.94°C) — while ConvCNP yielded 1.53°C. For precipitation, both models performed competitively, with ConvCNP (MAE = 1.11 mm) approaching SMACNP Localized (1.05 mm) and GriddedTNP (1.22 mm) approaching RK (1.24 mm). The asymmetric performance across variables reflects a fundamental architectural trade-off: grid discretization preserves the large-scale spatial patterns that drive precipitation but smooths the fine-scale topographic contrasts critical for resolving temperature lapse rates in complex terrain. SMACNP's station-to-station attention mechanism avoids this information loss, explaining its consistent superiority across both variables.

To complement the error analysis, we further examine how well the models quantify uncertainty and represent the full precipitation distribution (Figure 7~~Figure 6~~ and Table 2~~Table 2~~). For temperature (panel a), the z-score distributions show mean values generally close to zero (Table 2; $|\text{mean } z| \leq 0.14$), though SMACNP (Global) shows a slight positive bias (+0.14). The RK baseline exhibits a narrow z-score distribution (std ≈ 0.70) resulting in conservative, over-dispersive 95% intervals (95% coverage ≈ 0.99). SMACNP (Global) approaches a standard normal distribution (std ≈ 0.99) with slightly under-dispersive coverage (93.7%), while SMACNP (Localized) maintains a standard deviation of ≈ 0.87 and achieves coverage closest to nominal (1σ: 0.81, 95%: 0.96). These improvements are reflected in the temperature CRPS, where both

SMACNP configurations outperform RK (0.59–0.62 vs 0.72, with a slight advantage for the localized model). For precipitation amounts on wet days (panel b), all models exhibit negative mean z-scores, indicating a systematic underestimation of the intensity of rainy events. This bias is strongest for RK (mean z ≈ −0.70) and reduced for the SMACNP variants (≈ −0.46 to −0.47). The spread of z-scores is much larger than one for all models, reflecting the strong skewness and heavy tails of the wet-day distribution, but again the deep models show substantial gains: the standard deviation drops from ≈ 3.15 (RK) to 2.73 (SMACNP Global) and 1.92 (SMACNP Localized). The 95% coverage on wet days increases from 0.90 for RK to about 0.93 for both SMACNP variants, indicating that extreme wet events are captured more reliably, though some under-dispersion remains. Consistently, the CRPS for wet-day precipitation decreases from 2.30 (RK) to 2.13 and 2.12 for the global and localized SMACNP models, respectively. Panel (c) focuses on precipitation occurrence, implicitly derived from the marginal predictive distributions. All models maintain high hit rates for rainy days (≈ 0.95–0.96), but they differ notably in how often they falsely predict rain. RK strongly overestimates the frequency of wet days (predicted wet fraction ≈ 0.57 vs observed ≈ 0.34), leading to a high false-alarm rate of about 0.38 and a relatively low correct-dry rate of 0.63. Both SMACNP configurations bring the predicted wet fraction much closer to the observed value (≈ 0.47–0.48), reduce the false-alarm rate to about 0.23, and increase the correct-dry rate to ≈ 0.77. This indicates fewer spurious light-rain estimates while preserving sensitivity to true rain events. The coverage summary in panel (d) integrates these results. For temperature, RK over-covers at the 95% level (0.99), whereas SMACNP (Global) slightly under-covers (0.94) and SMACNP (Localized) is near-ideal (0.96). For wet-day precipitation, the RK intervals are too narrow in the tails (95% coverage ≈ 0.90), while SMACNP models increase the 95% coverage to ≈ 0.93 and modestly over-cover at 1σ (0.84), consistent with somewhat wider but more reliable uncertainty intervals. Finally, the CRPS scores in panel (e) summarise overall probabilistic skill in a distribution-agnostic way. For both temperature and precipitation (all days and wet days only), the two SMACNP variants consistently achieve lower CRPS than RK, with the localized version marginally outperforming the global one. Taken together, these diagnostics show that the deep learning approaches, and particularly SMACNP (Localized), not only reduce deterministic errors but also provide better-calibrated and more informative uncertainty estimates, especially for precipitation occurrence and intensity.

To further characterize precipitation performance, we computed seven climate-relevant diagnostics at the held-out test stations (Table 3). Wet-day frequency bias is the ratio of predicted to observed wet-day counts (threshold ≥ 1.0 mm, WMO standard), where 1.0 indicates unbiased frequency. Mean intensity bias is the ratio of mean predicted to mean observed precipitation on wet days, capturing whether the model systematically under- or overestimates rainfall amounts. Wet-day MAE and wet-day RMSE measure the average absolute and root-mean-square errors, respectively, computed only on observed wet days. R95p bias and R99p bias are the percentage deviations in total precipitation from days exceeding the 95th and 99th percentiles of the observed wet-day distribution, respectively, computed per station per year following the ETCCDI standard (Klein Tank et al., 2009) and then averaged across all station-years. Rx1day bias is the percentage deviation in the mean annual maximum one-day precipitation, computed per station per year following the same standard. All models slightly overpredict wet-day frequency (~20% overestimation) and underestimate wet-day intensity, consistent

with the smoothing inherent in spatial prediction methods. SMACNP Localized shows the smallest biases across all diagnostics, including frequency (1.193), intensity (0.762), aggregate extreme indices (R95p bias: −51%; R99p bias: −61%), and Rx1day (−27%). The underestimation of extreme precipitation totals is a known limitation of spatial interpolation from sparse networks, where individual convective events may not be captured by the nearest context stations.

