https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

Spatial Predictor Selection for Next-Day Minimum Temperature Forecasting: An Automated Machine Learning Framework Applied Across European Climate Regimes

Eric Duhamel
Independent researcher

Correspondence to: Eric Duhamel (edilia12380@gmail.com)

Abstract. Accurate prediction of near-surface air temperature remains a central challenge in geoscientific modeling, particularly when integrating high-dimensional spatial predictors derived from reanalysis datasets. While Model Output Statistics (MOS) approaches have been widely used, the systematic selection of spatially distributed predictors remains an open methodological issue.

This study proposes a genetic algorithm (GA) framework for automated predictor selection in daily minimum temperature forecasting. The method operates on spatially structured inputs derived from ERA5 reanalysis and is evaluated using observed temperature data from multiple European locations. The GA is designed to explore high-dimensional predictor spaces while controlling model complexity and ensuring compatibility with non-linear learning algorithms.

The approach is assessed using a one-day-ahead forecasting setup and compared against a LASSO-based baseline. Results show that the GA identifies compact predictor subsets that achieve predictive performance comparable to, or slightly better than, the baseline. Across test locations, mean absolute error values remain stable and indicate robust generalization.

Analysis of selected predictors highlights the existence of stable variable categories, although individual spatial selections exhibit variability across runs, reflecting the stochastic nature of the optimization process. These results suggest that predictor relevance should be interpreted in terms of distributions rather than fixed sets.

The proposed framework provides a flexible and reproducible approach to spatial feature selection in geoscientific applications. Its compatibility with complex models and high-dimensional inputs makes it a promising tool for improving forecasting systems based on reanalysis data. A key finding of this study is that spatial predictor selection is inherently non-unique, yet exhibits stable statistical structures at the variable level, suggesting that predictor relevance should be interpreted in probabilistic rather than deterministic terms.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

1 Introduction

1.1 Background and motivation

Accurate prediction of near-surface air temperature remains a central objective in meteorology, with applications ranging from operational forecasting to climate services. Among temperature variables, daily minimum temperature (Tmin) presents specific challenges due to its strong dependence on nocturnal processes such as radiative cooling, boundary layer decoupling, and local surface–atmosphere interactions (Stull, 1988; Oke, 1987). These processes often occur at spatial scales that are only partially resolved by numerical weather prediction (NWP) systems, leading to systematic forecast errors, particularly in complex terrain or coastal environments.

To address these limitations, statistical post-processing methods—commonly referred to as Model Output Statistics (MOS)—have long been used to improve forecast accuracy by learning empirical relationships between predictors and observed variables (Glahn and Lowry, 1972; Wilks, 2011). With the increasing availability of large-scale reanalysis datasets such as ERA5 (Hersbach et al., 2020), these approaches can now leverage spatially distributed atmospheric variables with high temporal and spatial consistency.

However, this abundance of data introduces a key methodological challenge: selecting relevant predictors from high-dimensional spatial fields. When multiple variables are available at thousands of grid points, the number of potential predictors grows combinatorially. Even within a limited geographic radius, the candidate predictor space can reach tens of thousands of variables, making exhaustive exploration infeasible.

Traditional feature selection methods provide only partial solutions to this problem. Filter-based approaches, such as correlation analysis or mutual information, evaluate predictors individually and may fail to capture interactions between variables (Guyon and Elisseeff, 2003). Embedded methods, such as LASSO regression (Tibshirani, 1996), enforce sparsity through regularization but remain inherently linear and may select correlated predictors that are suboptimal for non-linear models. More generally, many feature selection techniques are not explicitly designed to handle spatially structured data, where redundancy and autocorrelation play a central role (von Storch and Zwiers, 1999).

Recent advances in machine learning have demonstrated the potential of non-linear models, such as gradient boosting, for geoscientific applications (Chen and Guestrin, 2016; Ke et al., 2017; Reichstein et al., 2019). These models can capture complex interactions between predictors but remain sensitive to the quality and structure of the input feature set. As a result, the problem of feature selection remains critical, particularly in high-dimensional spatial contexts.

Despite the growing availability of high-resolution reanalysis data, the spatial dimension of predictor selection remains insufficiently formalized in most statistical forecasting frameworks. In practice, predictor selection is often limited to predefined locations or local variables, implicitly assuming that the most relevant information is located at or near the target site. However, atmospheric processes such as advection, cloud propagation, and large-scale circulation patterns suggest that relevant signals may originate from distant locations and propagate across space.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

This raises a fundamental methodological question: how to systematically explore spatial predictor configurations without imposing strong prior assumptions on location or variable importance. The difficulty is compounded by the strong spatial autocorrelation of atmospheric fields, which induces redundancy among neighboring grid points while preserving multiple alternative representations of the same physical signal. As a result, different spatial configurations may yield similar predictive performance, making the identification of a single "optimal" predictor set inherently ambiguous.

1.2 Objective and scope

The objective of this study is to develop and evaluate a methodology for automated selection of spatially distributed predictors for daily minimum temperature forecasting. Rather than focusing on maximizing predictive performance in an operational sense, the study aims to identify compact and informative subsets of predictors that capture the underlying structure of the problem. Beyond its application to temperature forecasting, the main contribution of this study is conceptual: it provides a structured interpretation of feature selection in spatially autocorrelated systems, highlighting its inherently non-unique nature and the relevance of probabilistic rather than deterministic perspectives on predictor importance.

More specifically, the proposed approach seeks to:

- explore high-dimensional spatial predictor spaces in a systematic manner,
- identify subsets of predictors that are both informative and parsimonious,
- assess the robustness of selected predictors across multiple runs and locations,
- evaluate the compatibility of selected predictors with non-linear regression models.

The analysis is conducted using a one-day-ahead forecasting framework and applied to multiple stations across Western Europe, covering a range of climatic conditions. This multi-site setup allows assessment of the generality of the proposed methodology.

This work is therefore positioned at the intersection of feature selection and spatial data analysis. Rather than seeking a unique optimal solution, the objective is to characterize the structure of the predictor space and to identify robust patterns that persist across different realizations of the selection process. This perspective shifts the focus from individual predictors to classes of predictors and from deterministic solutions to statistical regularities.

The main contribution of this work is to demonstrate that predictor selection in spatially structured atmospheric data is fundamentally non-unique, while still exhibiting robust statistical regularities. This perspective shifts the focus from identifying a single optimal predictor set to characterizing the structure of the predictor space.

