THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Unité de recherche : IAB : Institute for Advanced Biosciences (UGA / Inserm U1209 / CNRS UMR 5309)

Cadres hybrides d'apprentissage automatique pour la modélisation à haute résolution de l'exposition environnementale : applications à la température de l'air et aux particules fines en Israël et en France

Hybrid machine learning frameworksfor high resolution environmental exposure modelling: Applications to air temperature and particulate matter across Israel and France

Présentée par :

Juan BRICENO GUERRERO

Direction de thèse :

Johanna LEPEULE — Directrice de thèse
DIRECTRICE DE RECHERCHE, INSERM DELEGATION AUVERGNE-RHONE-ALPES

Itai KLOOG — Co-directeur de thèse
FULL PROFESSOR, BEN-GURION UNIVERSITY OF THE NEGEV

Rapporteurs :

RAANAN RAZ
ASSOCIATE PROFESSOR, THE HEBREW UNIVERSITY OF JERUSALEM

PETTER LJUNGMAN
ASSOCIATE PROFESSOR, KAROLINSKA INSTITUTET

Thèse soutenue publiquement le **18 décembre 2025**, devant le jury composé de :

GAËLLE UZU, DIRECTRICE DE RECHERCHE, IRD DELEGATION REGIONALE SUD EST	Présidente
ITAI KLOOG, FULL PROFESSOR, BEN-GURION UNIVERSITY OF THE NEGEV	Co-directeur de thèse
RAANAN RAZ, ASSOCIATE PROFESSOR, THE HEBREW UNIVERSITY OF JERUSALEM	Rapporteur
PETTER LJUNGMAN, ASSOCIATE PROFESSOR, KAROLINSKA INSTITUTET	Rapporteur

Université Grenoble Alpes

Doctoral School Health, cognition and environment engineering (EDISCE)
Institute for Advanced Biosciences (IAB)

Doctoral Thesis: Hybrid machine learning frameworks for high-resolution environmental exposure modelling: Applications to air temperature and particulate matter across Israel and France

Juan David Briceño Guerrero

To obtain the degree of Doctor of Philosophy (PhD)

Discipline: Environmental Epidemiology

Prepared under the supervision of

Dr. Itai Kloog, Full Professor

Department of Environmental, Geoinformatics and Urban Planning Sciences

Ben-Gurion University of the Negev, Be'er Sheva, Israel

Dr. Johanna Lepeule, Director of Research at Inserm

Institute for Advanced Biosciences, INSERM, CNRS

Université Grenoble Alpes, Grenoble, France

Acknowledgements

This research was made possible through the generous financial support of the U.S.-Israel Binational Science Foundation (BSF, Grant No. 2021197) and the French National Research Agency (ANR) under the "Investissements d'avenir" program (Grant No. ANR-15-IDEX-02 IRGA 2021 PAPex project). I am deeply grateful for this scholarship support, which enabled me to pursue my doctoral studies and conduct this research in an enriching international framework.

Besides, I would like to sincerely thank my thesis directors, Itai Kloog and Johanna Lepeule, for their invaluable guidance, support, and encouragement throughout this research. Their expertise, patience, and advice have been instrumental in shaping this work and helping me navigate the challenges of my thesis journey.

I am also deeply grateful to my parents for their unwavering support in all aspects of my life, as well as to my girlfriend and sisters for their constant encouragement and understanding.

Special thanks go to the institute for advanced biosciences (IAB) and Ben-Gurion University of the Negev for providing an inspiring research environment and the resources necessary to carry out this study.

Abstract

Global temperatures and urbanization are both rising rapidly, while air pollution remains a key environmental concern due to its strong links with air quality and human health, highlighting the need for reliable, fine-scale assessments of environmental exposures. This thesis develops hybrid machine learning (ML) frameworks to generate high-resolution estimates of two key environmental exposures: air temperature (Ta) and particulate matter (PM). The main objective of this work is to enhance exposure assessment for epidemiological applications by integrating diverse environmental datasets from multiple sources (e.g., NASA satellite imagery, Copernicus services, EUMETSAT, etc.) with innovative statistical learning approaches.

To achieve this, we built high-resolution models for Ta (Israel and France) and PM (Israel) by combining various spatial and spatio-temporal predictors with feature engineering strategies ranging from classic to advanced methods commonly used in the data science domain. Multistage procedures were applied separately for each target, managing data flow and producing daily predictions at fine spatial resolutions (≤1 km). Feature engineering included the creation of geographically weighted predictors and the imputation of critical covariates, such as Land Surface Temperature (LST) and Aerosol Optical Depth (AOD), which are particularly relevant due to their fine spatio-temporal resolution and correlation with the studied targets.

Our models demonstrated excellent predictive performance under advanced cross-validation schemes using root-mean-squared error (RMSE). For Ta, RMSE values ranged from 0.5 °C to 1.0 °C in Israel and from 0.4 °C to 1.1 °C in France. For PM in Israel, RMSE values ranged from 10.2 to 26.9 μg/m³ for PM10 (particulate matter with aerodynamic diameter ≤10 μm) and from 2.7 to 13.9 μg/m³ for PM2.5 (particulate matter with aerodynamic diameter ≤2.5 μm). These models were carefully cross-validated using multiple strategies, ensuring no data leakage during feature engineering and guaranteeing robustness. Our results demonstrate that the framework effectively captures fine-scale variability in environmental exposures and outperforms previous studies in similar contexts, demonstrating clear improvements over the literature.

By producing continuous, high-resolution exposure surfaces, this work reduces errors in epidemiological exposure assessment and supports public health research, urban planning,

and climate adaptation strategies. The methodological advances—particularly the integration of satellite data, feature engineering, and hybrid ML models—can be generalized to other pollutants and regions, strengthening the scientific basis for environmental and public health policies.

Résume (French Abstract)

Les températures mondiales et l'urbanisation augmentent rapidement, tandis que la pollution de l'air demeure une préoccupation environnementale majeure en raison de ses liens étroits avec la qualité de l'air et la santé humaine, soulignant la nécessité d'évaluations fiables et à fine échelle des expositions environnementales. Cette thèse développe des cadres hybrides d'apprentissage automatique (ML) afin de générer des estimations à haute résolution de deux expositions environnementales clés : la température de l'air (Ta) et les particules en suspension (PM). L'objectif principal de ce travail est d'améliorer l'évaluation des expositions pour des applications épidémiologiques en intégrant des ensembles de données environnementales provenant de multiples sources (par ex. imagerie satellite de la NASA, services Copernicus, EUMETSAT, etc.) à l'aide d'approches d'apprentissage statistique innovantes.

Pour y parvenir, nous avons construit des modèles à haute résolution pour Ta (Israël et France) et PM (Israël) en combinant divers prédicteurs spatiaux et spatio-temporels avec des stratégies de *feature engineering* allant des méthodes classiques aux méthodes avancées couramment utilisées en science des données. Des procédures en plusieurs étapes ont été appliquées séparément pour chaque variable cible, permettant de gérer le flux de données et de produire des prédictions quotidiennes à des résolutions spatiales fines (≤1 km). Le *feature engineering* a inclus la création de prédicteurs pondérés géographiquement et l'imputation de covariables, telles que la température de surface terrestre (LST) et la profondeur optique des aérosols (AOD), particulièrement pertinentes en raison de leur haute résolution spatio-temporelle et de leur corrélation avec les variables étudiées.

Nos modèles ont démontré d'excellentes performances prédictives grâce à des schémas de validation croisée avancés utilisant l'erreur quadratique moyenne (RMSE). Pour Ta, les valeurs de RMSE variaient de 0,5 °C à 1,0 °C en Israël et de 0,4 °C à 1,1 °C en France. Pour PM en Israël, les valeurs de RMSE variaient de 10,2 à 26,9 µg/m³ pour PM_{10} (particules avec un diamètre aérodynamique ≤10 µm) et de 2,7 à 13,9 µg/m³ pour $PM_{2.5}$ (particules avec un diamètre aérodynamique ≤2,5 µm). Ces modèles ont été soigneusement validés en utilisant plusieurs stratégies, garantissant l'absence de fuite de données lors de l'intégration du feature engineering et assurant leur robustesse. Nos résultats montrent que ce cadre capture efficacement la variabilité à fine échelle des expositions environnementales et surpasse les

études précédentes dans des contextes similaires, démontrant des améliorations nettes par rapport à la littérature.

En produisant des surfaces d'exposition continues et à haute résolution, ce travail réduit les erreurs de mesure dans l'évaluation des expositions et soutient la recherche en santé publique, l'urbanisme et les stratégies d'adaptation au climat. Les avancées méthodologiques—en particulier l'intégration des données satellitaires, le feature engineering et les modèles hybrides de ML—peuvent être généralisées à d'autres polluants et régions, renforçant ainsi la base scientifique des politiques environnementales et de santé publique.

Table of Contents

Glossary of Key Terms

Chapter 1: General Introduction ……..……..1

1.1 What's Ta? ...……..…….....2
1.2 What's PM? ……..……...…...3
1.3 Structure of the Thesis ……...…..……..5
1.4 Global Environmental Context ……..……6
1.5 Research Contribution ...……...……...9
1.6 Research Questions ……...…10

Chapter 2: Literature Review ……..……..11

2.1 Evolution of Spatiotemporal Approaches in Environmental Modeling …….……..11
2.2 Air temperature Modeling with Satellite data ……..17
2.3 PM Modeling with Satellite data ……..…...……..35
2.4 Conclusions from the Literature Review on Satellite-Based Models for Ta and PM 49

Chapter 3: Methodology ……59

3.1 Overview of Modelling Methodology ……59
3.2 Data ……...…61
3.3 Data Preprocessing ……...67
3.4 Feature Engineering ……70
3.4.1 From Basic to Advanced Feature Engineering…….……74
3.4.2 How to integrate Feature Engineering During Cross-Validation……..90
3.4.3 Similarities and Differences between FEs and Interpolation Techniques……98
3.4.4 Conclusions from feature Engineering and its integration within validation techniques……103

Chapter 4: Articles ……..…105

4.1 Article 1 ……..105
4.2 Article 2 ……..137
4.3 Article 3 ……..178

Conclusions and Future Work ……221

Glossary of Key Terms

Machine Learning & AI

Neural Network (NN): A model inspired by the human brain, consisting of interconnected nodes (neurons) that transform input data through layers to capture complex patterns and relationships for prediction or classification tasks.

Artificial Intelligence (AI): The broad discipline of creating systems capable of performing tasks that normally require human intelligence, including reasoning, learning, and decision-making.

Machine Learning (ML): A field of artificial intelligence focused on designing algorithms that can learn from data and improve their predictions or decisions without being explicitly programmed.

Feature Engineering (FE) – Feature Generation: The process of creating new input variables (features) from raw data to improve model performance by making implicit patterns more explicit.

Classification Tasks: A type of supervised learning problem where the goal is to assign input observations into discrete categories or classes based on their attributes.

Random Forest (RF): An ensemble learning method that builds multiple decision trees and aggregates their predictions to improve accuracy, reduce overfitting, and handle complex, non-linear relationships.

Extreme Gradient Boosting (XGBoost): A scalable and efficient gradient boosting framework that builds sequential decision trees while minimizing prediction errors, often yielding high performance in tabular data problems.

Support Vector Regression (SVR): A regression method that seeks a function within a specified tolerance of the training data, using a hyperplane to maximize prediction accuracy and generalization.

Gaussian Markov Random Fields (GMRF): A statistical model capturing spatial dependencies by assuming that each location's value is conditionally dependent on its neighbors, commonly used in environmental modeling.

Bi-LSTM Neural Network: A type of recurrent neural network that processes sequences in both forward and backward directions, improving the model's ability to capture temporal dependencies.

CNN Neural Network: A neural network designed to recognize patterns in grid-like data (e.g., images), using convolutional layers to extract spatial hierarchies of features.

RNN (Recurrent Neural Network): A neural network that processes sequential data by maintaining a memory of previous inputs, suitable for time series prediction and other sequential tasks.

Ensemble Learning Modeling: A machine learning approach that combines predictions from multiple models to improve overall performance and robustness.

Bootstrap Aggregating (Bagging): A technique in ensemble learning where multiple versions of a model are trained on different bootstrap samples of the data, and predictions are aggregated to reduce variance.

Transfert Learning: A method in machine learning where a model trained on one dataset or task is adapted to another related task, reducing training time and leveraging previously learned patterns.

Labeled Data**:** Data where each instance has a known output, used to train supervised machine learning models.

KNN Model (K-Nearest Neighbors)**:** Predicts values or classes for a data point based on the attributes of its nearest neighbors.

Statistical Methods & Validation

Linearity**:** The assumption that predictors relate to the response variable through a straight-line relationship, a key requirement for many regression models.