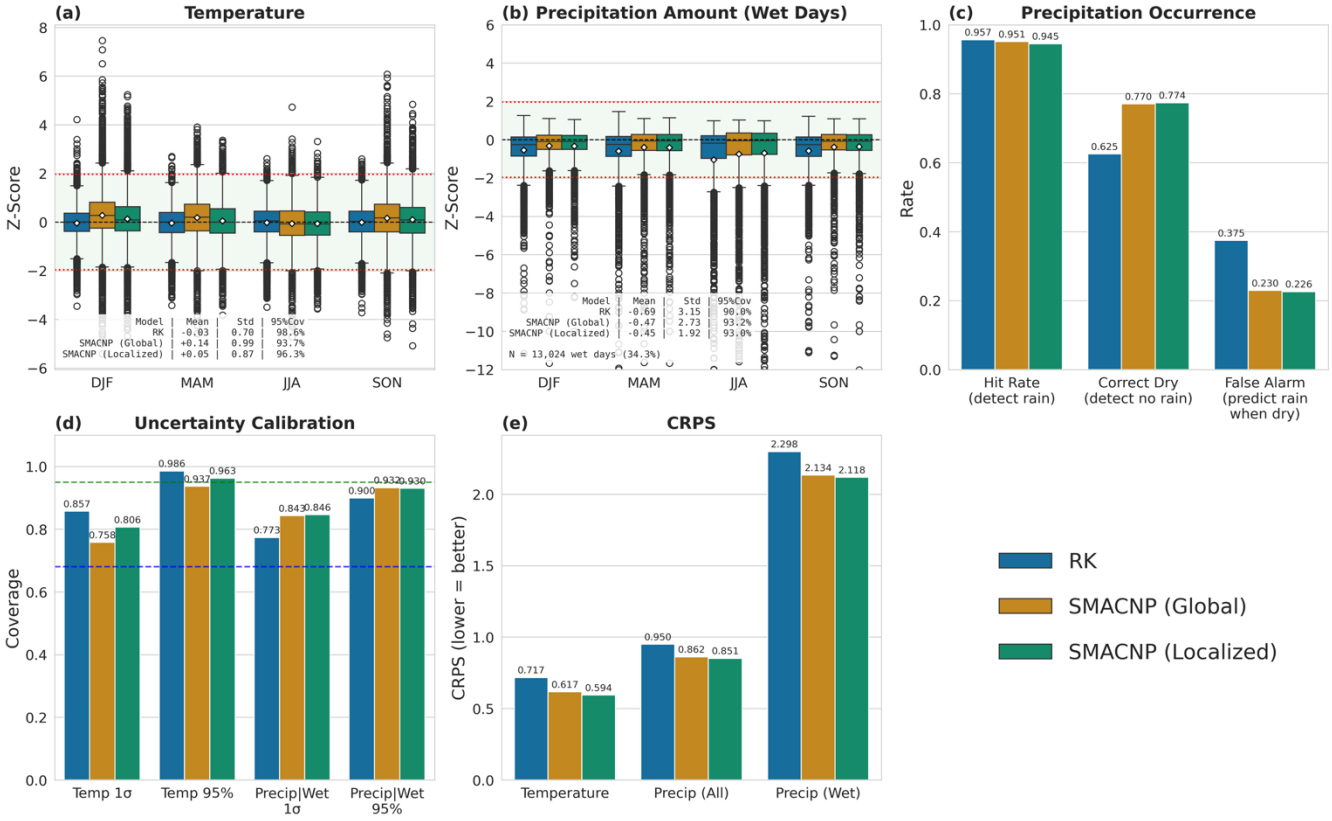

Figure 7~~6~~ Seasonal uncertainty calibration and overall probabilistic verification. (a) Seasonal distributions of standardized errors (z-scores) for temperature. (b) As in (a), but for precipitation amounts on wet days only (observed precipitation > 0.1 mm; sample size and wet-day fraction indicated in the panel). Z-scores are defined as (prediction − observation) / predicted_std; the red horizontal lines mark ±1.96, corresponding to the 95% interval of a standard normal distribution and used as a visual reference for well-calibrated predictions (mean ≈ 0, std ≈ 1). (c) Precipitation occurrence scores implicitly derived from the marginal predictive distributions: hit rate for rain detection, correct-dry rate, and false-alarm rate (predicting rain when it is dry). (d) Empirical coverage of the nominal 1σ and 95% predictive intervals for temperature and precipitation (wet days), with dashed horizontal lines indicating the expected values of 0.68 and 0.95. (e) Continuous Ranked Probability Score (CRPS) for temperature, all-day precipitation, and wet-day precipitation (lower values indicate better probabilistic skill).

Table 2 Summary of probabilistic verification metrics for temperature and precipitation. Metrics are aggregated over all seasons and stations for the three methods (RK, SMACNP Global, SMACNP Localized). For temperature, CRPS quantifies overall probabilistic skill (lower is better), "95% Cov" is the empirical coverage of the nominal 95% predictive interval, and "Mean Z" is the mean standardized error, with an ideal value of 0. For precipitation detection, scores are computed using an occurrence threshold of 0.1 mm: Hit Rate is the fraction of wet days correctly detected, FAR (false alarm rate) is the fraction of dry days on which rain is predicted, and True Dry is the fraction of dry days correctly forecast as dry. For precipitation intensity, CRPS, 95% coverage and Mean Z are evaluated on wet days only (observed precipitation > 0.1 mm). An ideal model has low CRPS and FAR, high Hit Rate and True Dry, 95% Cov close to 0.95, and |Mean Z| close to zero.