1.3 Methodological approach

To address the combinatorial nature of the predictor selection problem, this study adopts a genetic algorithm (GA) as a stochastic optimization framework. Genetic algorithms are well-suited for high-dimensional feature selection problems, as

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

they explore the search space through population-based evolution and can capture interactions between predictors (Goldberg, 1989; Xue et al., 2016).

In the proposed framework, each candidate solution represents a subset of spatial predictors defined by variable–location pairs. The GA iteratively refines these subsets based on their predictive performance, evaluated through a regression model. This approach enables the identification of non-linear and spatially distributed predictor combinations that would be difficult to detect using deterministic or greedy methods.

To ensure computational tractability, a two-stage pipeline is employed. During the selection phase, a computationally efficient linear regression model is used to evaluate candidate predictor subsets. In a second stage, the selected predictors are assessed using more expressive models, including gradient boosting methods such as XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017).

The proposed method is compared with a baseline based on LASSO regression (Tibshirani, 1996), providing a reference for linear sparse feature selection. This comparison allows assessment of whether stochastic search offers advantages in identifying predictor sets suitable for non-linear models.

A conceptual overview of the framework is provided in Fig. 1.

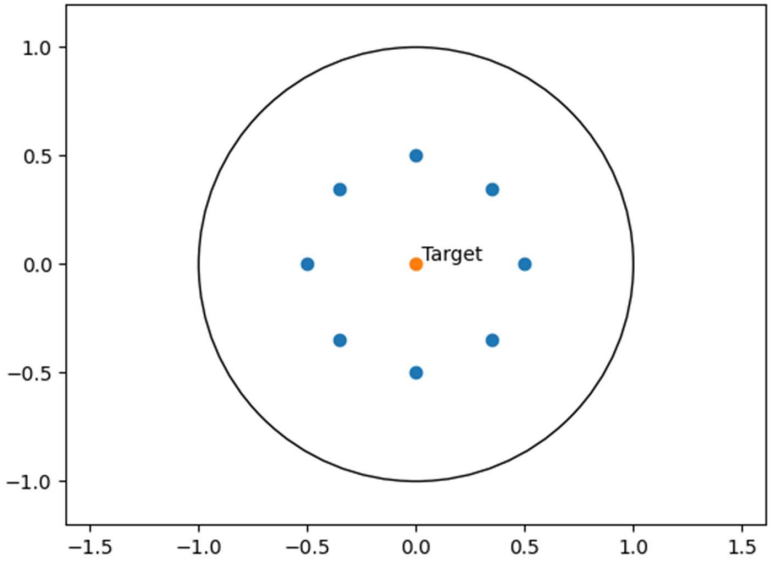

Figure 1. Conceptual illustration of the spatial predictor selection framework. The target site (red) is associated with candidate predictors defined as ERA5 grid points within a predefined search radius. Each predictor corresponds to a variable–location pair. The selection algorithm explores combinations of these spatial predictors to identify informative subsets.

Fig. 1 illustrates the definition of the spatial search domain.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

An important aspect of this approach is that the optimization process does not aim to identify a single globally optimal solution, but rather to explore a set of near-optimal solutions within a highly redundant search space. This distinction is critical in spatial contexts, where multiple predictor combinations can encode similar information due to the smoothness and coherence of atmospheric fields.

Consequently, the relevance of predictors is assessed not only through their contribution to predictive performance, but also through their consistency across independent runs. This approach allows the identification of stable structures within the predictor space, even when individual selections differ.

1.4 Data limitations and interpretation scope

The analysis relies on ERA5 reanalysis data (Hersbach et al., 2020) as predictor inputs and observational station data as target variables. While reanalysis datasets provide physically consistent and spatially complete representations of the atmosphere, they do not correspond to real-time forecast conditions. As a result, the predictive performance reported in this study should not be interpreted as representative of operational forecasting performance.

Instead, ERA5 is used as a controlled environment to investigate the structure of predictor relevance. The relationships identified between predictors and target variables reflect underlying atmospheric processes but may be affected by the characteristics of the reanalysis system.

In addition, the stochastic nature of the genetic algorithm implies that different runs may yield different predictor subsets with similar predictive performance. This non-uniqueness is a fundamental property of the problem, particularly in the presence of spatial autocorrelation. Consequently, the interpretation of results focuses on robust patterns—such as variable-level stability—rather than on individual predictor configurations.

2 Data

2.1 Study area and reference sites

The study focuses on Western Europe, covering latitudes 36.45°N to 63.16°N and longitudes 15.19°W to 18.44°E. This domain encompasses a range of climatic regimes, including Mediterranean, oceanic, and semi-continental conditions.

Eight reference sites were selected to represent this diversity (Table 1, Fig. 2), including both coastal and inland locations across France and the United Kingdom. This multi-site configuration enables evaluation of the proposed methodology under contrasted meteorological conditions.

Site	Lat. (°N)	Lon (°E)	Country	Climate type	Setting
Birmingham	52.42	-1.83	UK	Oceanic (Cfb)	Inland
Brest	48.44	-4.41	France	Oceanic (Cfb)	Coastal
Edinburgh	55.97	-3.21	UK	Oceanic (Cfb)	Coastal

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

Lyon	45.73	5.08	France	Semi-continental	Inland
Nice	43.65	7.21	France	Mediterranean	Coastal
Paris	48.72	2.38	France	Oceanic (Cfb)	Inland
Plymouth	50.35	-4.12	UK	Oceanic (Cfb)	Coastal
Strasbourg	48.55	7.64	France	Semi-continental	Inland

Table 1. Reference stations used in this study.

Figure 2 shows the geographic distribution of the selected stations within the ERA5 domain.

Figure 2. Geographic distribution of the eight reference sites across Western Europe. The background shows the ERA5 grid domain used for predictor extraction.

2.2 Observational data

Daily minimum temperature (Tmin) observations were obtained from the NOAA National Centers for Environmental Information (NCEI) Global Summary of the Day (GSOD) dataset (NCEI, 1999). The observation period spans 2004–2024, providing 7,671 daily values per station.

These observations are independent from the predictor dataset and serve as the target variable for model training and evaluation.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

2.3 ERA5 predictor data

Predictor variables are derived from the ERA5 reanalysis produced by the European Centre for Medium-Range Weather Forecasts. ERA5 provides globally complete and physically consistent atmospheric fields at a spatial resolution of 0.25° × 0.25°.