Homoscedasticity**:** The condition where residuals have constant variance across predictor values, ensuring accurate standard errors and statistical inference.

Normality of Residuals**:** The assumption that residuals follow a normal distribution, crucial for hypothesis testing and confidence interval calculations.

Euclidean Distance**:** The straight-line distance between two points in multidimensional space, widely used in clustering and KNN algorithms.

One-Hot Encoding**:** Converting categorical variables into binary vectors, allowing machine learning algorithms to process non-numeric features.

K-Means Model**:** An unsupervised learning algorithm that divides data into k clusters, minimizing variance within clusters.

Ordinary Least Squares (OLS): A linear regression technique that estimates coefficients by minimizing the sum of squared differences between observed and predicted values, commonly used for modeling continuous outcomes.

Linear Mixed-Effects Model with Day-Specific Random Effects: A regression model that accounts for both fixed effects and random effects, allowing daily variations to be modeled while capturing overall trends in longitudinal data.

Generalized Additive Model (GAM): A flexible regression approach that models relationships between predictors and response variables using smooth functions, capturing non-linear trends without assuming a specific parametric form.

Generalized Additive Mixed Model (GAMM): An extension of GAM that incorporates random effects, allowing modeling of both non-linear trends and hierarchical or grouped structures in data.

K-Fold Out-of-Sample Cross-Validation: A method for evaluating model performance by dividing the dataset into k subsets, sequentially training on $k-1$ folds and testing on the remaining fold, to reduce overfitting and improve generalization.

Spatial/Monitor-Blocked Cross-Validation: A cross-validation approach where spatially separated areas or monitoring stations are held out during training to assess model performance realistically in environmental contexts.

RMSE (Root Mean Squared Error): A metric that measures the average magnitude of prediction errors by taking the square root of the mean squared differences between predicted and observed values, highlighting large deviations.

MAE (Mean Absolute Error): A metric calculating the average absolute difference between predicted and observed values, providing an interpretable measure of prediction accuracy.

Log-Cosh Error: A performance metric for regression that approximates the mean squared error for small differences but is less sensitive to outliers than traditional squared error.

MAD (Median Absolute Deviation) for Performance Evaluation: A robust metric that quantifies the median absolute deviation from the median prediction, useful when data contain outliers.

R^2 Metric-Score: A statistical measure indicating the proportion of variance in the observed data explained by the model, with higher values representing better fit.

Odds Ratio (OR): A measure commonly used in epidemiology to quantify the association between an exposure and an outcome, representing the odds of an event occurring in one group relative to another.

SHAP Values Performance of Predictors: A method from explainable AI that assigns an importance score to each predictor, quantifying its contribution to the model's output for individual predictions.

Stratified Cross-Validation: A variation of cross-validation that ensures each fold has a similar distribution of target classes or key covariates, improving representativeness of performance estimates.

Out-of-Fold (OOF) Predictions / Out-of-Sample Predictions: Predictions generated for the validation folds during cross-validation. For each iteration, the model is trained on the training folds and evaluated on the

held-out validation fold. This ensures that the OOF predictions are made on data that the model has never seen, closely simulating real-world performance.

Exploratory Data Analysis (EDA): A set of techniques to summarize, visualize, and understand data distributions, relationships, and anomalies before formal modeling.

Satellite Data, Environmental Variables & GIS

Land-Use Regression (LUR): A modeling approach linking spatially distributed predictors such as land use, traffic, and population density to observed pollutant concentrations, commonly used for air pollution exposure assessment.

GIS-Derived Predictors: Variables extracted from Geographic Information Systems (GIS), including distances, land cover, elevation, and infrastructure features, used as covariates in environmental models.

ASTER Global Digital Elevation Model (GDEM): A satellite-derived digital elevation dataset providing high-resolution terrain information, used for topography-based modeling and hydrology studies.

Geographically Weighted Regression (GWR): A local regression technique that estimates spatially varying relationships between predictors and outcomes, allowing model coefficients to change across locations.

Inverse Distance Weighting (IDW): A simple spatial interpolation method where the influence of known points decreases with distance, used for estimating environmental variables at unsampled locations.

Kriging / Universal Kriging: An extension of kriging that accounts for both spatial autocorrelation and deterministic trends in data, improving interpolation accuracy for environmental variables.

Assimilation Techniques in Environmental Modelling**:** Methods that merge observational data with numerical models to refine predictions of environmental variables like temperature or air pollution, continuously updating models as new measurements become available.

Interpolation**:** The process of estimating values at locations where no measurements exist, using nearby observed data, often assuming spatial or temporal continuity.

First-Order Fourier Curve**:** A mathematical representation of periodic phenomena using sine and cosine terms, often applied to capture seasonal cycles in environmental variables.

Multi-Stage Monitor-Blocked Cross-Validation**:** A method to evaluate model performance where data is grouped by spatial regions. A stage is one step in which the model is trained on most of the blocks and tested on a single held-out block. The process is repeated for each block, and theresults from all stages are combined to give a realistic estimate of performance on unseen locations.

CAMS AOD**:** Aerosol Optical Depth measurements from the Copernicus Atmosphere Monitoring Service, used to estimate particulate matter and track atmospheric aerosols.

The Urban Heat Island Effect: The phenomenon where urban areas are warmer than surrounding rural areas due to human activities, built surfaces, and reduced vegetation.

Stratification of Data**:** Dividing a dataset into meaningful subgroups to maintain balanced representation during modeling or cross-validation.

NOAA Climate Region in USA**:** Standardized geographic zones used by the National Oceanic and Atmospheric Administration for climate monitoring and analysis.

Landsat Thermal Data**:** Satellite-derived surface temperature measurements used in urban heat, environmental, and climate studies.

NASA's Aura Satellite**:** An Earth-observing satellite focused on atmospheric composition, measuring ozone, aerosols, and trace gases for climate and air quality monitoring.

EUMETSAT**:** European organization providing continuous meteorological satellite data for weather forecasting, climate monitoring, and environmental research.

NASA TIRS – Thermal Infrared Sensor aboard Landsat 8, measuring thermal radiation to provide surface temperature data for environmental monitoring, agriculture, and climate studies.

LSA SAF – Land Surface Analysis Satellite Applications Facility from EUMETSAT, producing satellite-derived products on soil moisture, snow cover, and surface temperature for research and operational applications.

Leading U.S. Product of PM – EPA's Fused Air Quality Surface**:** A high-resolution dataset integrating ground monitors, satellite observations*, and models to provide daily particulate matter estimates across the U.S.*

MODIS (Moderate Resolution Imaging Spectroradiometer): A satellite instrument aboard NASA's Terra and Aqua satellites, providing high-resolution data on land, ocean, and atmospheric variables, commonly used for environmental monitoring.

Land Surface Temperature (LST): The radiative skin temperature of the Earth's surface derived from satellite measurements, important for climate, weather, and urban heat studies.

ERA5-ECMWF: A high-resolution global reanalysis dataset from the European Centre for Medium-Range Weather Forecasts, combining observations and models to produce consistent climate and weather variables over time.

Numerical Weather Prediction (NWP): The use of mathematical models of the atmosphere and oceans to predict future weather based on current observational data.

Urban Heat Islands (UHI): Urban areas exhibiting higher temperatures than surrounding rural areas due to human activity, infrastructure, and reduced vegetation, influencing local climate and health.

SEVIRI (Spinning Enhanced Visible and Infrared Imager): A satellite sensor aboard the Meteosat Second Generation satellites, providing high-frequency images for weather monitoring and environmental studies.

Reanalysis Products: Datasets combining past observational data with numerical models to generate consistent long-term environmental variables, widely used in climate and air quality studies.t

Shortwave Infrared (SWIR): Part of the electromagnetic spectrum with wavelengths between 1,000 nm and 2,500 nm, commonly used in satellite data to detect features such as water content, vegetation, and certain surface materials that are less visible in the visible or near-infrared ranges.

MeteoSwiss – Swiss Federal Office of Meteorology and Climatology: The national weather and climate service of Switzerland, responsible for monitoring, forecasting, and researching

atmospheric conditions. MeteoSwiss provides meteorological data, climate analyses, and early warnings to support public safety, environmental management, and scientific research.

Programming Languages / Software

Python Programming**:** A versatile scientific programming language with extensive libraries for data analysis, machine learning, and mathematical computing.

R Programming**:** A statistical computing language used for data analysis, visualization, and environmental modeling.

MATLAB**:** A numerical computing environment for algorithm development, matrix computations, and visualization.

Docker**:** A containerization platform that packages software and its dependencies for reproducible and portable deployments.

APIs (Application Programming Interfaces)**:** Tools that allow software programs to communicate and retrieve data, commonly used to access environmental datasets.

Spark or H2O**:** Distributed computing frameworks for handling large datasets efficiently, supporting scalable machine learning and model training.

Chapter 1: General Introduction

This dissertation is intended for scientists, geographers, epidemiologists and other researchers and practitioners involved in the study of environmental exposure, a multidisciplinary field that integrates meteorology, epidemiology and data science to characterize and estimate the spatial and temporal distribution of climatic factors affecting human health. To this end, exposure models draw on a range of mathematical approaches to predict exposure levels to these factor-variables based on well-characterized and highly relevant covariates.

In exposure modeling, researchers develop models to reproduce the behavior of a wide range of environmental variables. These commonly include air temperature (Ta); particulate matter (PM_{10} and $PM_{2.5}$, collectively referred to as PM); nitrogen dioxide (NO_2); ozone (O_3); carbon monoxide (CO); carbon dioxide (CO_2); humidity; volatile organic compounds (VOCs), such as benzene and formaldehyde; and other pollutants, including sulfur dioxide (SO_2), ammonia (NH_3), and heavy metals like lead (Pb) and mercury (Hg). To estimate these variables across space and time, models integrate data from multiple sources, including satellite observations, ground-based meteorological stations, and hybrid real-time data from assimilation techniques. This integration allows reconstruction of long-term environmental conditions over extensive areas, even in locations without direct monitoring. The increasing availability of high-resolution satellite imagery has been particularly influential, providing rich datasets that enhance the development, accuracy, and capabilities of these modeling approaches.

While these models are not limited to forecasting future conditions, they are primarily designed to generate historical estimates with continuous spatial coverage, addressing the needs of epidemiological studies. Such reconstructions allow researchers to evaluate associations between environmental variables and health outcomes, providing critical evidence to inform public health policies and strategies for mitigating exposure-related risks.

Exposure modeling has developed rapidly, initially relying on low- to medium-range approaches capable of predicting environmental variables over a few kilometers using techniques based on physical laws, data assimilation, land use regression, and interpolation. Over time, these methods have evolved into more sophisticated machine learning (ML) approaches, capable of capturing complex spatial and temporal relationships and enabling predictions across vast terrains without computational limitations.

The future of the field appears promising, with ongoing advancements in spatial and temporal resolution allowing for increasingly precise estimates that enhance the robustness of health-related epidemiological outcomes. By leveraging advanced algorithms for automated or manual feature generation, these new models will exploit the strengths of AI, being ultra-precise and reaching high levels of accuracy, as already observed in other fields where the integration of neural network architectures has achieved near-perfect accuracy, for example, in classification tasks for vehicle self-driving (Martínez, 2024).

First and foremost, considerable ambiguity can arise in these studies due to the similarity in terminology and the closely related logic underlying the variables considered. To address this, we begin by providing clear, practical definitions for key variables such as Ta, PM, and other predictors within the same family. Establishing these definitions from the outset helps eliminate potential confusion about the scope, purpose, and conceptual framework of this thesis.

1.1 What's Ta?

To avoid misunderstandings, we begin by defining Ta, due to its importance in this manuscript, as it is often defined differently across the literature, with multiple predictors referring to the concept of temperature without necessarily representing the same physical quantity.

Ta refers to the temperature of the air at a specified height above the surface, typically about two meters above ground level, where it is measured by meteorological stations. Air temperature is expressed in degrees, usually Celsius (°C) or Kelvin (K). Despite its apparent simplicity, air temperature is often confused with other types of temperature, such as surface or skin temperature.

Surface and Skin Temperature

Surface temperature usually refers to the temperature of the Earth's surface itself, such as land, water, or ice, representing the temperature of the bulk ground or water just beneath the very top layer (Li et al., 2023). In contrast, skin temperature refers to the very thin outer layer, only a few micrometers thick, which emits thermal infrared radiation (Jin et al., 1997). This is the layer that satellites mainly measure, and because it reacts quickly to sunlight, wind, and moisture, it can differ from the deeper surface temperature. Paradoxically, satellite products such as Land Surface Temperature (LST) from MODIS do not directly measure the bulk

ground temperature, but rather the skin temperature. The term LST was chosen because it represents the temperature of the land's radiative surface as seen from space, which is what satellites can physically observe: thermal infrared sensors measure emitted radiation from the topmost layer of the surface, not the subsurface heat. Skin temperature is extremely useful, as it informs heat fluxes, energy balance, and environmental exposure. To illustrate the difference, one can think of a lake on a sunny day: the bulk water below a few centimeters may remain relatively cool, representing surface temperature, while the very top layer that the sun directly heats is warmer and changes quickly, representing skin temperature.