Method	Temperature				Precipitation (Detection)			Precipitation (Intensity)			
	CRPS	**95% Cov**	**Mean Z**	**Std Z**	**Hit Rate**	**FAR**	**True Dry**	**CRPS**	**95% Cov**	**Mean Z**	**Std Z**
RK	0.717	0.986	**-0.031**	0.703	**0.957**	0.375	0.625	2.298	0.900	-0.695	3.152
SMACNP (Global)	0.617	0.937	0.136	**0.995**	0.951	0.230	0.770	2.134	**0.932**	-0.466	2.732
SMACNP (Localized)	**0.594**	**0.963**	0.050	0.874	0.945	**0.226**	**0.774**	**2.118**	0.930	**-0.455**	**1.916**

Table 3 Climate-relevant precipitation diagnostics at held-out test stations ($n = 26$), 2020–2023. Wet-day threshold = 1.0 mm (WMO standard). Frequency and intensity biases are expressed as ratios (1.0 = unbiased). R95p, R99p, and Rx1day biases are expressed as percentage deviations from observed values. Bold values indicate the best (closest to unbiased) performance for each diagnostic.

Diagnostic	RK	SMACNP Global	SMACNP Localized
Wet-day frequency bias (pred/obs)	1.200	1.235	**1.193**
Mean intensity bias (pred/obs)	0.633	0.746	**0.762**
Wet-day MAE (mm)	3.83	3.46	**3.43**
Wet-day RMSE (mm)	7.03	6.52	**6.36**

Diagnostic	RK	SMACNP Global	SMACNP Localized
R95p bias (%)	−58.2	−53.5	**−50.9**
R99p bias (%)	−66.1	−63.3	**−60.5**
Rx1day bias (%)	−31.0	−33.5	**−27.2**

Figure 8~~Figure 7~~ provides a time-series comparison of model performance for daily mean air temperature and total precipitation at the Bacău station (174 m a.m.s.l.) throughout 2023, selected from the test data subset. The figure shows the ground truth observations against the predictions from the three evaluated models.

For air temperature, all three models demonstrate a strong ability to capture the seasonal cycle and daily fluctuations, closely tracking the ground truth data. The shaded areas, representing the 95% confidence intervals, are relatively narrow for all models, indicating a high degree of certainty in their temperature predictions, though the RK intervals are noticeably wider than those of the deep learning models. The MAE values inset in the plot show that the deep learning models outperform the baseline, with SMACNP (Localized) achieving the lowest MAE of 0.468°C, followed by SMACNP (Global) at 0.488°C, both improving upon the RK model's MAE of 0.575°C.

The lower panel, illustrating total precipitation, reveals the challenges in modeling this discontinuous variable. While all models capture the timing of major precipitation events, the SMACNP (Localized) model demonstrates the best performance in predicting the magnitude of these events, as evidenced by its lower MAE of 0.663 mm. SMACNP (Global) follows closely with an MAE of 0.687 mm, while the RK model has the highest MAE at 0.789 mm. The 95% confidence intervals are visibly wider for precipitation than for temperature across all models, reflecting the greater inherent uncertainty in predicting this variable. Notably, the confidence intervals for the deep learning models, particularly SMACNP (Localized), often envelop the observed precipitation peaks more effectively than the RK model, suggesting a more reliable quantification of uncertainty. This is especially apparent during the intense precipitation events observed in September and November.

Corresponding time-series comparisons for all other stations in the test set are provided in the Supplement (Figures S1–S25), demonstrating consistent performance across diverse elevations and locations.