For the selected domain, this corresponds to approximately 14,000 grid points. ERA5 data are used to construct spatially distributed predictors representing atmospheric conditions surrounding each reference site.

2.4 Predictor definition

A predictor is defined as a triplet (variable, latitude, longitude) associated with a daily time series. The candidate predictor space is therefore given by the Cartesian product of variables and spatial grid points within a predefined radius around each station.

Depending on latitude, the number of candidate predictors ranges from approximately 41,000 to 53,000 per station.

2.5 Predictor variables

A total of 24 ERA5 variables were selected to represent the main physical processes influencing minimum temperature, including near-surface temperature, radiation fluxes, boundary layer properties, wind components, and soil conditions. These variables are listed in Table 2.

The previous day's observed Tmin at the target station is also included as a candidate predictor to account for temporal persistence.

No attempt is made to pre-filter variables based on physical assumptions; the selection process is entirely driven by the optimization procedure.

2.6 Data processing and storage

All datasets were integrated into a unified SQLite database to facilitate efficient data access and reproducibility. ERA5 data (NetCDF format) and observational data (CSV format) were harmonized into a consistent structure indexed by date and location.

This design enables direct querying of predictor time series and simplifies the implementation of the selection algorithm.

2.7 Evaluation framework

The objective of the study is to identify informative predictor subsets rather than to optimize absolute predictive performance. Model evaluation is therefore based on comparative metrics.

The mean absolute error (MAE) is used to rank predictor configurations: $MAE = \frac{1}{N} sum_{i=1}^{N} \left| T_{obs,i} - T_{pred,i} \right|$

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

Given the stochastic nature of the genetic algorithm, different runs may produce different predictor subsets with similar performance. As a result, the analysis focuses on the stability of performance and the structural properties of selected predictors rather than on individual configurations.

3 Methods

This section describes the predictor selection framework, including the definition of the search space, the genetic algorithm used for feature selection, and the evaluation protocol. The objective is to identify informative spatial predictors rather than to optimize predictive performance per se; the mean absolute error (MAE) is therefore used as a comparative criterion to guide the search.

3.1 Temporal configuration

The temporal depth of predictor information was evaluated by testing three configurations: (i) a single-day window (current day), (ii) a two-day window (current day + lag 1), and (iii) a three-day window (current day + lags 1–2).

Experiments were conducted on four representative stations (Birmingham, Brest, Nice, Strasbourg) using five independent runs per configuration. Performance was assessed on the test dataset (2023–2024) using XGBoost.

Results (Table 3, Fig. 3) show that the single-day configuration systematically yields the lowest MAE across all stations. Differences between one- and two-day windows remain within inter-run variability, while three-day windows consistently degrade performance.

Station	1-day	2-day	3-day
Birmingham	0.91 ± 0.02	0.92 ± 0.01	0.92 ± 0.02
Brest	1.32 ± 0.02	1.35 ± 0.01	1.36 ± 0.03
Nice	0.80 ± 0.01	0.81 ± 0.01	0.83 ± 0.02
Strasbourg	1.29 ± 0.02	1.29 ± 0.02	1.30 ± 0.02

Table 3. MAE (°C) for XGBoost according to temporal window size (mean ± std dev. over 5 runs).

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

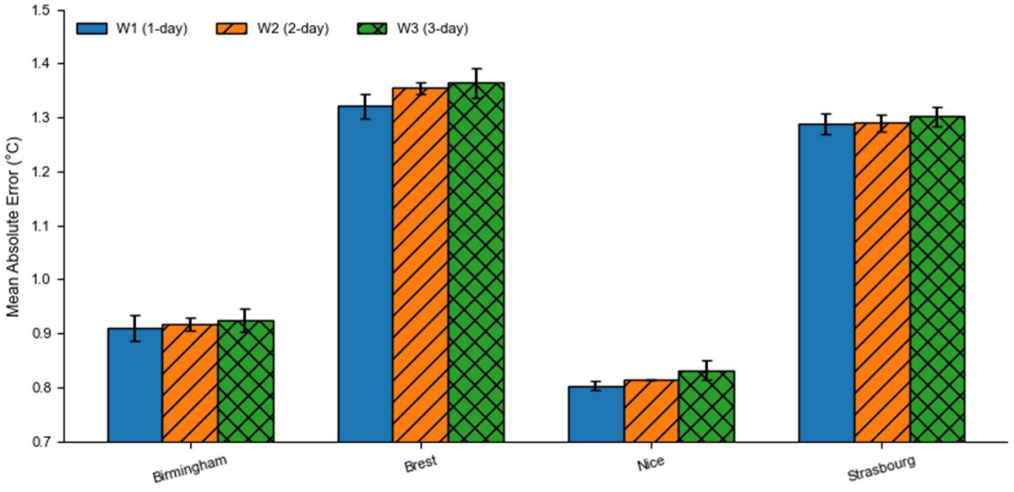

Figure 3. Mean absolute error (MAE, °C) of next-day minimum temperature forecasts obtained with the XGBoost model for three temporal window sizes (W1–W3) across four stations. Error bars represent ± one standard deviation computed over five random seeds.

Based on these results, the single-day temporal configuration was adopted for all subsequent analyses. This choice is consistent with the principle of parsimony and suggests that the relevant atmospheric information for next-day minimum temperature forecasting is largely contained in the current-day state.

3.2 Predictor search space

For each station, candidate predictors were defined as all combinations of ERA5 variables and grid points within a radius of 540 km. The set of candidate predictors is defined as the Cartesian product of the variable set and the spatial grid:

$$P = \{p = (v, x, y) \mid v \in V,\ (x, y) \in G\}$$

The 540 km radius was selected to capture synoptic-scale influences while maintaining computational tractability.

Due to the latitude–longitude grid structure of ERA5, the number of candidate predictors varies with latitude, ranging from approximately 41,000 to 53,000 per station. Despite this variability, all stations sample a comparable physical area.

In addition to ERA5 variables, the previous day's observed Tmin at the target station is included as a candidate predictor, allowing the algorithm to exploit temporal persistence when relevant.

3.3 Genetic algorithm for predictor selection

Given the size of the search space, exhaustive exploration is infeasible. Predictor selection is therefore formulated as a combinatorial optimization problem solved using a genetic algorithm (GA).