Even though surface and skin temperatures are also expressed in degrees (°C or K), they differ from Ta because air temperature measures the warmth of the air just above the ground, whereas surface and skin temperatures reflect the warmth of the ground, water, or ice beneath the top layer (surface temperature) or at the very thin outermost layer of these surfaces (skin temperature). The surface responds more quickly to sunlight, shading, snow, or nighttime cooling, so it can be significantly warmer or cooler than the air at the same location.

From other temperature sources, such as ERA5 reanalysis (ECMWF), these distinctions between the different temperature types are clearly defined. The 2-meter air temperature represents near-surface atmospheric conditions, whereas the skin temperature corresponds to the modeled radiative surface layer. ERA5 also provides soil and land temperatures at multiple depths, sea surface temperatures for the uppermost ocean layer, and atmospheric temperatures on various pressure levels (e.g., 1000, 850, or 500 hPa), which represent the free atmosphere rather than conditions near the surface. Together, these variables illustrate that air, surface, and skin temperatures are conceptually and physically distinct and should not be used interchangeably without careful consideration.

1.2 What's PM?

Particulate matter (PM) can be defined as a mixture of solid components of various sizes and tiny suspended liquids, which form when water vapor condenses around chemical substances in the air. These particles, varying in size and composition, are classified by their aerodynamic diameter: PM_{10} consists of particles with a diameter of 10 micrometers or less, while $PM_{2.5}$ includes particles with a diameter of 2.5 micrometers or less. Their concentrations are typically measured in micrograms per cubic meter ($\mu g/m^3$). The smaller the particle, the deeper it can penetrate into the respiratory system, increasing potential health

risks. $PM_{2.5}$ can even enter the bloodstream and is linked to serious health problems, including respiratory and cardiovascular diseases (U.S. Environmental Protection Agency, 2025). These particles originate from a variety of sources, both natural and human-made, including vehicle emissions, industrial processes, chemical evaporation, and dust storms. Understanding their distribution and concentration is essential for assessing air quality and evaluating impacts on public health.

Aerosol Optical Depth

Similar to what happens with Ta and other temperature types, one might be tempted to think of aerosol optical depth (AOD) as equivalent to PM. Although AOD and PM are often correlated, they describe fundamentally different aspects of airborne particles. PM measures the actual mass concentration of particles at the Earth's surface (PM_{10} and $PM_{2.5}$), categorized by their aerodynamic diameter. In contrast, AOD is an optical property that quantifies the extinction of sunlight by aerosols over a vertical column of the atmosphere, providing information on the total aerosol load rather than the mass at the surface. By aerosol we understand everything that can be considered as a tiny solid or liquid particle suspended in the atmosphere such as dust, sea salt, sulfate droplets, or condensed chemicals which have the capacity to scatter, absorb, or attenuate sunlight. Thus, the measure done in AOD is not about the particles themselves but about their effect on the transmission of the light (He et al., 2021).

AOD is wavelength-dependent, which allows scientists to infer aerosol size and composition. Wavelengths are measured in nanometers (nm), where 1 nm equals one billionth of a meter, and correspond to the color or energy of the light being scattered or absorbed. Shorter wavelengths, such as 440 nm in the blue range, are more sensitive to fine-mode aerosols, like smoke or sulfate particles, because these smaller particles scatter blue light more efficiently. Wavelengths in the green spectrum, around 500 nm, provide a general overview of aerosol presence, capturing contributions from both fine and coarse particles. Longer wavelengths, such as 675 nm in the red range and near-infrared wavelengths like 778 and 862 nm, are more sensitive to coarse particles, including desert dust and sea salt, which scatter longer wavelengths more effectively. Even longer wavelengths, around 1240 nm and beyond, are in the shortwave infrared (SWIR) region, which is less affected by fine aerosols and more useful for detecting larger particles, water vapor, and certain types of liquid aerosols. By analyzing AOD across this spectrum of wavelengths, researchers can estimate the size distribution,

composition, and dominant aerosol types in the atmosphere (NASA Earth Science Division, n.d.).

It is important to note that AOD does not measure the mass of particles directly and therefore cannot be used interchangeably with PM concentrations. PM represents particles at ground level, which directly affect air quality and human health. In contrast, AOD integrates aerosol effects over the entire atmospheric column, capturing both fine and coarse particles as well as other suspended components, including liquid aerosols formed when water vapor condenses around chemical substances. As a result, while a high AOD may indicate elevated particle presence in the atmosphere, it does not provide precise information on the amount or composition of PM at the surface.

Understanding these distinctions is essential for environmental exposure research and related health studies. AOD provides valuable insight into atmospheric aerosol loads, whereas PM measurements offer direct evidence of human exposure to harmful particles.

1.3 Structure of the Thesis

After introducing the key terminology for Ta and PM, we now highlight the main purpose of this thesis. Beyond reproducing the behavior of the previously defined variables, this work aims to contribute to the scientific field through the development, integration, and evaluation of feature engineering within hybrid machine learning frameworks for environmental exposure. Throughout the different sections, we describe the evolution of related studies, present the motivation that drove us toward this proposal, and provide the mathematical formulations necessary to ensure reproducibility. We also highlight the direct applications of our approach to the Israeli and French contexts. This dissertation is guided by several overarching research questions, which are addressed indirectly throughout the chapters and resolved directly in the final conclusions.

Chapters' Content

This dissertation is organized as follows. The remainder of Chapter 1 introduces the context of exposure modeling and explains the importance of generating high-resolution predictions of Ta and PM. Chapter 2 presents a comprehensive literature review of 16 studies on Ta and PM, tracing the evolution of modeling approaches for these variables and highlighting their main limitations. Chapter 3 details feature engineering, including its development, evaluation, and innovations, along with their impacts. Chapter 4 compiles our three research articles

developed during the work of this thesis, emphasizing the development of our strategies. These articles are presented in their final published form: modeling Ta in Israel, modeling Ta in France, and modeling PM in Israel. Finally, the conclusions discuss future research directions and present the overall findings of the PhD project.

1.4 Global Environmental Context

Our planet is experiencing rapid environmental changes driven by climate change, urbanization, and pollution. Global average temperatures have risen significantly over the past century, resulting in more frequent and severe heatwaves (IPCC, 2023). Simultaneously, urban expansion has increased emissions of air pollutants, including particulate matter (PM), which tend to accumulate in densely populated areas and amplify exposure risks (WHO, 2021). These converging trends highlight the urgent need to monitor key environmental factors, such as Ta and PM, using the latest technological advancements and measurement tools. Such monitoring is crucial for developing effective policies to protect populations, particularly those most vulnerable and sensitive to these environmental stressors.

Among environmental exposures, extreme air temperature and high particulate matter pollution pose major public health challenges. Extreme heat events are associated with increased mortality and morbidity, particularly among vulnerable populations such as older adults, children, and individuals with pre-existing cardiovascular or respiratory conditions (Gasparrini et al., 2015). Both long- and short-term exposure to fine particulate matter ($PM_{2.5}$ and PM_{10}) are linked to respiratory and cardiovascular diseases as well as premature mortality (Cohen et al., 2017). Advancements in technology have enhanced our ability to analyze these health risks. Traditional statistical approaches, often relying on local, ground-based measurements, can now be augmented with ML, allowing for broader, more precise assessments. These approaches minimize bias, enabling studies to leverage spatiotemporal inputs that reflect actual in situ conditions.

Environmental Modeling Across Space and Time

Guaranteeing spatial and temporal variability of environmental exposures is therefore crucial. Their interactions with socio-environmental factors make high-resolution assessments essential for epidemiological research, enabling more realistic exposure estimates, improved risk management, and the development of targeted public health interventions. For example, relying solely on measurements from fixed monitoring stations or coarse-resolution satellite

data may fail to capture local variability, especially between urban and rural areas. Air temperature illustrates this well: cities generally experience higher temperatures due to anthropogenic heat and slower heat dissipation, whereas rural areas tend to exhibit lower values. While temperature-related health effects have been observed in rural populations (Lee et al.), understanding these impacts is more difficult because high-resolution Ta estimates are often lacking in such regions (Hough et al., 2020). This limitation highlights the importance of developing fine-scale, spatio-temporally continuous datasets.

Importance of High-Resolution Data

In early attempts, the demand for high spatiotemporal resolution was addressed using mathematical formulations based on physical laws, known as Numerical Weather Prediction (NWP) models. Despite their value for operational weather forecasting, NWP outputs have limited utility for public health and exposure studies. Their standard resolutions, typically between 4 and 20 km depending on the model and domain, tend to smooth out sharp spatial gradients that are crucial for understanding human exposure (Masson et al., 2020). For Ta for example, this means that local variations related to urban heat islands or complex topography are often poorly represented. For particulate matter, the coarse resolution of NWP outputs cannot capture hotspots near traffic corridors or industrial sources, leading to underestimation of exposures in vulnerable communities (Hoek et al., 2008). In epidemiological research, such mismatches can translate into biased associations between Ta–PM exposures and health outcomes, while in urban planning they limit the capacity to design targeted mitigation strategies at the neighborhood scale.

Satellite Data as a Tool

Satellite data provides a complementary approach to overcome the limitations of traditional NWP models by offering spatially continuous observations at higher resolution. The Moderate Resolution Imaging Spectroradiometer (MODIS) has been particularly influential, supplying daily products at 1 km resolution. Land Surface Temperature (LST) retrievals serve as a proxy for near-surface Ta, while Aerosol Optical Depth (AOD) is widely used as a proxy for PM. These datasets have enabled the creation of finer-scale exposure models that better capture urban–rural contrasts and regional pollution gradients (Kloog et al., 2014; Chudnovsky et al., 2014; Stafoggia et al., 2016).

For example, MODIS LST has been shown to correlate strongly with Ta in Israel, France, and the USA, with hybrid ML models achieving RMSE values between 1 and 2 °C (Kloog et al.,

2014; Kloog et al., 2017; Rosenfeld et al., 2017; Hassani et al., 2024). Similarly, MODIS AOD products from the MAIAC algorithm have been successfully integrated into $PM_{2.5}$ models in Europe and the Middle East, improving predictions in areas with sparse monitoring coverage (de Hoogh et al., 2018; Lee et al., 2015).

However, important challenges remain: LST retrievals are limited by cloud cover, and AOD retrievals can suffer from missing data or errors over bright or heterogeneous surfaces. Thus, while MODIS products represent a major advance toward high-resolution exposure modeling, they must be integrated with ground observations and advanced statistical approaches to achieve the spatiotemporal continuity required for public health research.

Machine Learning in Exposure Modelling

Machine learning approaches are increasingly applied in exposure modeling, addressing the inherent limitations of traditional NWP methods, which rely heavily on physical laws. These approaches are often supported by networks of ground-based monitors allocated at multiple locations, providing labeled data from specific areas and at periodic intervals. ML can capture complex, nonlinear relationships among environmental variables and effectively leverage large, heterogeneous datasets. This is particularly valuable in exposure modeling where predictors such as satellite-derived land surface temperature (LST), aerosol optical depth (AOD), vegetation indices (NDVI), local topography, meteorological reanalysis, and wind patterns provide critical information at various fine spatial and temporal scales. By integrating these diverse data streams, ML models can provide accurate predictions at finer spatial and temporal resolutions, adapting to the local characteristics of the analyzed terrain. As a result, ML is particularly well-suited to improving exposure estimates of variables such as Ta and PM, especially in regions where the coverage of monitors is not dense.

Recent studies have demonstrated the effectiveness of ML approaches in predicting Ta across diverse regions. These models typically employ multiple predictors, including satellite-derived LST (e.g., from MODIS and SEVIRI), reanalysis data, NDVI, and other static variables such as population and elevation. To model Ta, various ML strategies have been applied, including advanced tree-based algorithms, ensemble techniques, and locally adaptive models, all trained on ground-based meteorological observations to capture the complex relationships between predictors and Ta. These approaches have enabled accurate Ta modeling across a wide variety of contexts, including Europe (Italy: Limoncella et al., 2025; Germany: Nikolaou et al., 2022; France: Hough et al., 2020; Switzerland: Flückiger et al.,

2022; Sweden: Jin et al., 2022), the Middle East (Israel: Zhou et al., 2018), Asia (China: Chen et al., 2021), and the United States (Just et al., 2024), illustrating the versatility and scalability of ML methodologies across different climates and terrains.