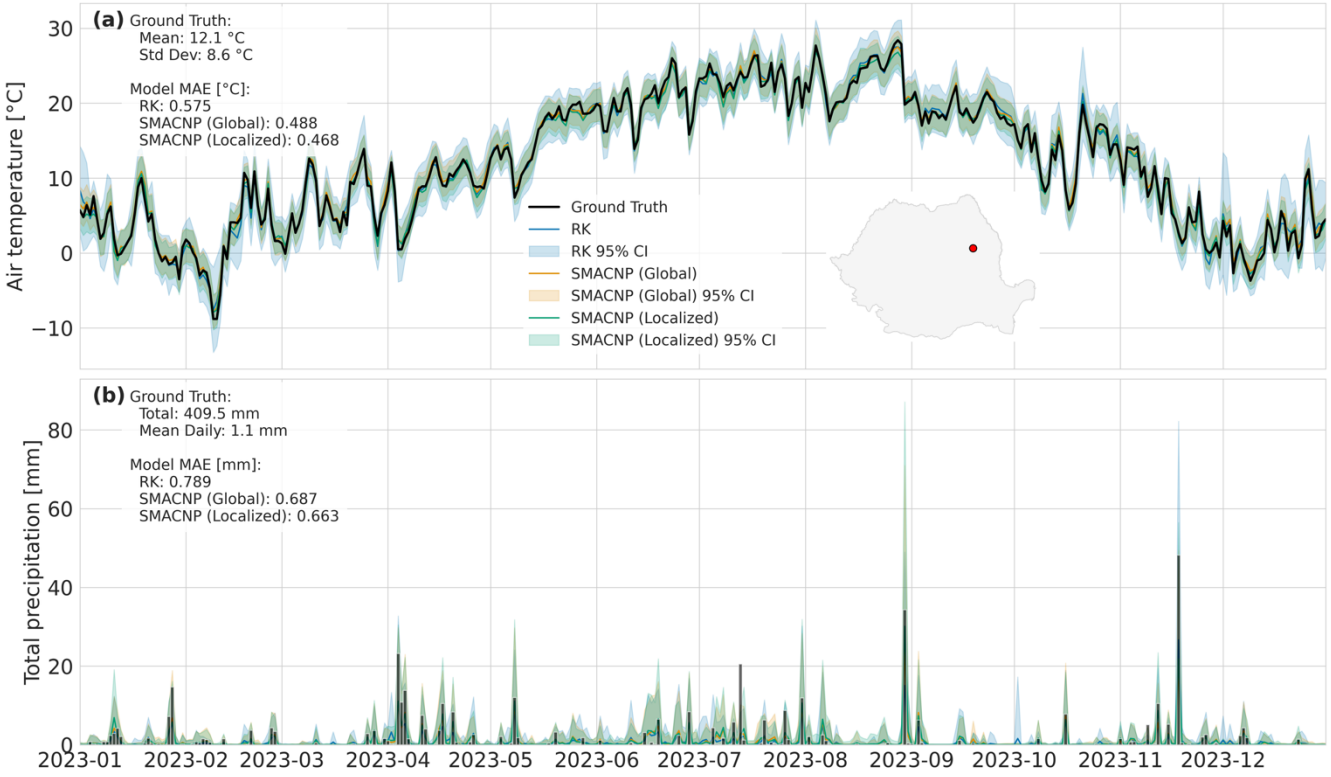

Figure 8~~7~~ Model performance and uncertainty at Bacău station (428 m a.m.s.l.). Time series comparison of (a) daily average temperature and (b) daily total precipitation for the year 2023. Shaded areas represent the 95% confidence interval for each model's prediction. The inset map shows the location of the station within the study area, marked by a red dot.

5 Discussions

This study evaluated the SMACNP architecture — an attentive Conditional Neural Process originally proposed for general meta-learning tasks — as a practical alternative to Regression Kriging for daily climate gridding from a sparse national station network over complex terrain. The superior performance of the neural process models, particularly for precipitation, can be attributed to their ability to learn complex, non-linear relationships directly from the data. Unlike RK, which relies on a pre-defined linear model for the trend and a stationary variogram for the residuals, deep learning architectures can capture the intricate interplay between topography, location, and climate variables. This is especially critical for precipitation, whose spatial distribution in mountainous terrain is often intermittent, highly localized, and non-stationary—conditions that violate the core assumptions of RK and explain its tendency to produce overly smooth and generalized outputs.

Among the two deep learning architectures, the good results of the SMACNP with a *localized* encoder highlights the benefit of incorporating a strong structural prior into the model design. This approach aligns with the principles driving recent breakthroughs in global weather forecasting, where models like Google's GraphCast and ECMWF AIFS have demonstrated

the importance of effectively capturing local spatial dependencies (Lam et al., 2023; Moldovan et al., 2025). In our application, by restricting the attention mechanism to a $k-NN$, the localized encoder is explicitly guided to learn representations based on spatial proximity. This provides a powerful inductive bias for modeling climate variables, contrasting with the more flexible, but less constrained, global attention mechanism of the standard SMACNP. Moreover, the localized configuration offers superior computational efficiency when generating high-resolution grids. By limiting interactions to the k nearest neighbors ($k-NN$) rather than computing global pairwise correlations, it avoids the quadratic complexity of the global encoder, making it significantly more scalable for producing dense gridded outputs. Building on this, future studies could further enhance the configuration by incorporating explicit graph structures within the encoder. Adopting Graph Neural Network (GNN) architectures would allow for sophisticated message passing between observation points, potentially capturing complex, non-linear spatial dependencies more effectively than the current attention-based neighborhood aggregation.

The practical importance of well-calibrated uncertainty cannot be overstated. For applications such as flood risk assessment or agricultural modeling, an unbiased estimate of the prediction's confidence is as valuable as the prediction itself. Our analysis reveals that only SMACNP (Localized) provides reliable uncertainty estimates across both temperature and precipitation. The severe overconfidence of the RK model's precipitation uncertainties renders them impractical for risk-aware decision-making. Furthermore, the uncertainty quantification of the deep learning models could be enhanced by adopting a deep ensemble approach. As demonstrated in the creation of a global canopy height map from satellite imagery (Lang et al., 2023), training an ensemble of models with different random initializations offers a robust approach to estimate epistemic uncertainty—the model's lack of knowledge in regions or conditions underrepresented in the training data. The variation among ensemble predictions provides a direct measure of model confidence. While the inherent probabilistic design of our SMACNP (Localized) already captures aleatoric uncertainty, combining it with an ensemble strategy would yield a more comprehensive and reliable characterization of total predictive uncertainty, further supporting the adoption of such frameworks for operational climate datasets.

Remarkably, the SMACNP models achieved superior accuracy compared to the RK baseline, even though it was constrained to model temperature and precipitation simultaneously. While RK had the advantage of fitting distinct models for each variable, the deep learning approach demonstrated that a unified architecture could leverage shared spatio-temporal features to outperform the specialized, single-variable geostatistical models.