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

Each individual represents a subset of predictors (chromosome), where each gene corresponds to a specific (variable, latitude, longitude) triplet. Each candidate solution corresponds to a subset of predictors of fixed size: $S \subseteq P, | S | = k$

The fitness of an individual is defined as the MAE obtained on a validation dataset using a linear regression model trained on the selected predictors.

The GA operates with a population of candidate solutions evolved through selection, crossover, and mutation. The process is terminated when no improvement in validation MAE is observed over a predefined number of generations.

Due to its stochastic nature, the GA produces different predictor subsets across runs. To account for this variability, all experiments are repeated with multiple random seeds, and results are analyzed in terms of performance stability and structural properties of the selected predictors.

The choice of a genetic algorithm for feature selection is motivated by the combinatorial structure of the predictor space. Here, candidate predictors are defined as variable–location pairs within a spatial domain, leading to tens of thousands of potential features per station. The number of possible subsets grows exponentially with the number of candidates: even for a moderate subset size, the number of combinations becomes computationally intractable, making exhaustive search impossible.

In this framework, each candidate solution is represented as a subset of predictors (chromosome), where each predictor corresponds to a specific variable–location pair. The objective is to identify subsets that minimize prediction error while maintaining a limited number of predictors.

Alternative feature selection approaches present limitations in this context. Filter methods evaluate predictors independently and may fail to capture interactions between variables. Embedded methods, such as LASSO regression, enforce sparsity within a linear framework but may retain correlated predictors that are not optimal for non-linear models. More generally, these approaches do not explicitly address the combinatorial nature of subset selection in spatially structured data.

Genetic algorithms provide a flexible framework for exploring such high-dimensional search spaces. By evolving populations of candidate subsets through stochastic operators, they allow interactions between predictors to be evaluated implicitly via the fitness function. This is particularly relevant in spatial contexts, where multiple predictors may jointly encode similar atmospheric signals.

The use of a linear regression model as the fitness function reflects a trade-off between computational efficiency and representational capacity. Evaluating candidate subsets with non-linear models such as gradient boosting would be computationally prohibitive given the number of evaluations required during the search process. Linear regression provides a fast and stable proxy for predictor relevance, enabling efficient exploration of the search space, while final evaluation with non-linear models ensures that predictive performance is assessed in a more expressive framework. This apparent mismatch is intentional and reflects a trade-off between computational tractability during the selection phase and representational capacity during the evaluation phase.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

3.4 Predictor count sensitivity

The impact of predictor subset size was evaluated by testing configurations ranging from 50 to 150 predictors. Experiments were conducted on four stations using five independent runs per configuration.

Results (Table 4, Fig. 5) indicate that predictive performance improves with increasing predictor count up to approximately 90 predictors, beyond which gains become marginal and comparable to inter-run variability.

Station	50	70	90	110	130	150
Birmingham	0.929 ± 0.017	0.925 ± 0.024	0.91 ± 0.023	0.897 ± 0.012	0.907 ± 0.014	0.891 ± 0.006
Brest	1.365 ± 0.003	1.346 ± 0.018	1.321 ±0.023	1.324 ± 0.024	1.316 ± 0.014	1.294 ± 0.012
Nice	0.831 ± 0.033	0.813 ± 0.011	0.803 ± 0.009	0.804 ± 0.006	0.805 ± 0.013	0.799 ± 0.011
Strasbourg	1.321 ± 0.015	1.336 ± 0.029	1.288 ± 0.02	1.284 ± 0.009	1.274 ± 0.013	1.285 ± 0.022

Table 4. Mean absolute error (°C) by station and number of predictors (XGBoost model).

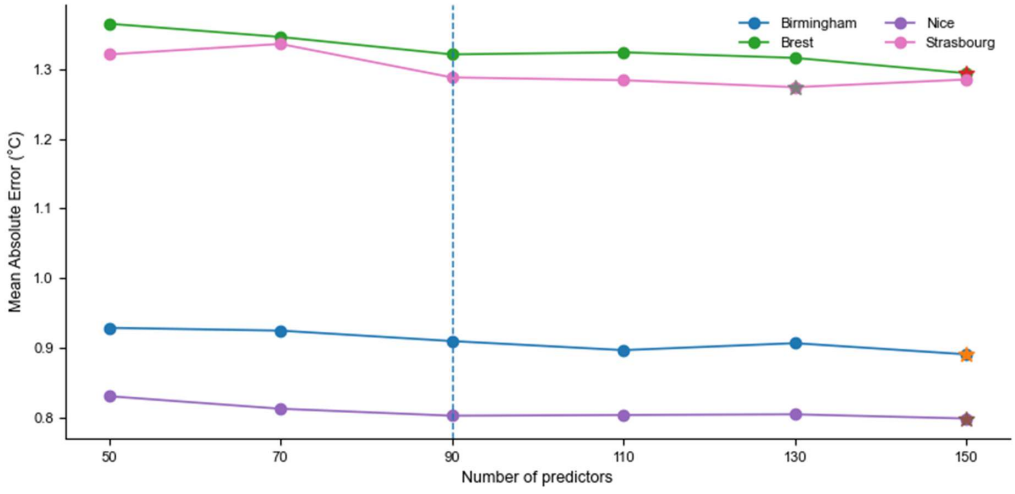

Figure 4. Sensitivity of prediction accuracy to predictor subset size (XGBoost model, test period 2023–2024). The vertical dashed line indicates the selected configuration (90 predictors); stars denote best performance per station.

Although the best performance is achieved with larger subsets (130–150 predictors), the improvement remains within inter-run variability and does not justify the additional increase in model complexity. The 90-predictor configuration captures most of the predictive signal while maintaining a lower dimensionality. This configuration was therefore adopted as a reference for comparative analyses.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

3.5 Model evaluation framework

Selected predictor subsets are evaluated using three regression models: linear regression, LightGBM, and XGBoost. XGBoost consistently provides the best predictive performance and is used as the primary model for reporting results.

The dataset (2004–2024) is partitioned chronologically into training (2004–2020), validation (2020–2022), and test (2023–2024) periods (Fig. 6). This separation ensures that test results reflect out-of-sample performance.