Machine learning has also been proven effective for modeling particulate matter (PM_{10} and $PM_{2.5}$) concentrations, addressing challenges posed by sparse monitoring networks. These models typically employ satellite-derived observations of aerosol optical depth (AOD) from MODIS, reanalysis CAMS AOD, along with meteorological data, land-use information, and proxies for anthropogenic emissions, such as proximity to major roads, highway density, and industrial areas. As with LST, a crucial step in the modeling process involves imputing missing AOD values to maintain the integrity of the PM models' inputs. Models are then calibrated using ground-based PM measurements to establish accurate relationships between predictors and PM. Various ML techniques have been applied, including tree-based algorithms and ensemble learning, to generate high-resolution PM estimates. These methodologies have been successfully implemented in multiple regions, including Europe (France: Hough et al., 2021; Italy: Shtein et al., 2019), the United States (Just et al., 2025), Asia (South Korea: S. Yang et al., 2023; Thailand: Logothetis et al., 2023; China: Tian et al., 2023), and the Middle East (Israel: Shtein et al., 2020), demonstrating the versatility and scalability of ML-based exposure modeling for environmental and public health applications.

1.5 Research Contribution

The main objective of this thesis is to develop hybrid Machine Learning approaches that integrate ground-based observations, satellite data, and other spatiotemporal covariates to predict Ta and PM at fine spatial and temporal scales (≤1 km, daily). The work is carried out in two geographical contexts: Israel and continental France. For Ta, we develop a framework applicable to both countries, demonstrating high performance under different climatic and geographical conditions. For particulate matter, the focus is on Israel, where previous studies provide a solid foundation, but where new methods are explored to improve the accuracy of exposure estimates. In all cases, multi-stage approaches are employed to produce continuous, high-resolution datasets, as it is traditionally done.

Moreover, this thesis introduces an innovation applying common ML techniques known as feature engineering, which allows the derivation of new covariates from existing data, enhancing model performance. As the term suggests, data is "engineered", creating additional variables. Furthermore, these new variables have been shown to substantially improve model

accuracy without sacrificing variance, particularly for environmental exposure models, where spatial relationships are often implicit and typically ignored in tabular data formats

1.6 Research Questions

This thesis is guided by a set of research questions:

1. How can hybrid Machine Learning frameworks be designed to improve the prediction of air temperature (Ta) and particulate matter (PM) at high resolution?
2. Is it worthwhile to invest in Feature Engineering to enhance exposure modeling approaches? How much can a model improve compared to those that do not use it?
3. To what extent do satellite data and spatial covariates, combined with Feature Engineering, enhance predictions of models trained on monitoring stations?
4. How well can these modeling strategies be transferred across regions with different environmental conditions? Can recent machine learning methods significantly advance Ta and PM modeling in Israel and France?
5. Since spatial context plays a crucial role in environmental exposure, what are the best ways to evaluate and validate the robustness of these models? How sensitive are these evaluations when using different validation approaches?

Chapter 2: Literature Review

2.1 Evolution of Spatiotemporal Approaches in Environmental Modeling

Early Phases and Physical Models

The need for exposure modeling has long been recognized, as air quality control and exposure assessment are essential for many real-world applications. Over the years, various approaches have been developed to address these concerns, illustrating both the evolution of exposure modeling and its continued relevance. The earliest models, for example, were designed to describe how pollutants such as particulate matter and gaseous emissions disperse in the atmosphere. Rooted in simplified physics, these approaches included the Gaussian plume formulation, which represented the spread of emissions from point sources, such as smokestacks, by accounting for wind speed, wind direction, and atmospheric stability (Sutton, 1947). These models enabled regulators to estimate pollutant concentrations and assess compliance with emerging air quality standards. However, their applicability was limited to relatively simple and localized scenarios, and they could not capture the more complex spatial patterns of air pollution observed in urban environments (Seinfeld and Pandis, 2016).

Geostatistics and the Need to Fill Spatial Gaps

As modeling approaches evolved beyond purely physical formulations, the focus shifted toward addressing the limitations of sparse monitoring networks, which often failed to capture spatial variations in air pollution and other environmental variables. This underscored the importance of methods capable of filling observational gaps and producing continuous exposure fields across large regions. By the late 20th century, health research increasingly linked pollutants such as $PM_{2.5}$, NO_2, and even Ta variations to respiratory and cardiovascular illnesses. To overcome these limitations on data, geostatistical methods such as kriging were employed to interpolate pollutant and temperature values across unmonitored areas using spatial correlations.

For example, in the early 1990s, the Japan International Cooperation Agency (JICA, 1990) conducted a study on air pollution in Bogotá, Colombia. This study demonstrated the limitations of sparse monitoring networks in capturing spatial variability of pollutants. To address these gaps, the researchers applied geostatistical methods, including kriging, to interpolate air quality data across the city. This approach enabled the creation of continuous

exposure maps, offering a more comprehensive understanding of urban air pollution distribution. The findings highlighted the importance of integrating geostatistical techniques to overcome data gaps and support effective air quality management strategies.

These continuous exposure surfaces became essential tools in environmental epidemiology, allowing researchers to link environmental exposures to health outcomes even in regions with limited monitoring (Jerrett et al., 2005; Hoek et al., 2008)

Land-Use Regression and Assimilation Techniques for Urban Health Studies

In the early 2000s, urban-scale health studies required detailed exposure estimates for pollutants such as NO_2 and $PM_{2.5}$, which vary strongly near roads and across local areas. Land-Use Regression (LUR) emerged as an effective solution, linking pollutant measurements to GIS-derived predictors such as traffic intensity, land cover, and population density. LUR was particularly important at that time because it captured fine-scale spatial variability that simpler interpolation methods or traditional multiple linear regression could not, while remaining interpretable and feasible given the relatively sparse monitoring networks.

Although early machine learning approaches—such as decision trees, Random Forests, and Gradient Boosting Machines—existed, they were not widely applied in exposure modeling due to the larger datasets and computational resources required. LUR's simplicity, scalability, and strong predictive performance made it the dominant approach for urban exposure assessment. For instance, in the large-scale European ESCAPE project, which included 20 towns across Europe (e.g., Amsterdam, Barcelona, Rome, and Stockholm), LUR models for NO_2 and PM achieved cross-validated R^2 values up to 0.83 and 0.81, respectively, enabling quantification of chronic exposure and investigation of cardiovascular and respiratory health outcomes (Wang et al., 2013).

To address the temporal and spatial limitations of Land-Use Regression (LUR) models, researchers began incorporating data assimilation techniques into exposure modeling. These methods combine observational data with model outputs to enhance the accuracy and resolution of exposure estimates. For example, Russell et al. (2011) developed a high-resolution nitrogen dioxide (NO_2) retrieval algorithm for the Ozone Monitoring Instrument (OMI) on NASA's Aura satellite. By applying advanced spectral fitting and air mass factor

calculations, they achieved a spatial resolution of approximately 50 km daily at mid-latitudes, improving the representation of NO_2. This approach enabled estimation of daily exposure levels, allowing researchers to examine short-term exposure spikes and their associated health impacts, insights that annual-average models alone could not provide.

On the other hand, in France, M. Bentayeb et al. (2012) produced daily (~2 km) $PM_{2.5}$ and other exposure datasets by combining ground-based measurements with chemistry-transport models using data assimilation techniques. Their study, covering 1989–2008, enabled the reconstruction of long-term exposure patterns and the assessment of associated health impacts across the country. Similarly, in the United States, researchers applied data assimilation to integrate satellite-derived aerosol optical depth (AOD) with a size-resolved sectional chemical transport model to estimate $PM_{2.5}$ concentrations. This approach allowed the reconstruction of both long- and short-term, spatially resolved exposures at 22 km resolution (hourly), while accounting for observational constraints, multi-wavelength retrievals, and fine-mode aerosol contributions (Saide et al., 2013).

One last example is the study by Qin et al. (2017). The researchers estimated daily ground-level NO_2 concentrations by integrating NO_2 columns from the Ozone Monitoring Instrument (OMI) with meteorological and land-use data using a geographically and temporally weighted regression (GTWR) model. While the original OMI data had a spatial resolution of approximately 13×24 km², the GTWR model produced predictions at a finer ~10 km² resolution across central-eastern China. The resulting daily NO_2 estimates showed strong agreement with ground-based measurements, demonstrating the model's utility for environmental health and epidemiological applications.

These integrated approaches have demonstrated the significant progress yielded by land-use regression and data assimilation techniques for environmental exposure purposes, enabling more precise quantification of chronic exposure and detailed examination of associations with cardiovascular and respiratory outcomes.

Hybrid Machine Learning in Exposure Modeling

By the mid-2010s, a methodological shift occurred in exposure modeling: researchers began leveraging high-resolution satellite products, such as MODIS LST and MODIS AOD, for variables like Ta and PM. These products enabled much finer spatiotemporal characterization

of environmental exposures, achieving high spatial resolutions (≤1 km) and daily temporal coverage.

Some of the earliest applications focused on air temperature and PM modeling. For instance, Kloog et al. (2012) developed one of the first hybrid models combining MODIS AOD with local land-use regression (LUR) and ground-based $PM_{2.5}$ measurements to estimate daily $PM_{2.5}$ concentrations across the Mid-Atlantic United States. The model also incorporated meteorological predictors, including temperature, wind speed, and humidity, to enhance temporal and spatial accuracy. By fusing these diverse data sources, the study produced high-resolution daily $PM_{2.5}$ estimates at 1 km, achieving a cross-validated R^2 of 0.8.

Zhang et al. (2013) combined MODIS LST time series with ground station observations and geographically weighted regression (GWR) to produce daily Ta estimates at 1 km resolution over the Tibetan Plateau. This study was among the first to demonstrate that MODIS LST could capture Ta variability in high-altitude, complex-terrain environments with sparse monitoring networks. In France, Laaidi et al. (2012) used 1 km MODIS LST as a proxy for Ta during the 2003 Paris heatwave, enabling epidemiological analysis of daily mortality impacts. Later, Kloog et al. (2017) also in France calibrated MODIS LST against ground meteorological observations using mixed regression and spatial smoothing, producing daily Ta estimates at 1 km resolution with a cross-validated R^2 of approximately 0.95.

Parallel developments occurred for particulate matter (PM). In Israel, Kloog et al. (2015) were among the first to use 1 km MODIS-MAIAC AOD in mixed-effects models to predict daily PM2.5 and PM10, achieving cross-validated R^2 ranging from 0.72 to 0.79, respectively. Around the same time, Sorek-Hamer et al. (2015) improved $PM_{2.5}$ prediction over bright desert surfaces by integrating Deep Blue and Dark Target AOD retrievals. These pioneering studies demonstrated the feasibility of using satellite products as high-resolution inputs for exposure modeling and provided the foundation for the development of the modern frameworks used to this purpose.

By this period, machine learning models, particularly Random Forest, became increasingly popular in exposure modeling due to their ability to integrate diverse predictors (e.g., satellite data, meteorology, land use) and capture nonlinear relationships. Building on the earlier work of Laaidi et al. (2012) and Kloog et al. (2017), Hough et al. (2020) developed a multi-resolution model combining MODIS LST with auxiliary predictors (elevation, NDVI, land

cover), generating daily minimum, mean, and maximum Ta estimates at 1 km resolution in France, plus the improvement of spatial resolution to 200 m in major French urban areas.

In Korea, a Random Forest model was used to estimate daily PM2.5 concentrations at 1 km resolution with good performance ($R^2 = 0.857$), enabling refined assessments of health effects linked to daily exposure variations (Kim, Lee, & Lee, 2023). In Italy, spatiotemporal Random Forest models were applied to daily PM_{10} and $PM_{2.5}$, achieving cross-validated R^2 values of up to 0.86 for 1 km resolution predictions, demonstrating their utility for both long-term and short-term health research (Stafoggia et al., 2019). Moreover, in China, a national-scale Random Forest model incorporating meteorological, satellite, chemical transport, and geographic predictors produced daily, monthly, and annual $PM_{2.5}$ estimates with R^2 values between 0.85 and 0.90, providing critical insights for health risk evaluations over extended periods (Wei et al., 2020).

Deep Learning and Novel Machine Learning Approaches for the Modeling of Spatiotemporal Dynamics

While traditional machine learning has advanced the modeling of environmental exposures considerably, emerging deep learning techniques are enabling even more sophisticated representations of spatiotemporal dynamics. For example, Convolutional Neural Networks (CNNs) have been applied to integrate ground-based monitoring data with satellite-derived aerosol optical depth (AOD) to generate highly accurate, high-resolution $PM_{2.5}$ surfaces across the United States (RMSE = 2.786 $\mu g/m^3$) at 1 km spatial and daily temporal resolution (Di et al., 2019). Similarly, hybrid deep learning frameworks have been developed to downscale land surface temperature and quantify urban heat island intensity at intra-urban scales by directly learning spatial features from remote sensing imagery, with spatial resolution varying by city (Zheng et al., 2021). In addition, Long Short-Term Memory (LSTM) networks, though often designed for predictive tasks, have been adapted to reconstruct missing spatiotemporal exposure data in monitoring networks, thereby improving $PM_{2.5}$ coverage in regions such as Beijing, China (Cao et al., 2018).