An advantage of the multivariate SMACNP framework is its ability to learn shared latent representations across climate variables. Because temperature and precipitation are predicted jointly through shared encoder pathways, the model can implicitly capture cross-variable dependencies, for instance, the coupled effects of topography on temperature lapse rates and orographic precipitation enhancement. In classical geostatistics, such dependencies can be exploited through co-kriging, which models cross-variograms between variables. The deep learning approach achieves a similar effect but extends to nonlinear relationships and scales naturally to additional variables (e.g., wind speed, humidity) through additional output heads, without requiring separate models or the estimation of additional cross-variogram parameters. While we do not

present a formal univariate ablation in this study, we note that training separate single-variable models is straightforward within the SMACNP framework, and such a comparison would help quantify the benefit of joint variable learning for climate gridding.

The SMACNP model is designed for temporal transferability. Because the model is conditioned on actual station observations at each prediction step, it learns a spatial interpolation function that is largely independent of the specific training period. The topography-climate relationships captured by the encoder are expected to remain valid for periods outside the 2020–2023 training window, provided that the station network density and the underlying spatial structure of the climate fields do not change substantially. Extending the training period to include additional years would improve the model's exposure to rare events and extreme values, and is a priority for future development.

It is also pertinent to address the architectural selection for this study. During the initial phases, we evaluated other neural process variants, including the ConvCNP and GriddedTNP. However, these models failed to yield satisfactory results under the strict constraints of our experimental setup; the combination of a very sparse observation network (130 stations) and the exclusion of dynamic covariates appeared to hinder the ability of these architectures to resolve local features effectively. While GriddedTNP is designed to scale efficiently to large datasets, it likely struggled to construct a robust latent grid representation from such sparse inputs. The SMACNP architecture proved suited to this data-scarce regime because its attention mechanism explicitly models directly the relationship between specific context and target points based on topographic similarity, rather than relying on grid-based projections or convolutions that typically require denser data to perform optimally.

It is important to acknowledge the limitations of this study. Our analysis was intentionally restricted to static, topographically-derived predictors to ensure the developed models are applicable to long-term historical datasets where dynamic covariates like satellite imagery products are unavailable. The inclusion of dynamic predictors could further enhance model performance for modern periods. Also, the evaluation is restricted to a single country. Although Romania presents a demanding combination of topographic complexity, station sparsity, and mixed precipitation regimes, the generalizability of the results to other climatic regions and station network configurations remains to be demonstrated. In particular, the relative advantage of the localized attention mechanism may depend on factors such as station density, the dominant spatial scales of variability, and the degree of topographic control on the target variables. Extending the framework to additional countries or continental-scale networks, and assessing its performance under different data availability conditions, is a priority for future work.

6 Conclusions

In this study, we evaluated and compared a traditional geostatistical interpolation method, Regression Kriging (RK), with two advanced deep learning models, the Spatial Multi-Attention Conditional Neural Process (SMACNP) with Localized and

Global encoders, for producing daily gridded fields of air temperature and precipitation from a very sparse station data in Romania.

The primary conclusions of this work are as follows:

1. Deep learning models substantially outperform the geostatistical baseline**.** Both SMACNP configurations demonstrated significantly higher accuracy than RK for both climate variables. The performance gains were particularly large for precipitation, where the neural process models were better able to capture the complex, non-linear, and heterogeneous spatial patterns inherent to rainfall in a topographically diverse region.
2. SMACNP (Localized) is the most robust and highest-performing model. Our study identifies this specific configuration as a novel and highly effective approach for sparse data interpolation. By explicitly leveraging local neighborhood structure via a localized encoder, the SMACNP (Localized) model consistently achieved the best or near-best performance across most evaluation metrics and seasons. It excelled at creating detailed and realistic spatial fields while minimizing prediction errors.
3. Probabilistic deep learning provides superior uncertainty quantification**.** A key advantage of the neural process framework is its native ability to provide probabilistic predictions. The SMACNP (Localized) model, in particular, produced well-calibrated and reliable uncertainty estimates for both temperature and precipitation. In contrast, the RK model's precipitation uncertainty estimates showed signs of overconfidence, reducing their reliability.

Our findings demonstrate that a unified deep learning framework can outperform specialized univariate geostatistical models, effectively leveraging shared spatio-temporal patterns to improve accuracy across multiple climate variables simultaneously. The ability of these models to learn from data to represent complex spatial dependencies, combined with their capacity for reliable uncertainty estimation, makes them exceptionally well-suited for the development of high-quality gridded climate datasets. The results of this study provide a strong methodological foundation for the future production of a new, long-term, high-resolution daily climate dataset.

Appendix A Regression Kriging Formulation

The temperature Z at a location s is modeled as the sum of a deterministic trend $m(s)$ and a stochastic, spatially correlated residual $R(s)$:

$$Z(s) = m(s) + R(s) \tag{A1}$$

The prediction $\hat{Z}(s_0)$ at a target location s_0 is derived in three steps:

(1) The trend $m(s)$ is estimated using a multiple linear regression model based on a set of p predictor variables $x_k(s)$. The estimated trend $\hat{m}(s)$ is:

$$\hat{m}(s) = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k \, x_k(s) \qquad (A2)$$

where $\hat{\beta}_k$ are the regression coefficients fitted using the data at the station locations.

(2) The residuals at the n station locations s_i are calculated as:

$$R(s_i) = Z(s_i) - \hat{m}(s_i) \qquad (A3)$$

These residuals are then interpolated to the target location s_0 using ordinary kriging (OK). The kriged residual $\hat{R}(s_0)$ is a weighted average of the observed residuals:

$$\hat{R}(s_0) = \sum_{i=1}^{n} w_i \, (s_0) R(s_i) \qquad (A4)$$

where the weights w_i come from the kriging system.~~are determined by solving the kriging equations based on a fitted spherical variogram model of the residuals~~.