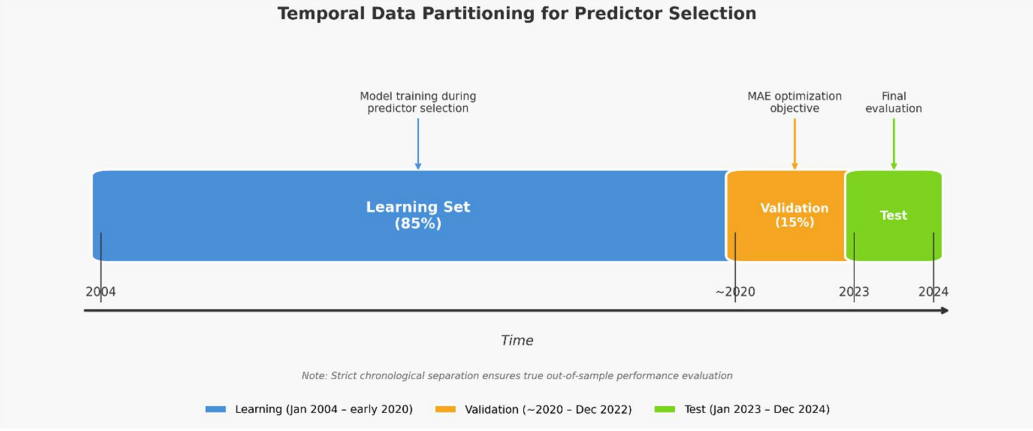

Figure 5. Schematic diagram of temporal data partitioning. The chronological split ensures strict separation between training, validation, and test periods.

Baseline models, including persistence and climatology, are used for contextual evaluation (see Section 4.1). In addition, LASSO regression is used as a benchmark for linear feature selection (see Section 4.2).

3.6 Reproducibility and computational setup

All experiments are conducted within a containerized environment using Docker to ensure reproducibility. The execution environment includes fixed dependency versions and controlled numerical settings.

To ensure reproducibility of individual runs, numerical libraries are constrained to single-thread execution, and all experiments are performed with fixed random seeds. Under these conditions, identical runs produce identical results across different hardware platforms.

The full computational workflow, including data processing, model training, and predictor selection, is designed to be reproducible from the provided code and dataset (see Code and data availability section).

All experiments involving the genetic algorithm were systematically repeated using five fixed random seeds (1, 2, 3, 4, 5). These seeds control all stochastic components of the optimization process, including initialization, crossover, and mutation. This ensures that all reported results can be exactly reproduced.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

4 Results

4.1 Overall performance

The proposed framework was applied to eight stations spanning diverse European climate regimes. Table 5 summarizes predictive performance on the independent test period (2023–2024) using XGBoost with 90 selected predictors.

Station	Climate	MAE (°C)	R^2	Bias (°C)	≤1°C (%)	≤2°C (%)	P95 (°C)
Nice	Mediterranean	0.80	0.972	-0.09	69.1	93.7	2.13
Birmingham	Oceanic	0.91	0.938	-0.03	62.7	91.5	2.23
Plymouth	Oceanic	1.07	0.907	+0.00	53.6	86.3	2.64
Paris	Semi-continental	1.19	0.925	-0.33	51.0	82.3	2.95
Lyon	Semi-continental	1.23	0.944	-0.21	49.5	81.1	3.13
Edinburgh	Oceanic	1.24	0.891	-0.04	49.8	79.6	3.13
Strasbourg	Continental	1.29	0.932	-0.03	46.5	79.9	3.27
Brest	Oceanic	1.32	0.849	+0.14	46.3	77.6	3.31

Table 5. Performance metrics on the independent test set (2023–2024, XGBoost with 90 predictors; averages over 5 GA runs).

MAE values range from 0.80°C (Nice) to 1.32°C (Brest), with intermediate performance for other stations. Coefficients of determination (R^2) exceed 0.84 for all sites, indicating that a substantial fraction of variance is explained by the selected predictors.

Performance differences across stations reflect variations in predictability rather than methodological limitations. In particular, sites with lower MAE tend to exhibit higher temporal persistence and lower variability.

Comparison with baseline methods (Table 6) shows that the proposed approach reduces MAE by 36–54% relative to persistence and by 52–64% relative to climatology. These results demonstrate that incorporating spatial predictors provides substantial added value compared to purely local or univariate approaches.

Station	Model	Persistence	Climatology	vs Pers.	vs Clim.
Nice	0.80	1.25	1.93	-36%	-59%
Birmingham	0.91	1.99	2.56	-54%	-64%
Plymouth	1.07	2.00	2.46	-47%	-57%
Paris	1.19	2.24	2.97	-47%	-60%
Lyon	1.23	2.25	2.98	-45%	-59%
Edinburgh	1.24	2.29	2.59	-46%	-52%
Strasbourg	1.29	2.42	3.09	-47%	-58%
Brest	1.32	2.40	2.73	-45%	-52%

Table 6. MAE improvement (%) versus baselines

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

Seasonal performance (Table 7) shows moderate variability across stations, with no consistent pattern across all sites. Differences between seasons remain within a limited range relative to overall error levels, suggesting that the method maintains stable performance across varying meteorological conditions.

Station	Winter	Spring	Summer	Autumn	Most difficult
Nice	**0.86**	0.79	0.75	0.81	Winter
Birmingham	0.89	0.98	0.78	**0.99**	Autumn
Plymouth	1.09	1.04	1.01	**1.16**	Autumn
Paris	**1.30**	1.19	1.15	1.13	Winter
Lyon	**1.47**	1.21	1.06	1.20	Winter
Edinburgh	1.13	1.29	1.15	**1.39**	Autumn
Strasbourg	1.24	**1.53**	1.09	1.29	Spring
Brest	1.38	1.25	1.22	**1.44**	Autumn

Table 7. Seasonal MAE (°C) for each station. Bold values indicate the most challenging season.

Beyond aggregate performance metrics, the variability observed across stations provides insight into the structure of the prediction problem. Differences in MAE between sites do not appear to reflect inconsistencies in the methodology, but rather variations in the intrinsic predictability of local temperature dynamics.

In particular, stations exhibiting lower prediction errors tend to be associated with stronger temporal coherence and smoother spatial gradients in the predictor fields. In such contexts, the information relevant for next-day temperature forecasting is more easily captured by a limited set of predictors. Conversely, higher error levels are observed at sites where atmospheric conditions exhibit greater variability or weaker spatial coherence, making the identification of stable predictive relationships more challenging.