Beyond deep learning, newer strategies such as feature engineering and ensemble learning remain central to exposure modeling. Gradient boosting machines and random forests that incorporate meteorological, land-use, and traffic covariates have produced daily PM2.5 surfaces at 1 km resolution across the U.S. and Europe (Di et al., 2019; Chen et al., 2022),

achieving strong predictive performance (e.g., RMSE = 2.786 μg/m³, MAE = 3.54 μg/m³). Similarly, Barbalat et al. (2022) developed an ensemble model to predict daily NO_2 concentrations at 1 km resolution in France, supported by advanced spatial-blocking cross-validation to reduce overfitting.

The previously highlighted studies played a key role in tracing the development of environmental exposure research over time and in demonstrating their contribution to epidemiological studies. A precise representation of short-term exposure peaks is particularly important for examining spikes in respiratory illnesses or heat-related mortality, and this became possible through advances in modeling approaches. Figure 1 presents a summary timeline highlighting the most important milestones in the evolution of exposure models for Ta and PM.

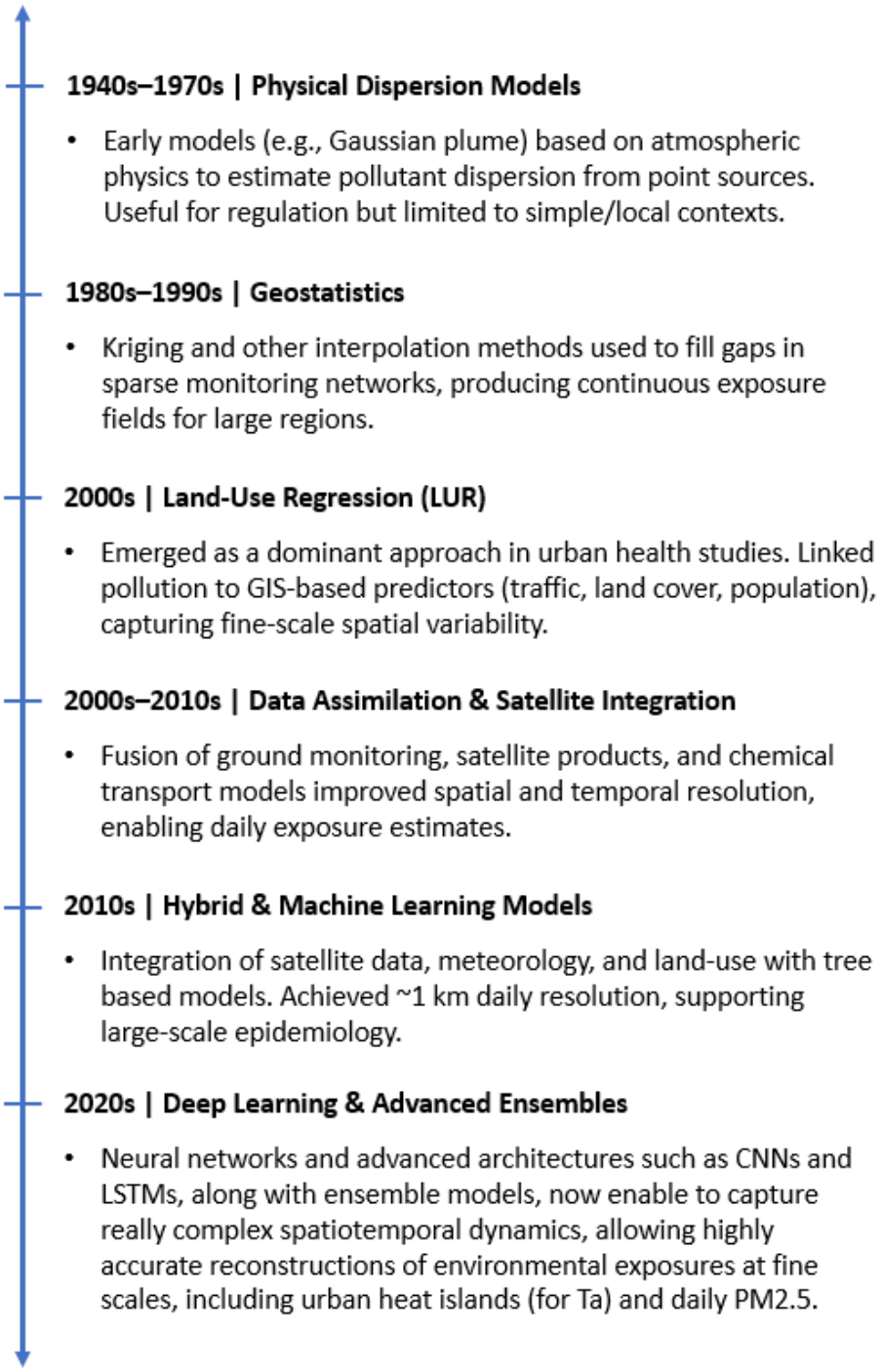


Fig. 1. Timeline of environmental exposure modelling for Ta and PM models.

2.2 Air temperature Modeling with Satellite Data

As outlined in the preceding section, numerous studies have been developed to predict Ta using models that incorporate satellite-derived LST. The following studies highlight prior

research across diverse geographical contexts, illustrating how the integration of satellite data has advanced exposure models for this variable. We comment on them principal strengths and limitations

The impact of heat islands on mortality in Paris during the August 2003 heat wave (Laaidi et al., 2012)

The early 2000s were marked by studies employing regression-based approaches that examined the relationship between satellite-derived LST and Ta. These investigations produced a range of methodological frameworks, each adapted to regional conditions and the availability of ground monitoring networks.

In parallel, local epidemiological studies highlight the serious impact of extreme temperatures in countries like France, where summers have historically posed significant risks due to heatwaves and limited protective measures for vulnerable populations. The elderly, particularly those living alone, are often most affected (Fouillet et al., 2006). Laaidi et al. (2012) investigated the use of surface temperature (Ts) from NOAA (National Oceanic and Atmospheric Administration) and AVHRR (Advanced Very High Resolution Radiometer) to assess heat-related morbidity and mortality in the Paris metropolitan area. They combined daily Ts images with Ta records from weather stations during the summer heatwave that impacted the region from 1 to 13 August 2003.

Exposure was studied in a small case–control group composed by 241 individuals, aged 65 or more, who died during the wave. Their analysis included comparisons between satellite-derived and ground-based measurements across both urban and rural areas, which required estimating daily Ta (minimal, mean, maximal) at fine resolution (1 km). Although no specific statistical model was reported, it was evident the study employed a hybrid approach, integrating Ts images, Ta records, the normalized difference vegetation index (NDVI), and land-cover data.

Their results confirmed a strong association between Ts and Ta, but also revealed systematic biases in urban areas. These biases were attributed to the urban heat island effect and to the mismatch between Ts and Ta, since satellites measure the temperature of skin surfaces, such as roofs, roads, and vegetation, from satellites, while weather stations record Ta at 1.5–2 m above the ground. Using their estimated Ta, they reported that the highest daytime

temperature exceeded 40 °C after 1st of August on several days (3rd, 4th, 6th, and 10th), and that at night the heat island effect was concentrated in central Paris, whereas during the day, temperature anomalies appeared more scattered across densely built and industrial zones.

This study was particularly important for the French context, as it was among the first in the country to predict Ta using satellite-derived products. Although the authors did not apply formal validation strategies such as cross-validation or similar techniques to assess model performance, they evaluated the reliability of their results using established epidemiological metrics, drawing on their expertise in the field. For example, they found that a 0.41°C increase in average Ta min during the 1–13 August was associated with an odds ratio (OR) of 2.17 (with 95% confidence interval=1.14–4.16), indicating that even small increases in nighttime Ta substantially raised the likelihood of death. Similarly, a 0.51°C increase in the 6-day pre-death average Ta min corresponded to an OR of 2.24 (with 95% confidence interval=1.03–4.87). These findings led the authors to conclude that prolonged exposure of elderly individuals to elevated nighttime temperatures during heatwaves increases the risk of mortality, while daytime temperature exposure also plays a role, though less pronounced. This study further emphasized that with higher-resolution estimates, the obtained ORs would likely be even larger, this remarking on the importance for improving the spatial resolution of such models.

Finally, they researchers pointed out that when Ta exceeds 35°C, vegetation can provide a beneficial cooling effect by reducing both the urban heat island effect and air pollution, supporting the promotion of tree planting and green roofs in urban zones. Results were consistent with previous studies using high-resolution Landsat data, where a 2.13 °C increase in temperature yielded an OR of 2.98 (with 95% confidence interval=1.37–6.50). The confidence intervals indicate the uncertainty around each odds ratio, reflecting the reliability of the model's results.

MODIS-based estimation of air temperature of the Tibetan Plateau (Yao & Zhang, 2013)

Another pioneering study linking satellite data and Ta was conducted in Asia. Zhang et al. (2013) investigated the relationship between land surface temperature (LST) and Ta on the Tibetan Plateau, a high-altitude region identified as a major heat source influencing the climate of the Eurasian continent. They modeled Ta for the period 2001–2007 using 1 km

daily LST (Ts), together with elevation data from the ASTER Global Digital Elevation Model (GDEM), surface topography predictors, and meteorological records from 137 monitoring stations. Their framework was structured into the following three stages:

- In the first stage, they performed a time series analysis to better understand the relationship between Ta and Ts.
- In the second stage, they applied an ordinary least squares (OLS) analysis to identify and exclude stations with outlier measurements.
- In the third stage, they used a geographically weighted regression (GWR) to estimate Ta, calculating location-specific parameters like elevation to adjust their predictions. For the scope of the study, they found it relevant using elevation as location-specific parameter within the GWR regression.

We do not provide a summarized formulation of all the revised studies within this dissertation. However, we deemed it pertinent to illustrate how this kind of early approach was used to model the dynamics of Ta. Equation 1 presents the formulation used in this study, where Ta represents air temperature, Ts is MODIS LST, *h* denotes altitude, *u* indicates a specific spatial location, and $i = 1 \ldots n$ corresponds the spatial sites at which Ta is to be predicted.

$$Ta_i(\mathrm{u}) = \beta_{0i}(u) + \beta_{1i}(u) Ts_i(u) + \beta_{2i}(u) h_i(u) \quad \text{Eq. 1}$$

Though sites at higher elevations exhibited larger negative $\beta_{2i}(u)$ values, reflecting the expected decrease in Ta with altitude, regions with higher Ts generally showed large $\beta_{1i}(u)$, confirming the strong positive relationship between Ts and Ta. Overall, the model effectively simulated the temperature dynamics of the region.

To sum up, this study addressed the limitations of the epoch, particularly the limited capacity of the models to predict Ta at fine scales across large terrains, requiring the use of lower-resolution estimates to cover larger areas, while still successfully capturing the dynamics of Ta on the Tibetan Plateau. Monthly mean Ta was estimated for 2001–2007, achieving RMSE performance between 1.13 °C and 1.53 °C. This work represents one of the earliest attempts to employ satellite products for spatially modeling Ta in complex mountainous regions with sparse in-situ coverage.

The authors noted that substantial discrepancies emerged in the modeling under cloudy conditions and during periods of pronounced diurnal variability, particularly in valley areas. The study provided valuable insights for Ta modeling, but one aspect that was not emphasized at the time was cross-validation, as the only validation they performed involved holding out three stations out of 136, with no comments on their spatial distribution or similar.

It is also important to note that, although the use of various regression algorithms was not yet common in exposure modeling, by the early 2000s machine learning models such as Random Forests and Decision Trees had already become well-established. Breiman (2001) introduced Random Forests, an ensemble method combining multiple decision trees to improve predictive accuracy and reduce overfitting. These approaches offered enhanced performance and flexibility, addressing several limitations of earlier regression-based models. Although it took some time for them to gain widespread recognition, by 2004 the *randomForest* package was available in R, and in 2010 the *scikit-learn* library was released in Python.

Years later, subsequent studies sought to systematically address the issues with models' capacity, temporal and spatial resolutions, and validation drawbacks.

Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data (Kloog et al., 2017)

Seven years after the research done by Laaidi et al. (2012), a new study marked a significant milestone in the development of strategies using satellite-derived LST to estimate environmental variables such as Ta. This work built on earlier contributions by the lead author (I. Kloog), including the modeling of particulate matter concentrations in the U.S. (Kloog et al., 2011), and was among the first to expand the integration of satellite data into exposure modeling. The study introduced hybrid and multistage modeling approaches, which have since become widely adopted to address these challenges. Building on this framework, Kloog et al. (2017) developed a model in France using LST, defined in this study as Ts, to generate daily Ta surfaces at 1km resolution across the entire French metropolitan area, excluding Corsica.