The kriging weights are obtained from a spherical semi-variogram model fitted to the residuals for each daily field independently. The spherical model is defined as:

$$\gamma(h) = c_0 + c_1 \left[1.5 \left(\frac{h}{a}\right) - 0.5 \left(\frac{h}{a}\right)^3 \right] \quad \text{for } 0 < h \le a, \qquad \gamma(h) = c_0 + c_1 \quad \text{for } h > a \qquad (A5)$$

where c_0 is the nugget, c_1 is the partial sill, and a is the range parameter. These parameters are estimated via weighted least-squares fitting (Benjamin Murphy et al., 2025) separately for every daily residual field, allowing the spatial correlation structure to adapt to the day-to-day variability in the data. Because the variogram is re-fitted daily, we do not report fixed parameter values; rather, the semi-variogram captures the evolving spatial structure of the regression residuals across the time series.

(3) The final temperature prediction is the sum of the estimated trend and the kriged residual:

$$\hat{Z}(s_0) = \hat{m}(s_0) + \hat{R}(s_0) \qquad (A6)$$

A hurdle model is used for precipitation $Y(s)$, which separates the process into two parts: the probability of rain and the amount of rain if it occurs. The final prediction $\hat{Y}(s_0)$ is the statistical expectation:

$$\hat{Y}(s_0) = P(Y(s_0) > 0) \cdot E[Y(s_0) | Y(s_0) > 0] \qquad (A7)$$

Probability of precipitation $P(Y(s_0) > 0)$ is estimated using a logistic regression model. The probability $p(s)$ at a location s is modeled as:

$$\text{logit}\big(p(s)\big) = \log\left(\frac{p(s)}{1 - p(s)}\right) = \gamma_0 + \sum_{k=1}^{p} \gamma_k \, x_k(s) \qquad (A8)$$

The predicted probability $\hat{p}(s_0)$ at the target location is the output of this model. The amount of precipitation, conditioned on rain occurring $E[Y(s_0) | Y(s_0) > 0]$, is modeled using RK on log-transformed data for rainy stations only $L(s) = \log(Y(s) + 1)$. The process mirrors the RK model for temperature:

(1) A linear model is fitted for $L(s)$ at rainy stations:

$$\hat{m}_L(s) = \hat{\alpha}_0 + \sum_{k=1}^{p} \hat{\alpha}_k \, x_k(s) \qquad (A9\sout{8})$$

(2) Residuals $R_L(s_i) = L(s_i) - \hat{m}_L(s_i)$ are computed and kriged to get $\hat{R}_L(s_0)$

(3) Prediction (Log-scale):

$$\hat{L}(s_0) = \hat{m}_L(s_0) + \hat{R}_L(s_0) \qquad (A10\sout{9})$$

The predicted amount is obtained by back-transforming the result: $\exp\left(\hat{L}(s_0)\right) - 1$

(4) The final precipitation prediction is the product of the two parts:

$$\hat{Y}(s_0) = \hat{p}(s_0) \cdot \left(\exp\left(\hat{L}(s_0)\right) - 1\right) \qquad (A11\sout{0})$$

Appendix B Hyperparameter Optimization Details

SMACNP (Global Encoder):

- Embedding Dimensions (r_dim, v_dim): categorical [64, 128, 256, 512]
- Location Encoder Dimension (w_dim): categorical [64, 128, 256, 512]
- Hidden Size: categorical [256, 512, 1024]
- Attention Heads: categorical [4, 8, 16]
- MLP Layers: categorical [2, 3]
- Dropout Rate: float [0.1, 0.3] with a step of 0.05
- Laplace Distance (p): categorical [1.0, 2.0]
- Learning Rate: float [1e-4, 5e-4] (log scale)

SMACNP (Localized Encoder):

- Embedding Dimensions (r_dim, v_dim): categorical [64, 128, 256, 512]
- Location Encoder Dimension (w_dim): categorical [64, 128, 256, 512]
- Hidden Size: categorical [256, 512, 1024]
- Attention Heads: categorical [4, 8, 16]
- MLP Layers: categorical [2, 3]
- Dropout Rate: float [0.1, 0.3] with a step of 0.05
- Localized Neighbors (k): categorical [15, 20, 25, 30]
- Laplace Distance (p): categorical [1.0, 2.0]
- Learning Rate: float [1e-4, 5e-4] (log scale)

Table B1 Optimal hyperparameters selected via TPE optimization for the Global and Localized SMACNP configurations.

Hyperparameter	SMACNP (Global)	SMACNP (Localized)
Embedding Dimensions		
Representation (r_dim)	128	64
Location (w_dim)	64	256
Variance (v_dim)	256	512
Hidden Size	512	512
Architecture		
Attention Heads	16	8
MLP Layers	2	3
Regularization & Training		
Dropout Rate	0.1	0.1
Layer Normalization	False	False
Learning Rate	4.93×10^{-4}	4.19×10^{-4}
Encoder-Specific		
Neighbors (k)	N/A	15
Distance (p)	1.0 (Manhattan)	1.0 (Manhattan)

Appendix C SMACNP Mathematical Formulation

C.1 Laplace Attention (Spatial Pathway)