These differences suggest that the effectiveness of spatial predictor selection is partly conditioned by the structure of the underlying data. When relevant atmospheric signals are spatially coherent, multiple predictor configurations can capture similar information, leading to stable performance across runs. In contrast, when signals are more heterogeneous or less spatially organized, the selection process becomes more sensitive to stochastic variability, resulting in slightly reduced performance and greater dispersion across runs.

This interpretation is consistent with the observation that the genetic algorithm yields comparable performance across different predictor subsets, despite variations in their spatial composition. It reinforces the view that predictive skill is driven by the presence of informative structures in the data rather than by the identification of a unique optimal predictor configuration.

4.2 Comparison with LASSO-based feature selection

To assess the relevance of the proposed feature selection approach, results obtained with the genetic algorithm are compared with a baseline based on LASSO regression, which performs feature selection through L1 regularization.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

Two configurations are considered:

(i) LASSO constrained to select 90 predictors (LASSO-90), and

(ii) unconstrained LASSO selection (LASSO-full).

In both cases, the selected predictors are used as inputs to the same XGBoost model, ensuring a consistent evaluation framework.

Since both feature selection methods are evaluated using the same downstream model, the comparison may favor predictor subsets that are well suited to XGBoost. However, this design ensures a controlled comparison focused on feature selection rather than model choice.

Performance at equivalent sparsity

At identical dimensionality (90 predictors), the genetic algorithm yields lower MAE than LASSO-90 for all sites except Paris, where both methods achieve identical performance (Table 8). The average improvement is approximately 2.9%.

This comparison is performed under strictly controlled conditions, as both methods operate with the same number of predictors. Performance differences can therefore be attributed to differences in feature selection rather than model capacity.

The observed gains remain moderate but consistent across sites, and are obtained under strict sparsity constraints, which is relevant in operational contexts where model compactness is required. This may be explained by the ability of the genetic algorithm to retain predictors involved in non-linear interactions, although this mechanism is not explicitly quantified in the present study.

Table 8. Performance at equivalent sparsity (90 predictors, XGBoost model)

Site	GA+XGB (90 feat.)	LASSO-90	LASSO-90+XGB	Difference	Winner
Birmingham	0.91	0.94	0.98	+7.7%	GA
Brest	1.32	1.4	1.34	+1.5%	GA
Edinburgh	1.24	1.25	1.25	+0.8%	GA
Lyon	1.23	1.18	1.26	+2.4%	GA
Nice	0.80	0.79	0.81	+1.3%	GA
Paris	1.19	1.20	1.19	+0.0%	Tie
Plymouth	1.07	1.11	1.09	+1.9%	GA
Strasbourg	1.24	1.32	1.33	+7.3%	GA
Mean	1.125	—	1.1563	+2.8%	GA

Performance with unconstrained LASSO

When LASSO is allowed to select predictors without constraint, the number of selected variables increases substantially, ranging from 346 to 1191 depending on the site (Table 9).

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

The substantial increase in the number of predictors in the LASSO-full configuration may impact training time, storage requirements, and interpretability, particularly in large-scale applications.

In this configuration, performance becomes very similar to that of the genetic algorithm, with a mean MAE difference of approximately 0.5%. LASSO slightly outperforms the genetic algorithm at some sites (e.g. Brest, Lyon, Paris), while the opposite is observed at others (e.g. Nice).

These results indicate that both approaches capture a comparable predictive signal when no constraint is imposed on the number of predictors.

However, this comparable performance is obtained with a much larger number of variables in the LASSO case, which may have implications in terms of model complexity, computational cost, and interpretability. The main advantage of the genetic algorithm is therefore not raw predictive performance, but its ability to identify compact predictor subsets under strict sparsity constraints.

Table 9. Performance comparison (GA vs LASSO-full, XGBoost model)

Site	GA+XGB	LASSO-full + XGB	Features (GA/LASSO)	GA advantage
Birmingham	0.91	0.92	90/855	+1.1%
Brest	1.32	1.31	90/915	-0.8%
Edinburgh	1.24	1.24	90/675	0.0%
Lyon	1.23	1.25	90/738	-0.8%
Nice	0.80	0.82	90/346	+2.5%
Paris	1.19	1.17	90/694	-0.8%
Plymouth	1.07	1.08	90/1026	+0.9%
Strasbourg	1.24	1.26	90/431	+1.6%
Mean	1.125	1.131	90/774	+0.53%

Discussion

Overall, the results show that the genetic algorithm and LASSO lead to similar predictive performance when LASSO is allowed to select a large number of predictors.

Under sparsity constraints, however, the genetic algorithm consistently achieves better performance, indicating a more efficient use of the available predictors.

This suggests that both methods identify a similar underlying predictive signal, but differ in how compactly this information is represented. The genetic algorithm tends to produce more compact subsets, whereas LASSO relies on larger predictor sets to reach comparable performance.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

It should be noted that this comparison is conducted within a unified modeling pipeline, where both feature selection methods are evaluated using the same downstream model. The results therefore reflect differences in feature selection strategies rather than fully optimized end-to-end configurations for each method.

LASSO hyperparameters were selected using standard cross-validation procedures. However, the objective of this comparison is not to fully optimize each method independently, but to evaluate feature selection strategies within a unified modeling framework.

4.3 Stability and structure of selected predictors

This analysis is based on the five independent runs defined by the fixed random seeds described in Section 3.6.

Due to the stochastic nature of the genetic algorithm, individual runs produce different predictor subsets. However, this variability primarily affects spatial locations rather than variable types.

Across repeated runs, most ERA5 variables are consistently selected, indicating that predictor relevance is stable at the variable level. In contrast, the exact spatial positions of selected predictors vary substantially, reflecting the spatial autocorrelation structure of atmospheric fields.

This distinction suggests that predictor selection should be interpreted in terms of distributions or classes of variables rather than fixed spatial configurations. Multiple spatially distinct predictor sets can provide comparable predictive performance.

A complementary analysis was conducted to examine the distribution of selected predictors across variable types. Results show that certain categories, particularly near-surface temperature and radiation-related variables, are consistently represented among the most frequently selected predictors across stations and runs.

While the exact spatial locations of selected predictors vary, the relative importance of variable types remains stable. This suggests that the predictive signal is primarily associated with broad physical processes rather than specific geographic points.