Their methodology employed a linear mixed-effects model with day-specific random effects, integrating Ts with ground monitoring data from Météo-France, a French national meteorological institution, alongside other relevant predictors. A key innovation of the study

was the inclusion of a dedicated stage to address gaps in satellite data caused by cloud cover in the MODIS LST instruments, which enabled the generation of continuous spatiotemporal Ta estimates for the period 2000–2011. These estimates were subsequently used for epidemiological exposure assessments. Model performance was excellent, with a mean out-of-sample R^2 of 0.95.

Validation was conducted using robust approaches, including 10-fold out-of-sample cross-validation at both modeling stages: in stage 1, during the calibration of satellite-derived Ts against ground Ta, and in stage 2, when predicting Ta in grid cells with missing Ts, where the Ta–Ts relationship from stage 1 could not be directly leveraged. Both stages achieved a mean R^2 of approximately 0.95 with a low RMSPE (root mean squared prediction error—similar to RMSE but calculated specifically on unseen data) ranging from 1.52 to 1.68 °C, demonstrating the high accuracy and reliability of their model outputs.

The methodology of this study was structured into two modelling stages:

1. Stage 1: Monitored Ta was calibrated against Ts. This calibration incorporated all relevant spatiotemporal predictors, including NDVI, elevation, population density, and percent urban cover, among others.
2. Stage 2: For locations where Ts data was missed, for example, due to cloud cover, a 60 km buffer around each prediction site was used to calculate the daily average from nearby Ta monitors. These averaged values were then used as inputs in a regression model with a fixed intercept and a slope varying by spatial location.

The final Ta predictions combined outputs from both stages: sites with available Ts were predicted using the Stage 1 model (denoted on Equation 2), while sites with missing Ts were predicted using the Stage 2 model (denoted on Equation 3). Given their relevance, this dissertation includes Equations 2 and 3 to clearly summarize and understand this study's formulation.

Stage 1 modelling equation in Kloog et al. (2017)

$$Ta_{ij} = (\alpha + u_j) + (\beta_1 + v_j)Ts_{aij} + \beta_2 Percent\ urban_i + \beta_3 Elevation_i + \beta_4 NDVI_{ij} + e_{ij}(u_j, v_j) \sim [(0,0), \Sigma] \quad \text{Eq. 2}$$

where Ta_{ij} represents the ground measured Ta at a spatial site i on a day j; α and u_j are the fixed and random intercepts. Ts_{ij} denotes the Ts value of site i on a day j; β_1 and v_j denote the fixed and random slopes on day j, respectively. Percent urban is the percent of urbanicity at site i, $NDVI_{ij}$ denotes the monthly averaged NDVI value at site i in which day j falls. Elevation denotes mean elevation of the site i. Σ is the unstructured variance–covariance matrix for the random effects and e_{ij} is the error term at site i on a day j.

Stage 2 modelling equation in Kloog et al. (2017)

$$predTa_{ij} = \alpha_i + \beta_i mTa_{jr} + E_{ij} \quad \text{Eq. 3}$$

where $Pred\text{Ta}_{ij}$ is the prediction of Ta at site i on a day j from the calibration stage; $m\text{Ta}_{jr}$ is mean Ta in the specific buffer r on a day j; α denotes the fixed intercepts and β_1 denotes the fixed slopes that varies by spatial site i, and E_{ij} is the error term at site i on a day j (in this stage).

As discussed earlier, the methodology exposed in this study was essentially hybrid, addressing predictions through multiple approaches in a synergistic manner. With the advent of approaches like this, exposure modeling using hybrid approaches has become the baseline for many subsequent studies. This study introduced several innovations, including the application of formal validation approaches evaluating model error across different regions and time periods. These advancements not only supported high-performing studies in smaller areas, such as New England in the U.S. (Kloog et al., 2011), but also demonstrated that these methods could be successfully extended to larger territories such as France.

To conclude, this work was particularly relevant for the fields of epidemiology and public health, which had previously relied on assigning population exposure based on the nearest monitoring station. One of the study's key findings was a significant association between estimated Ta and reduced birth weight across multiple exposure windows, adjusting also for PM2.5 levels, as done in other studies (e.g., Wells & Cole, 2002; Kloog et al., 2012b). The authors emphasized that, in contrast to the results obtained using their modeling approach, none of the exposure periods derived from nearest-station datasets obtained significant associations; in some cases, the associations were even positive. Which supports the idea of environmental modelling following hybrid approaches for better epidemiological conclusions.

A multi-resolution air temperature model for France from MODIS and Landsat thermal data (Hough et al., 2020)

A subsequent study under the direction of I. Kloog further advanced the modeling of Ta in France. Hough et al. (2020) developed a multi-resolution daily Ta model spanning 2000–2016, integrating MODIS LST with a suite of spatiotemporal predictors, including NDVI, land cover, and reanalysis temperature data. Similar to the previous study in this context, linear mixed-effects models were used to calibrate daily Ta values (minimum, mean, and maximum) at a 1 km base resolution, achieving excellent performance ($R^2 \geq 0.92$, RMSE < 2 °C). To capture finer spatial detail in urban areas, the authors downscaled these 1 km estimates to 200 m resolution using Landsat thermal data, producing high-resolution Ta surfaces that more accurately represented urban microclimates.

The work of Hough et al. (2020) builds upon the hybrid approach introduced by Kloog et al. (2017), extending it through a multistage framework that includes the following stages:

- In the first stage, they calibrated Ta at each monitoring station as a function of 1 km LST and emissivity (the latter not included in previous studies), monthly 1 km NDVI, elevation, population, and land cover. A linear mixed model was used to regress Ta against these spatiotemporal variables, capturing the relationship between Ta and LST—similar to the approach in Kloog et al. (2017).
- In stage 2, they used inverse distance weighting (IDW) of observed Ta as input to a secondary linear mixed model at sites where LST was unavailable, regressing Ta against the IDW estimates. Applying both stages at their corresponding sites—those with available and unavailable LST—they produced the final 1 km Ta model.
- Stages 3 and 4 focused on enhancing spatial resolution to 200 m in urban areas. In stage 3, the residuals from the 1 km model were regressed against higher-resolution predictors—including monthly averaged Landsat LST, climatic regions, latitude, longitude, Landsat NDVI, population, elevation, and land cover, using random forest (RF) and XGBoost models. In stage 4, the predictions from RF and XGBoost were ensembled using a generalized additive model, leveraging the strengths of each base learner at different locations.

We considered equations 4 and 5 particularly relevant and include them in this dissertation to summarize and understand the formulation of this study's downscaling stages.

Stage 3 modelling equation in Hough et al. (2020).

$$R_{ij} = f(Ta_{p_{1km_{ij}}}, Tb_{im} + NDVI_{im} + Land\ Cover_{ily}, Climate_i,$$

$$Elevation_i, Population_i$$

$$X_i, Y_i, j) + e_{ij} \quad \text{Eq. 4}$$

where R_{ij} is the residual of the 1 km Ta model associated with 200 m grid cell *i* on day *j*, *f* designates the random forest or extreme gradient boosting function, $Ta_{p1km_{ij}}$ is the 1 km T_a model prediction associated with 200 m grid cell *i* on day *j*, Tb_{im} is the Landsat top-of-atmosphere brightness temperature, the others are the high-resolution predictors mentioned for this model, and e_{ij} is the error term at cell *i* on a day *j*.

Stage 4 modelling equation (ensemble learner) in Hough et al. (2020).

$$R_{ij} = \mathrm{t}(X_i, Y_i)R_{p_rf_{ij}} + \mathrm{t}(X_i, Y_i)R_{p_gb_{ij}} \quad \text{Eq. 5}$$

where R_{ij} is the residual of the 1 km T_a model associated with 200 m grid cell *i* on day *j*, *t*(x_i, y_i) is a tensor product smooth of the x and y coordinates of cell *I*, R_{p_rfij} and R_{p_gbij} are the predicted residuals of the 1 km T_a model from the stage 3 random forest and XGBoost model, respectively, for cell *i* on day *j*, and ε_{ij} is the error for cell *i* on day *j*.

This study introduced several innovations, including the use of novel tree-based algorithms, the downscaling of spatial resolution, and the application of a formal validation approach using 10-fold cross-validation, similar to the methodology applied by Kloog et al. (2015).

Estimation of hourly near-surface air temperature across Israel using an ensemble model (Zhou et al., 2020)

Once machine learning became more widely adopted, and with the recognition of strategies that performed well for environmental exposure assessment, models began to implement more comprehensive methodologies addressing finer spatial and temporal resolutions. Studies increasingly relied on freely available spaceborne data and other periodically maintained datasets (e.g., reanalysis products, SEVIRI) to improve the spatial and temporal resolution of their estimates. This was illustrated by the following study conducted in Israel, whose main innovation was producing estimates at fine spatial (1 km) and hourly resolution.

In the Middle East, under the guidance of I. Kloog, Bin Zhou et al. (2020) evaluated the utility of MODIS LST for estimating Ta across Israel's diverse climatic zones, including coastal Mediterranean, semi-arid, and desert regions, over the period 2004–2017.

The study used daily 1 km MODIS LST data, integrating it with weather station records and other predictors to develop regression models designed to capture Israel's diverse climatic and geographic conditions. The results showed that, although MODIS LST provided a useful baseline, model performance varied across these contrasting environments, reflecting Israel's sharply changing topography. Their models performed best in coastal Mediterranean areas but were less reliable in arid zones, where substantial discrepancies exist between LST, called by them Ts, and Ta.

This study demonstrated the need for models that adapt to local conditions in heterogeneous landscapes. The authors also improved time coverage of their models by using hourly 4 km Ts data from the SEVIRI instrument, a satellite sensor launched by the European Space Agency, providing frequent observations of the Earth's surface and atmosphere.

To achieve this, Ta was modeled using a three-stage framework, which is detailed as follows:

- In stage 1, missing values in the hourly 4 km SEVIRI Ts were imputed by training both a Random Forest (RF) and an XGBoost model using spatiotemporal predictors like latitude, longitude, hour, day, day of the year (DoY), and synoptic variables from ERA5 reanalysis data. To manage the large volume of data, the models were trained separately for each month of the year. Finally, a generalized additive model (GAM) was used to ensemble the predictions from the two learners.
- In stage 2, Ta was estimated using the hourly gap-free SEVIRI Ts data, combined with hourly Ta values from monitoring stations, at a spatial resolution of 4 km. Random Forest (RF) and XGBoost models were trained and ensembled to estimate hourly Ta at 4 km resolution.
- In stage 3, Ta predictions were downscaled to 1 km resolution by estimating residuals. To address the limited temporal coverage of MODIS LST at 1 km, the authors used a Random Forest approach based on the daily MODIS overpass hours (10 am–2 pm and 9 pm–2 am) to create a multi-annual monthly mean diurnal pattern at 1 km. They then applied this monthly mean hourly MODIS Ts pattern, along with a first-order Fourier curve, to estimate Ts values for hours not covered by MODIS overpasses (3 am–9 am

and 3 pm–8 pm). Finally, the residuals from the 4 km Ta estimates of stage 2 were modeled using both RF and XGBoost, and the predictions were ensembled to produce the final hourly Ta at 1 km resolution.

For each stage of their modeling, performance was evaluated using standard 5-fold cross-validation. Key innovations of this study were the temporal downscaling method and the ensemble of commonly used machine learning models, including RF and XGBoost.

From this study, it is important to note that although ensemble learning has the potential to improve predictive performance, its application is complex and requires careful consideration. The authors implemented ensemble methods at multiple stages, but did not quantify in advance the contribution of each ensemble to prediction improvement. Typically, residual analysis, assessing correlations or systematic patterns, is performed beforehand to evaluate whether ensembling is likely to provide meaningful gains. As Breiman (1994) demonstrated with bootstrap aggregating (bagging), reducing prediction variance depends on understanding the residual behavior of individual models before combining them.

Modelling daily air temperature at a fine spatial resolution dealing with challenging meteorological phenomena and topography in Switzerland (Fluckiger et al., 2022)

Up to this point, methodologies for exposure modeling of Ta largely followed a standardized protocol, almost as if adhering to a strict recipe. This underscored the need for innovation and spurred the development of new approaches aimed at improving the reproduction of generalization capacity at previously unobserved locations. In this context, it became evident that traditional cross-validation schemes could yield overly optimistic evaluations, which led to the adoption of spatial blocking strategies to provide a more realistic assessment of model performance.