The spatial pathway computes distance-based attention weights using a Laplace kernel with a configurable L_p norm:

$$w(\mathbf{s}^*, \mathbf{s}_i) = \text{softmax}_i\left(-\tau \cdot \| \text{MLP}(\mathbf{s}_i, y_i) - \text{MLP}(\mathbf{s}^*, \cdot) \|_p\right) \quad (A12)$$

where τ is a learnable temperature parameter and p selects between L_1 (Manhattan, $p = 1$) or L_2 (Euclidean, $p = 2$) distance. The weighted representation is:

$$\mathbf{w}^* = \sum_i w(\mathbf{s}^*, \mathbf{s}_i) \cdot \text{MLP}(\mathbf{s}_i, y_i) \quad (A13)$$

C.2 Multi-Head Cross-Attention (Attribute and Variance Pathways)

The attribute and variance pathways use standard Multi-Head Attention. For a single head, the operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (A14)$$

where d_k is the dimension of the keys. The input sets for Q, K, and V differ between the Global and Localized configurations and between the attribute and variance pathways, as described in Section 3.2 (Table). For the variance pathway, observations (y) are excluded from V in both configurations.

C.3 Output Distributions

The temperature head parameterizes a Gaussian distribution:

$$p\left(y_{\text{temp}} | \mathbf{s}^*, \mathbf{x}^*\right) = \mathcal{N}\left(y_{\text{temp}};\ \mu_{\text{temp}}(\mathbf{s}^*, \mathbf{x}^*),\ \sigma^2_{\text{temp}}(\mathbf{s}^*, \mathbf{x}^*)\right) \quad (A15)$$

The mean decoder uses a linear output activation; the variance decoder applies Softplus:

$$\sigma^2 = \log\left(1 + \exp(\cdot)\right) \quad (A16)$$

The precipitation head employs a hurdle model:

- Occurrence Model (Bernoulli):

$$p(y > 0 | \mathbf{s}^*, \mathbf{x}^*) = \sigma\left(\text{logit}(\mathbf{s}^*, \mathbf{x}^*)\right) \quad (A17)$$

  where σ is the sigmoid function.

- Amount Model (Log-Normal):

$$p(y | y > 0, \mathbf{s}^*, \mathbf{x}^*) = \text{LogNormal}\left(y;\ \mu_{\text{prec}}(\mathbf{s}^*, \mathbf{x}^*),\ \sigma^2_{\text{prec}}(\mathbf{s}^*, \mathbf{x}^*)\right) \quad (A18)$$

C.4 Loss Functions

The total training loss is a weighted combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{precip)}} \quad (A19)$$

Temperature loss *uses a* standard Gaussian negative log-likelihood:

$$\mathcal{L}_{\text{temp}} = \frac{1}{N}\sum_i \left[\frac{\log(\sigma_i^2)}{2} + \frac{(y_i - \mu_i)^2}{2\sigma_i^2}\right] \quad (A20)$$

Precipitation loss uses *a* weighted sum of the occurrence and amount components:

$$\mathcal{L}_{\text{precip}} = \lambda_{\text{occ}} \cdot \mathcal{L}_{\text{BCE}} + \lambda_{\text{amt}} \cdot \mathcal{L}_{\text{amount}} \quad (A21)$$

where:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_i \left[w_{\text{pos}} \cdot \mathbb{I}_i \cdot \log(p_i) + (1 - \mathbb{I}_i) \cdot \log(1 - p_i)\right] \quad (A22)$$

$$\mathcal{L}_{\text{amount}} = \frac{1}{N_{\text{rain}}} \sum_{i:\, y_i > 0} \left[\frac{\log(\sigma_i^2)}{2} + \frac{(\log(y_i + 1) - \mu_i)^2}{2\sigma_i^2} \right] \quad (A23)$$

Here $\mathbb{I}_i$ is an indicator function equal to 1 if precipitation is observed ($y_i > 0$), BCE denotes binary cross-entropy, and λ_{occ}, λ_{amt} are weights balancing the two components.

To address the imbalance between rainy and dry days, the BCE loss incorporates a positive class weight:

$$w_{\text{pos}} = \min\left(\frac{N_{\text{dry}}}{N_{\text{rainy}}},\ w_{\text{max}} \right) \quad (A24)$$

where w_{max} is a cap to prevent instability when very few rainy days are present in a batch.

Code and data availability

The source code developed for this study (SMACNP architectures: Global and Localized; Regression Kriging baseline; and training scripts) is publicly available on GitHub and archived on Zenodo: https://github.com/alexdum/climate-gridder; https://doi.org/10.5281/zenodo.18763498 (Dumitrescu, 2026). The homogenized daily air temperature and precipitation dataset (2020–2023) used for model development and evaluation, together with station metadata, is openly available via Zenodo at https://zenodo.org/records/14880417 (Dumitrescu, 2025).

Supplement link

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

Author contributions

AD is responsible for all aspects of this work, including conceptualization, methodology, software, validation, formal analysis, data curation, and writing – original draft preparation.

Competing interests

The author declares that there is no conflict of interest.

Disclaimer

Copernicus Publications adds a standard 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."

Please feel free to add disclaimer text at your choice, if applicable.

Acknowledgements

This research was supported by the project "Cross-sectoral Framework for Socio-Economic Resilience to Climate Change and Extreme Events in Europe (CROSSEU)" funded by the European Union Horizon Europe Programme, under Grant agreement n° 101081377.

The author used Grammarly to refine the phrasing and correct grammatical errors in the text. The author reviewed and takes full responsibility for the final content.