Although a detailed variable-level comparison with LASSO-based feature selection was not performed, the comparable predictive performance of both approaches suggests that they capture a similar core signal, potentially through different predictor configurations.

4.4 Predictor importance and dimensionality

Analysis of feature importance (based on XGBoost gain) shows that predictive performance is typically dominated by a small subset of variables, primarily near-surface temperature-related predictors.

A limited number of predictors accounts for the majority of total gain, while the remaining predictors contribute marginally. This concentration of importance indicates that the effective dimensionality of the problem is significantly lower than the nominal number of selected predictors.

This result is consistent across stations and supports the interpretation that the genetic algorithm identifies redundant but interchangeable predictors within spatially correlated fields.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

4.5 Synthesis

The results highlight three main findings:

- Spatially distributed predictors provide substantial improvements over baseline approaches.
- Different feature selection methods (genetic algorithm and LASSO) identify comparable predictive structures, with the genetic algorithm offering a more compact representation.
- Predictor selection is inherently non-unique due to spatial autocorrelation, and should be interpreted in terms of stable variable classes rather than fixed predictor sets.

These findings support the use of stochastic search methods for exploring high-dimensional spatial predictor spaces, while emphasizing the importance of robustness analysis in interpreting the results.

5 Discussion

5.1 Interpretation of predictor selection results

The results highlight a fundamental property of spatial predictor selection in atmospheric data: the non-uniqueness of optimal predictor configurations. While individual runs of the genetic algorithm produce different subsets of predictors, these subsets exhibit similar predictive performance and share common structural characteristics.

This behavior can be explained by the strong spatial autocorrelation of atmospheric fields (von Storch and Zwiers, 1999). Neighboring grid points often contain highly redundant information, allowing multiple alternative representations of the same underlying physical processes. As a result, different predictor combinations can encode equivalent predictive signals, leading to a multiplicity of near-optimal solutions.

This finding has important methodological implications. In high-dimensional spatial contexts, feature selection should not be interpreted as identifying a single optimal set of predictors, but rather as characterizing a space of equivalent solutions. Consequently, robustness analysis across multiple runs becomes essential for distinguishing stable patterns from stochastic variability.

At the variable level, the results indicate a high degree of consistency, with most ERA5 variables being selected across runs. This indicates that predictor relevance is primarily determined by variable type rather than precise spatial location. Spatial variability in the selected predictors reflects the interchangeable nature of nearby grid points rather than instability in the underlying signal.

5.2 Comparison with linear feature selection methods

The comparison with LASSO-based feature selection provides additional insight into the structure of the predictor space. Both approaches identify predictor sets that achieve comparable predictive performance, indicating that a common core signal is captured regardless of the selection method.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

However, differences emerge in terms of sparsity and structure. The genetic algorithm tends to produce more compact predictor sets, while LASSO may retain a larger number of correlated features. This difference can be attributed to the optimization mechanisms of the two methods: LASSO operates through coefficient shrinkage in a linear framework, whereas the genetic algorithm evaluates subsets of predictors as a whole.

These results suggest that stochastic search methods may be better suited for identifying predictor subsets that are compatible with non-linear models. In particular, the subset-based evaluation inherent to the genetic algorithm allows interactions between predictors to be taken into account during selection, even when the fitness function itself is based on a linear model.

At the same time, the relatively small performance differences between methods indicate that the choice of feature selection technique may be less critical than the definition of the predictor space itself. Ensuring that relevant variables and spatial domains are included appears to be a prerequisite for effective model performance.

5.3 Limitations

Several limitations of the present study should be acknowledged.

First, the use of ERA5 reanalysis data as predictors provides an idealized representation of atmospheric conditions. While this allows for a controlled analysis of predictor relevance, it does not reflect the uncertainties and biases associated with operational numerical weather prediction systems. As a result, the absolute performance levels reported here cannot be directly translated to operational forecasting contexts. This limitation is inherent to the use of reanalysis data and highlights the need for future validation using operational forecast inputs. In addition, a comparison with models trained on the full predictor set without prior feature selection would provide further insight into the relative contribution of feature selection versus model capacity.

Second, the genetic algorithm relies on a stochastic search process, which introduces variability in the selected predictor subsets. Although this variability is partly addressed through multiple runs, it also reflects the intrinsic non-uniqueness of the problem. The interpretation of results must therefore focus on robust patterns rather than individual selections.

Third, the fitness function used during the selection phase is based on linear regression for computational efficiency. While this choice enables exploration of large search spaces, it may bias the selection process toward predictors that perform well under linear assumptions. However, the subsequent evaluation with non-linear models partially mitigates this limitation.

Finally, the study focuses on a specific temporal configuration (one-day-ahead forecasting) and a limited set of stations. Although the results are consistent across the selected sites, further validation would be required to assess the generality of the approach in different climatic regions and forecasting horizons.

5.4 Implications for geoscientific modelling

The findings of this study have broader implications for feature selection in geoscientific applications.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

First, they highlight the importance of considering spatial structure explicitly in predictor selection. Ignoring spatial information or restricting predictors to local variables may overlook relevant signals that originate from larger-scale atmospheric processes.

Second, the results emphasize the need to account for redundancy in spatial datasets. High-dimensional predictor spaces derived from gridded data inherently contain correlated variables, and effective feature selection methods must be able to navigate this redundancy without overfitting.

Third, the observed non-uniqueness of predictor selection suggests that interpretability should be approached with caution. Rather than attributing physical meaning to individual predictors, it may be more appropriate to interpret results in terms of variable classes or aggregated patterns.

Finally, the combination of stochastic search methods with machine learning models provides a flexible framework for exploring complex predictor spaces. This approach is not limited to temperature forecasting and could be extended to other geoscientific variables, such as precipitation, wind, or air quality indicators.

5.5 Perspectives

Several directions for future work can be identified.

One important extension would be to apply the proposed methodology to operational forecast data, allowing direct assessment of its performance in real-world forecasting conditions. This would provide insight into how the selected predictor structures transfer from reanalysis-based environments to predictive systems.

Another avenue concerns the integration of additional constraints into the selection process, such as spatial coherence or physical interpretability criteria. Incorporating such constraints could help reduce variability in the selected predictors and improve interpretability.

Finally, further investigation of the relationship between feature selection and model architecture would be valuable. In particular, exploring how different machine learning models interact with selected predictor subsets could provide a more comprehensive understanding of the role of feature selection in geoscientific forecasting.