In Switzerland, Flückiger et al. (2022) used MODIS LST to reconstruct daily Ta across the country's mountainous terrain. They combined these satellite products with high-altitude meteorological stations, showing that while the data captured broad heating gradients, model performance declined in narrow valleys and under frequent cloud contamination. In this study, the authors filled gaps in MODIS LST data and then calibrated ground-based Ta using the gap-filled predictor along with other spatiotemporal variables, including NDVI, impervious surface fraction, and elevation.

The methodology in this study was conducted in two stages.

- In the first stage, missing values in MODIS LST from all Aqua and Terra overpasses were imputed. For each year and satellite overpass, a RF model was trained using 5-fold cross-validation on 80% of the available data. The model was then evaluated on a 20% independent hold-out set, and the resulting best-performing model was used to impute missing MODIS LST values.
- In stage two, the authors used the gap-filled MODIS LST along with other spatiotemporal predictors to calibrate ground-based Ta and generate predictions. Unlike stage one, they stratified the monitoring data such that the cross-validation procedure consisted of 10 folds, each representing a different area of the region, with approximately 10% of station measurements in each fold.

The cross-validation scheme in this study was designed to leave out an entire area in each iteration, simulating the real performance of their models spatially. The authors compared their models with a leading local product from MeteoSwiss, which produces daily 2 km Ta, and found that their approach outperformed it in both instances: the downscaled MeteoSwiss model to 1 km using interpolation, and the upscaled version of the study's model aggregated to 2 km. Their models were particularly better at predicting high temperatures in valleys, where the MeteoSwiss model exhibit underestimations. Based on these findings, they recommended allocating more monitoring stations in Switzerland urban areas to reduce bias in epidemiological studies.

In sum, their models achieved fine spatial resolution, producing daily estimates at 100 m for the period 2003–2018. Model performance, measured by RMSE, ranged from 1.05 to 1.86 °C using RF as their regression learner across the different stages of modeling. This study highlighted the importance of constructing validation sets to ensure spatial robustness, with homogeneous sample sizes of selected monitoring stations covering the entire study area. It also implemented new types of cross-validation approaches. In stage 1, traditional 5-fold cross-validation was applied to 80% of randomly selected grid cells, with 20% held out for independent validation. In stage 2, spatially stratified cross-validation was used, leaving out one of ten geographic areas at a time to account for spatial correlation and test the model on unseen regions. This hybrid approach ensured both an overall performance assessment and spatial validation robustness.

Predicting spatiotemporally-resolved mean air temperature over Sweden from satellite data using an ensemble model (Jin et al., 2022).

Similar to Flückiger et al. (2022), recent studies have employed machine-learning frameworks that improve exposure modeling of Ta while addressing the limitations of prior approaches. Jin et al. (2022) applied such methods in Sweden, integrating daily 1 km MODIS LST (defined by them as Ts like in previous studies) with other spatiotemporal predictors, including land cover type from MCD12Q1 (reclassified into five categories following Hough et al., 2020), NDVI, elevation, and meteorological reanalysis data. They implemented a three-stage framework using generalized additive models (GAMs) to ensemble predictions from four base learners each year. This approach allowed them to generate daily continuous Ta surfaces (2001–2019), achieving high accuracy under challenging conditions such as snow cover and seasonal transitions, with an overall root mean square error (RMSE) of 1.38 °C.

To build their models, authors used the following three-stage methodology:

- In the first stage, they predicted missing values of Ts using averaged hourly ERA5 2 m Ta at ~10 km spatial resolution, applying a simple linear mixed-effects (LME) model.
- In the second stage, they predicted 1 km daily Ta using predictors such as elevation, NDVI, land cover, coordinates, and day of the year, training 4 base learners, including GAM, GAMM, RF, and XGBoost.
- In stage 3, they combined the out-of-sample predictions from the four base learners using another GAM to generate the final Ta predictions.

They validated their final estimates using a hybrid spatial cross-validation scheme, similar to Flückiger et al. (2022). The cross-validation followed a two-stage methodology: first, the four base models (GAM, GAMM, RF, and XGBoost) were trained on eight station groups and tested on one, while a separate group was held out for ensemble validation. Then, the ensemble model was trained on the out-of-sample predictions from the base models and tested on the held-out group. This spatially structured cross-validation provided an unbiased evaluation of model performance, analogous to the spatially blocked cross-validation method used in the previous reviewed Swiss study. Predictions were generally more accurate in southern Sweden than in the north, and model performance was better in summer than in winter.

It is important to note that the choice of base learners in this study was deliberate rather than random. Unlike many other studies in the field, the authors carefully considered the ensembling process they intended to implement. In machine learning, diversity among base models is known to improve ensemble performance (Dietterich, 2000).

By combining models from different families, the ensemble can leverage complementary strengths arising from their distinct architectures and optimization strategies. For instance, the authors noted that RF excels at capturing long-term variation in Ta, whereas XGBoost is better at modeling short-term fluctuations. This deliberate selection of complementary models improves both the accuracy and robustness of the final predictions. Linear models are better suited for extrapolation, predicting beyond the range of the monitor stations' values used for training, while tree-based algorithms typically perform well for interpolation. As previously mentioned, a key advantage of ensembling models from different families lies in the diversity of errors they produce, which helps reduce overall prediction bias.

To conclude, their findings highlight that even modest improvements in model accuracy at 1 km resolution can significantly enhance performance on the downscaling of Ta to higher resolutions, showing that prioritizing precision for such environmental models is worthwhile. Beyond the spatial validation, they also compared predicted Ta values with nearby monitoring stations using histograms, confirming that the predicted distributions closely matched the observed distributions, as intended.

XIS-temperature: A daily spatiotemporal machine-learning model for air temperature in the contiguous United States (Just et al., 2024)

In the United States, Just et al. (2024) developed the XIS-temperature model, representing one of the most comprehensive efforts to date. A key innovation in this study was the use of advanced machine-learning strategies, such as feature engineering, to improve the accuracy of daily, continuous, point-based Ta predictions (minimum, mean, and maximum air temperature; T_{min}, T_{mean}, T_{max}) across the contiguous U.S (2003-2023). This approach enabled the model to generate predictions at arbitrary geographic locations, rather than being restricted to a fixed grid, by leveraging inverse distance weighting (IDW). Although the model is based on MODIS data at a 1 km spatial resolution, an internal interpolation step allows it to provide estimates at any desired point.

Their modeling incorporates an IDW-engineered variable derived from averaging Ta monitor measurements within buffers of 100km from the site of interest. The key distinction of this approach is that, to enable predictions at arbitrary geographic locations, the model only uses predictors that are available everywhere. Consequently, variables such as Ts from SEVIRI or reanalysis products are not included. Instead, the model relies mostly on the IDW-interpolated Ta variable, along with latitude, longitude, enhanced vegetation index, surface imperviousness, population density, and distance from watercourses.

Their model achieved weighted root mean square errors (RMSE) for 2023 of 1.78 K for daily T_{min}, 1.19 K for T_{mean}, and 1.48 K for T_{max}, which represented improvements of 40% for T_{min} and 7% for T_{max} in comparison with other leading products tested in the US. Their results were validated using a station-wise cross-validation (CV) approach.

This study shows how satellite observations, combined with advanced statistical learning, can be used to estimate Ta across a continental scale, achieving much higher accuracy than traditional methods. The cross-validation (CV) strategy is also noteworthy: data were split by weather station rather than randomly by day, ensuring that all observations from a given station were either in the training or testing set, thus avoiding "data leakage" from seeing the same location twice.

The engineered IDW features were likewise calculated for each training fold, with stations in the test set excluded in the averages of those used for training to guarantee realism in the performance assessment. Given the large dataset, a 5-fold CV was used, stations were divided into five groups and each group rotated as the test set. To account for the denser distribution of stations in urban areas, observations were weighted by the area each station represents, ensuring fair performance assessment across both rural and urban regions. Predictions were evaluated using RMSE, bias, and standard deviation, and results were analyzed by year, season, and NOAA climate region, including checks for extreme hot/cold days and isolated stations. Additionally, SHAP values were computed during cross-validation (CV) to quantify the contribution of each predictor to the model. This methodology provided a realistic, spatially balanced evaluation of Ta predictive performance across the contiguous U.S.

Key Takeaways from Ta Modeling with MODIS LST

Collectively, all these previous studies demonstrated that integrating MODIS LST for the estimation of Ta constituted a significant and progressive advancement in the field. Its strengths included consistent global coverage, finer spatial resolution (≤1 km), and daily revisit cycles (temporal resolutions ≤ daily), features that help diminishing the burdens of exposure misclassification, and are especially valuable in regions-countries with limited meteorological infrastructure.

As a result, most recent studies model Ta with satellite data and other complementary predictors via advanced statistical approaches (including ML). Looking ahead, improvements in satellite technology and the availability of higher-resolution observations are likely to further enhance the precision of model outputs, making future epidemiological analyses even more accurate. Nevertheless, limitations remain, such as cloud contamination, discrepancies between surface and air temperature, and reduced performance in complex terrains. In addition, satellite data collection is time-consuming and prone to significant errors, underscoring the importance of identifying a limited set of key predictors that can reliably explain the behavior of Ta (Just et al., 2024).

Summary of reviewed studies in this section

A synthesis of the key innovations and limitations of the studies discussed in this section is presented in Table 1.

Table 1. Key innovations, contributions, and limitations of the studies linking Ta with satellite data**.**

Study	Innovations-contributions	Limitations
1. Laaidi et al. (2012)	• First study using satellite data to estimate Ta in France. • Identification of spatial bias due to the urban heat island effect, which demands high-resolution estimates in urban zones.	• No formal estimation approach was reported. • The extent of the study covers only Paris.

Study	Innovations-contributions	Limitations
2. Yao & Zhang, 2013	• One of the first studies using satellite data to estimate Ta.	• Low spatio-temporal resolution in the model. • Minimal validation approach (3 held-out stations). • Old model in comparison to what was already available at the epoch.
3. Kloog et al., 2017	• Proposed a strategy to address the issue of gaps in satellite data. • Use of formal validation approaches that implicitly avoid overfitting. • Large-scale predictions (whole continental France) at fine resolutions (1km, daily).	• Limited ability to capture microclimate variations at very fine spatial scales, highlighting the need for Ta predictions at resolutions finer than 1 km for epidemiological studies. • Still relying on linear regression approaches.
4. Hough et al., 2020	• Applied advanced machine learning algorithms, including RF and XGBoost, and combined their outputs through an ensemble approach. • Downscaled 1 km Ta predictions to 200 m in urban areas, supporting epidemiological studies.	• The core methodology is built on the previous study (Kloog et al., 2017), with new methods applied primarily for downscaling. • Did not directly address gaps in satellite data, instead following the approach of Kloog et al. (2017).
5. Zhou et al., 2020	• Improved temporal resolution from daily to hourly, a challenging and relatively uncommon achievement in the field. • Like other common studies of the epoch used ensemble learning.	• Ensemble learning was applied at multiple stages, but the contribution of each ensemble to improving the modeling was not explicitly quantified before its repeated application. Model ensembling is often preceded by residual analysis, such as

Study	Innovations-contributions	Limitations
		assessing correlations between residuals, to quantify potential improvements in advance (Breiman, 1994). • They did not impute MODIS LST, so the multi-annual monthly mean is likely biased toward seasons other than winter, when the majority of gaps occur.
6. Fluckiger et al., 2022	• Achieved an impressive spatial resolution (100m). • Directly addressed gaps in MODIS LST by imputing missing values using other spatiotemporal variables. • Improved local performance metrics and validated results using robust spatial cross-validation.	• The study only used RF for all the modelling. Although RF has been widely used in spatiotemporal Ta modeling (Zhou et al., 2020), no single model consistently outperforms others, and performance depends on the study context. XGBoost has also been proven effective in environmental exposure modeling (Chen & Guestrin, 2016; Di et al., 2019).
7. Jin et al., 2022	• Ensembled models of different families, which further improves performance. • Their spatial cross-validation approach provides a realistic assessment of model performance.	• For imputing MODIS LST, they used an LME, whereas more recent studies have applied newer methods like tree-based models. • For epidemiological studies, predictions at 1km are not enough.
8. Just et al., 2024	• Their model is easy to reproduce while maintaining strong performance due to careful	• Their models may perform poorly in areas with sparse monitors, as they exclude universally available predictors (e.g., SEVIRI Ts, reanalysis products).

Study	Innovations-contributions	Limitations
	feature selection and evaluation. • For cross-validation, feature engineering was applied usin only training stations at each fold when computing buffer averages, preventing data leakage when testing.	• Relies heavily on IDW-interpolated Ta, which may limit sensitivity to local microclimates.