Financial support

This research has been supported by the European Union Horizon Europe Programme, under Grant agreement n° 101081377.

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, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., and Nauss, T.: Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania, Spat. Stat., 14, 91–113, https://doi.org/10.1016/j.spasta.2015.05.008, 2015.

Ashman, M., Diaconu, C., Langezaal, E., Weller, A., and Turner, R. E.: Gridded Transformer Neural Processes for Large Unstructured Spatio-Temporal Data, https://doi.org/10.48550/arXiv.2410.06731, 10 October 2024.

Bao, L.-L., Zhang, C.-X., Zhang, J.-S., and Guo, R.: A two-stage spatial prediction modeling approach based on graph neural networks and neural processes, Expert Syst. Appl., 258, 125173, https://doi.org/10.1016/j.eswa.2024.125173, 2024a.

Bao, L.-L., Zhang, J.-S., and Zhang, C.-X.: Spatial multi-attention conditional neural processes, Neural Netw., 173, 106201, https://doi.org/10.1016/j.neunet.2024.106201, 2024b.

Benjamin Murphy, Roman Yurchak, and Sebastian Müller: GeoStat-Framework/PyKrige: v1.7.3, , https://doi.org/10.5281/ZENODO.17372225, 2025.

Carr, A. and Wingate, D.: Graph Neural Processes: Towards Bayesian Graph Neural Networks, https://doi.org/10.48550/arXiv.1902.10042, 1 October 2019.

Daly, C., Neilson, R. P., and Phillips, D. L.: A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain, J. Appl. Meteorol., 33, 140–158, https://doi.org/10.1175/1520-0450(1994)033%253C0140:ASTMFM%253E2.0.CO;2, 1994.

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.

Diggle, P. J. and Ribeiro, P. J.: Model-based Geostatistics, Springer New York, New York, NY, https://doi.org/10.1007/978-0-387-48536-2, 2007.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://doi.org/10.48550/ARXIV.2010.11929, 2020.

Dumitrescu, A.: RoCliHom - Long-term homogenized air temperature and precipitation datasets in Romania, 1901 - 2023 (1.0.0), https://doi.org/10.5281/ZENODO.14880417, 2025.

Dumitrescu, A.: Climate Gridder: SMACNP and Regression Kriging for Climate Field Construction, , https://doi.org/10.5281/ZENODO.18763498, 2026.

Dumitrescu, A., Micu, D., Guijarro, J., Manea, A., and Cheval, S.: Long-term homogenized air temperature and precipitation datasets in Romania, 1901–2023, Sci. Data, 12, 1116, https://doi.org/10.1038/s41597-025-05371-4, 2025.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. A.: Conditional neural processes, in: International conference on machine learning, 1704–1713, 2018.

Gordon, J., Bruinsma, W. P., Foong, A. Y. K., Requeima, J., Dubois, Y., and Turner, R. E.: Convolutional Conditional Neural Processes, https://doi.org/10.48550/ARXIV.1910.13556, 2019.

Hengl, T., Heuvelink, G. B., and Rossiter, D. G.: About regression-kriging: from equations to case studies, Comput. Geosci., 33, 1301–1315, 2007.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas, Int. J. Climatol., 25, 1965–1978, https://doi.org/10.1002/joc.1276, 2005.

Iwase, K. and Takenawa, T.: Interpolation of mountain weather forecasts by machine learning, https://doi.org/10.48550/arXiv.2308.13983, 14 August 2024.

Jarvis, A., Reuter, H. I., Nelson, A., and Guevara, E.: Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT), 2008.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W.: Attentive Neural Processes, https://doi.org/10.48550/ARXIV.1901.05761, 2019.

Klein Tank, A. M. G., Zwiers, F. W., and Zhang, X.: Guidelines on Analysis of Extremes in a Changing Climate in Support of Informed Decisions for Adaptation, World Meteorological Organization, Geneva, 2009.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, N., Jetz, W., Schindler, K., and Wegner, J. D.: A high-resolution canopy height model of the Earth, Nat. Ecol. Evol., 7, 1778–1789, https://doi.org/10.1038/s41559-023-02206-6, 2023.

Li, J. and Heap, A. D.: Spatial interpolation methods applied in the environmental sciences: A review, Environ. Model. Softw., 53, 173–189, https://doi.org/10.1016/j.envsoft.2013.12.008, 2014.

Moldovan, G., Pinnington, E., Nemesio, A. P., Lang, S., Bouallègue, Z. B., Dramsch, J., Alexe, M., Cruz, M. S., Hahner, S., Cook, H., Theissen, H., Clare, M., O'Brien, C., Polster, J., Magnusson, L., Mertes, G., Pinault, F., Raoult, B., Rosnay, P. de, Forbes, R., and Chantry, M.: An update to ECMWF's machine-learned weather forecast model AIFS, https://doi.org/10.48550/arXiv.2509.18994, 23 September 2025.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Sekulić, A., Kilibarda, M., Protić, D., and Bajat, B.: A high-resolution daily gridded meteorological dataset for Serbia made by Random Forest Spatial Interpolation, Sci. Data, 8, 123, https://doi.org/10.1038/s41597-021-00901-2, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, https://doi.org/10.48550/ARXIV.1706.03762, 2017.

Vaughan, A., Tebbutt, W., Hosking, J. S., and Turner, R. E.: Convolutional conditional neural processes for local climate downscaling, Geosci. Model Dev., 15, 251–268, https://doi.org/10.5194/gmd-15-251-2022, 2022.