5.6 Generalization and methodological implications

The results obtained in this study extend beyond the specific case of minimum temperature forecasting and point to more general properties of feature selection in high-dimensional spatial datasets.

First, they illustrate that in the presence of strong spatial autocorrelation, feature selection problems are inherently underdetermined. Multiple predictor subsets can provide equivalent predictive performance, as they encode similar information through different spatial configurations. This challenges the common assumption that feature selection should identify a single optimal subset, and instead supports a probabilistic or ensemble-based interpretation of predictor relevance.

Second, the study highlights the importance of distinguishing between variable-level relevance and spatial-level specificity. While variable types exhibit strong stability across runs, spatial selections remain flexible, suggesting that the predictive signal

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

is distributed rather than localized. This distinction is particularly relevant for geoscientific applications, where spatial redundancy is a fundamental property of the data.

Third, the combination of stochastic search methods with deterministic learning algorithms provides an effective strategy for navigating high-dimensional predictor spaces. The genetic algorithm enables exploration of the combinatorial structure of the feature space, while the regression model provides a consistent evaluation criterion. This separation between exploration and evaluation may be applicable to a wide range of problems involving structured data.

Finally, these findings suggest that future developments in feature selection for geoscientific applications may benefit from explicitly incorporating spatial structure and uncertainty. Approaches based on ensembles of predictor subsets, or on probabilistic representations of feature importance, could provide more robust and interpretable solutions than methods relying on a single deterministic selection.

6 Conclusions

This study introduces a framework for automated selection of spatial predictors in the context of daily minimum temperature forecasting. The approach relies on a genetic algorithm to explore high-dimensional predictor spaces derived from ERA5 reanalysis data and identifies compact subsets of informative predictors.

Results show that spatially distributed predictors substantially improve predictive performance compared to baseline approaches. The comparison with LASSO-based feature selection indicates that similar predictive structures can be identified using different methods, with the genetic algorithm providing more compact predictor sets at comparable performance levels.

A key result of this study is that spatial predictor selection is inherently non-unique: while individual spatial selections vary across runs, the relevance of variable types remains stable. This finding suggests that predictor importance should be interpreted in terms of robust statistical structures rather than fixed configurations, with implications for feature selection in high-dimensional geoscientific datasets.

The proposed framework is generic and applicable to other geoscientific forecasting problems involving high-dimensional spatial data. Its design also supports integration with non-linear models commonly used in operational contexts. More broadly, this study demonstrates that in spatially autocorrelated systems, feature selection is not an identification problem but a characterization problem.

Code and data availability

The exact version of the code used in this study is archived on Zenodo (Duhamel, 2026a, https://doi.org/10.5281/zenodo.19570174) and is publicly available. This repository contains the full computational workflow, including the genetic algorithm implementation, data processing pipeline, and model training scripts required to reproduce the experiments.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

The dataset used for training and evaluation, consisting of ERA5 reanalysis variables and observational temperature records, is also archived on Zenodo (Duhamel, 2026b, https://doi.org/10.5281/zenodo.19401697). It is provided as a precomputed SQLite database (≈12 GB), which constitutes the primary entry point for reproducibility. Scripts allowing reconstruction of the dataset from the original sources are included in the code repository; however, due to the computational cost of reconstruction, users are encouraged to rely on the archived dataset.

The execution environment is fully containerized using Docker and is available at:
https://hub.docker.com/r/grezac/edilia/tags (tag: v1.0)

The Docker image ensures consistent dependencies and enables full reproducibility of the experiments across platforms. All Python dependencies are additionally specified in a requirements.txt file included in the code repository.

The original data sources are publicly available:

- ERA5 reanalysis data from the Copernicus Climate Data Store
- Observational data from the NOAA NCEI Global Summary of the Day

Licensing information for all external datasets is provided in the code repository.

Author contributions

The author designed the study, developed the methodology, implemented the code, performed the analyses, and wrote the manuscript.

Competing interests

The author declares that there is no conflict of interest.

Acknowledgements

The author acknowledges the European Centre for Medium-Range Weather Forecasts for providing ERA5 reanalysis data through the Copernicus Climate Change Service, and the National Centers for Environmental Information for maintaining the observational datasets used in this study.

The author used an AI-based language model (ChatGPT) to assist in drafting and improving the clarity of the manuscript. The use of this tool was limited to language and presentation. All scientific content, methodology, and conclusions are the sole responsibility of the author.

Financial support

This research received no external funding.

https://doi.org/10.5194/egusphere-2026-2136
Preprint. Discussion started: 22 June 2026

References

Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Duhamel, E.: EDILIA code repository (version 1.0), Zenodo, https://doi.org/10.5281/zenodo.19570174, 2026a.

Duhamel, E.: EDILIA dataset (precomputed SQLite database), Zenodo, https://doi.org/10.5281/zenodo.19401697, 2026b.

Glahn, H. R. and Lowry, D. A.: The use of Model Output Statistics (MOS) in objective weather forecasting, Journal of Applied Meteorology, 11, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203>2.0.CO;2, 1972.

Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, USA, 1989.

Guyon, I. and Elisseeff, A.: An introduction to variable and feature selection, Journal of Machine Learning Research, 3, 1157–1182, 2003.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnóti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree, Advances in Neural Information Processing Systems, 30, 3146–3154, 2017.

Oke, T. R.: Boundary Layer Climates, 2nd edn., Routledge, London, UK, 1987.

NOAA National Centers for Environmental Information (NCEI): Global Surface Summary of the Day (GSOD), version 1.0, https://www.ncei.noaa.gov/access/search/data-search/global-summary-of-the-day, last access: 15 April 2026, 1999.

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.

Stull, R. B.: An Introduction to Boundary Layer Meteorology, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1988.

Tibshirani, R.: Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society Series B, 58, 267–288, 1996.

von Storch, H. and Zwiers, F. W.: Statistical Analysis in Climate Research, Cambridge University Press, Cambridge, UK, 1999.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 3rd edn., Academic Press, Amsterdam, the Netherlands, 2011.

Xue, B., Zhang, M., Browne, W. N., and Yao, X.: A survey on evolutionary computation approaches to feature selection, IEEE Transactions on Evolutionary Computation, 20, 606–626, https://doi.org/10.1109/TEVC.2015.2504420, 2016.