2.3 PM Modeling with Satellite Data

Similar to the previous section, numerous studies have focused on predicting particulate matter (PM_{10} and $PM_{2.5}$) using models supported by remote sensing data. The following studies illustrate how AOD has been linked to PM across different regions, highlighting the significant progress made in the field. We comment in these literature studies strengths and limitations.

Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements (Kloog et al., 2011)

The foundational contributions of I. Kloog in Ta modelling were actually started to studies addressing air pollution. Kloog et al. (2011) developed a first 10 km high-resolution model for daily exposure to fine particulate matter ($PM_{2.5}$) across New England, USA (2000–2008). Their approach combined satellite-derived aerosol optical depth with ground-based monitoring data, meteorological variables, and land-use characteristics (e.g., major roads and highways, elevation, percent open space) within a three-stage mixed-effects modeling framework. Their study has, surprisingly, been cited more than 400 times.

The authors provided a detailed explanation of why including satellite-derived AOD is relevant, arguing that it offers broader spatial coverage in areas where $PM_{2.5}$ monitors are sparse. They also described AOD as the extinction of light due to aerosol scattering and absorption in the atmospheric column, which is influenced by factors such as aerosol mass concentration, humidity, and effective scale height. The authors further noted that previous attempts to model the relationship between $PM_{2.5}$ and AOD using linear models, mixed-

effects models, or GAMs showed poor predictive performance. In contrast, day-specific calibrations significantly improved the ability of AOD to predict $PM_{2.5}$.

Their prediction methodology is summarized in the following stages:

- In Stage 1, they calibrated AOD against monitored $PM_{2.5}$ to estimate $PM_{2.5}$ at sites with missing AOD values (due to cloud cover, high surface reflectance, etc.) using a mixed-effects regression model, similar to (Kloog et al., 2017).
- In Stage 2, they predicted $PM_{2.5}$ for grid cells with available AOD values using the calibrated model from Stage 1.
- In Stage 3, the goal was to predict $PM_{2.5}$ at sites with missing AOD. They used neighboring predicted $PM_{2.5}$ values from locations with available AOD and regional $PM_{2.5}$ averages, applying spatial smoothing functions such as universal kriging to generate predictions at these sites.
- The following was not reported as a particular stage yet they also applied a downscaling resolution improvement were local land use predictors at higher resolution were used to estimate residuals of the 10km model allowing them mapping $PM_{2.5}$ at 100mts.

To end, they formally validated their models across the different stages. In Stage 1, using 10-fold cross-validation, they obtained year-to-year R^2 values of 0.725–0.904. In Stage 3, model performance was also strong, with year-to-year R^2 values ranging from 0.692 to 0.887.

Mapping particulate matter in alpine regions with satellite and ground-based measurements: An exploratory study for data assimilation (Emili et al., 2011)

The following is another pioneering model in the field that explored the integration of satellite AOD for particulate matter modeling.

Emili et al. (2011) investigated the combination of satellite-derived aerosol optical depth (AOD) and in-situ PM_{10} measurements to produce daily PM_{10} surfaces over the European Alpine region, specifically in Switzerland and South Tyrol, Italy. They compared two approaches for producing PM_{10} surfaces: one using satellite-derived AOD from SEVIRI and MODIS, and another based on interpolation of ground-based measurements. Their results showed that satellite-derived maps generally underperformed, particularly in complex terrain

and urban areas, with RMSE values of ~10–12 μg/m³ for satellites versus ~5–7 μg/m³ for interpolation. They also argued that an assimilation technique combining both approaches would not significantly improve the estimates, as satellite-derived PM_{10} at epoch was too inaccurate to enhance the representation of daily spatial patterns at this regional scale.

Hence, they also applied a vertical correction method using the digital elevation model and the height of the planetary boundary layer, which improved the accuracy of interpolated maps in mountainous regions. Data assimilation combining satellite and ground measurements showed limited benefits under the existing dense monitoring network, but provided modest improvements (~1 μg/m³ in RMSE).

This study was particularly important because it demonstrated that, although AOD data can facilitate spatiotemporal estimation of PM_{10} at finer resolutions, simple interpolation approaches can outperform satellite approaches in complex terrains when monitoring coverage is dense. Satellite-derived PM_{10} is particularly valuable in regions with sparse monitoring, such as areas where stations are spaced ~100 km apart.

Estimating daily $PM_{2.5}$ and PM_{10} across the complex geo-climate region of Israel using MAIAC satellite-based AOD data (Kloog et al. 2015)

Years later, the study of Kloog et al. (2015) took advantage of upgraded satellite products from MODIS to develop a high-resolution (1 km) daily PM model ($PM_{2.5}$ and PM_{10}) in Israel for 2003–2013. The model relied on MODIS–MAIAC AOD data and successfully addressed the challenges posed by the country's complex geo-climatic conditions and sparse ground monitoring network.

Their three-stage modeling framework consisted of the following stages:

- A first stage using a daily mixed-effects model to calibrate PM against AOD.
- In Stage 2, PM is predicted using the calibrated model from the previous stage.
- In Stage 3, 30 km buffers were created around each site with unavailable AOD to calculate the mean PM from nearby monitors. These values were then used in a final PM estimation model using generalized additive mixed models, allowing robust PM estimation.

The modelling of this study involved detailed land-use, meteorological, and boundary layer predictors, and included inverse probability weighting to correct for nonrandom missing AOD. Model performance was strong despite the challenging setting, with cross-validated R^2 values of 0.79 for PM_{10} and 0.72 for $PM_{2.5}$, representing a substantial improvement over the nearest previously published study using MODIS 10 km AOD to predict PM in Israel (Sorek-Hamer et al., 2015; $R^2 = 0.45$).

By exploiting MAIAC's fine spatial resolution and spectral retrieval advantages over bright surfaces, the study captured gradients across urban, rural, coastal, mountain, and desert regions, enabling reconstruction of long- and short-term population exposures with unprecedented coverage. Results were validated using a 10-fold cross-validation approach.

This study provided a benchmark for air pollution exposure assessment in complex environments and offered a valuable basis for both epidemiological studies and environmental policy applications. Moreover, it highlighted that PM satellite-derived models are sensitive to the spatial distribution of ground monitors, particularly in regions of Israel with very sparse coverage (the Negev, areas near Beersheva, and the Golan and Kinneret) where some locations have only one or no monitors allocated.

Estimating daily and intra-daily PM10 and PM2.5 in Israel using a spatio-temporal hybrid modeling approach (Shtein et al., 2018)

Once it was scientifically clear that PM models can be significantly improved by incorporating satellite data, new innovations emerged, focusing on enhancing both the spatial and temporal resolution of predictions. By leveraging the hourly availability of satellite AOD, as in Zhou et al. (2020), it became possible to develop intra-daily models characterizing hourly PM exposure. Combining with strategies to enhance spatial resolution using high-resolution predictors such as Landsat data, this study enabled the development of semi-hourly, ultra-precise PM models at 200 m resolution.

The study by Shtein et al. (2018), building on the work of Kloog et al. (2015) in Israel, exemplifies these advancements. They developed high-resolution daily and intra-daily PM_{10} and $PM_{2.5}$ models (1 km-200 m) using a multi-stage modeling framework. Their approach relied on MODIS–MAIAC AOD data, land-use, meteorological, and pollution predictors, as well as ground PM_{10} and $PM_{2.5}$ measurements (2005–2015). Satellite AOD was linked to

monitored PM using mixed-effects models with day-specific random effects, refined locally through support vector regression, and gap-filled to account for missing AOD data

The modelled framework of this study was conducted with the following stages:

- In Stage 1, they calibrated 1 km AOD with grid-level PM monitor observations using a linear mixed-effects model. Models were fitted to predict PM across three time windows—daily, intra-daily 10 AM–4 PM, and 8 AM–2 PM—corresponding to the hourly overpasses of the MODIS Terra and Aqua satellites. Once the 1 km predictions were generated, a local refinement stage improved the resolution to 200 m by incorporating high-detail spatial variables (e.g., NDVI, elevation, distance to roads, and percent land use), and training the residuals from the 1 km stage using support vector regression.
- In the second stage, 1 km predictions were generated using the calibrated model from Stage 1 at sites with available AOD.
- In Stage 3, for grid cells where AOD was missing, the relationship between the estimated PM and inverse-distance-weighted (IDW) interpolations of PM observations from monitoring stations was modeled using a mixed-effects approach.

To validate model performance, authors applied 10-fold cross-validation by randomly splitting the monitoring data into 10 groups. Cross-validated performance was strong, with R^2 values of 0.78–0.94 and RMSE of 10.36–56.09 μg/m³ for PM_{10}, and R^2 values of 0.58–0.94 and RMSE of 4.41–14.76 μg/m³ for $PM_{2.5}$. The fine spatial detail of the models revealed urban, industrial hotspots beyond those visible in daily averages. Moreover, the study showed that PM levels during satellite overpasses were often higher than 24-hour means, highlighting the added value of intra-daily modeling. In general, performance deteriorated on dust days due to extreme PM values due to dust-storms.

One important contribution of this study was the identification of the main sources of high PM exposure. The authors emphasized the challenges of predicting PM concentrations in a country like Israel, where, in addition to anthropogenic emissions, recurrent natural events such as dust storms sharply elevate particulate levels. During these events, daily average PM_{10} concentrations can exceed 2,000 μg/m³, with hourly concentrations reaching up to 5,000 μg/m³ (Vodonos et al., 2015), making accurate prediction extremely difficult. This explains the relatively high RMSE values for PM_{10} in their models. To avoid misinterpretation of

model robustness, the authors classified days as "Dust" or "Non-Dust" using the methodology of Vodonos et al. (2015), which defines dust storm days as those when PM_{10} levels exceed two standard deviations above the background mean. Based on this approach, they established a threshold of 71 μg/m³ to identify dust days.

By integrating satellite AOD data and ground measurements with advanced statistical approaches, this study substantially improved PM exposure assessment in the Middle Eastern region, characterized by complex topography and sparse environmental monitoring networks, providing a valuable tool for both epidemiological research and policy applications.

Estimating Daily $PM_{2.5}$ and PM_{10} over Italy Using an Ensemble Model (Shtein et al., 2020)

Alexandra Shtein and other researchers applied their expertise by exploring, for the first time in Italy, the potential of satellite-derived models to estimate PM concentrations. This study is especially relevant because PM_{10} levels in Italy often exceed the limits set by the European Union, and the country is characterized by diverse geo-climatic regions and mountain ranges such as the Alps and the Apennines.

This study integrates the predictions of different base learners; linear mixed-model, random forest, extreme gradient boosting, and the Italian CTM FARM (known as the flexible air quality regional model) into a single geographically weighted GAM to estimate daily $PM_{2.5}$ and PM_{10} concentrations. Rather than assigning fixed weights to each model, the GAM allows their contributions to vary across space and concentration levels, better reflecting Italy's complex geo-climatic conditions and diverse emission sources.

The main innovation of this study is the ensemble learning approach they used, which at that time incorporated RF and XGBoost as novel methods compared to the commonly used LMM. To train the GAM meta-learner, they used out-of-fold predictions from each base learner as the training set, ensuring that no data leakage occurred. From this study, we found that some models, such as the CTM FARM, were included in the ensembling process yet contributed little to the final model's performance. This suggests that a preliminary assessment is necessary before constructing an ensemble model, in order to evaluate the potential contribution of each individual model to the overall performance. This study also explored the imputation of $PM_{2.5}$ with PM_{10} to expand the coverage of in-situ measurements. Another

important highlight of this study was the implementation of an advanced cross-validation scheme to prevent spatial leakage and overly optimistic results, unlike conventional approaches that rely on standard k-fold cross-validation.

Their cross-validation procedure consisted of a 10-fold split of the monitoring database into 10 groups of stations. For each fold, 80% of the data was used for training, 10% for testing the base learners, and 10% for final validation. The final model made predictions only on this validation set, which was never used during the training of any base learner, allowing an unbiased assessment of performance. To prevent overfitting and overestimation of model performance, the spatial locations of monitors used for training did not overlap with those used for testing.

Final predictions were generated on a 1km resolution grid across Italy for 2013–2015, the first period with sufficient nationwide $PM_{2.5}$ monitoring data. Cross-validated results of the ensemble model showed clear improvements over any individual model, with R^2 values of 0.79–0.81 and RMSE of 5.3–6.6 $\mu g/m^3$ for $PM_{2.5}$, and R^2 of 0.73–0.76 and RMSE of 8.5–9.4 $\mu g/m^3$ for PM_{10}.

Gaussian Markov random fields improve ensemble predictions of daily 1 km PM2.5 and PM10 across France (Hough et al, 2021)

Ian Hough and colleagues developed a spatiotemporal model to predict daily $PM_{2.5}$ and PM_{10} at 1 km resolution across continental France from 2000 to 2019, marking the first long-term, high-resolution model of this kind in the country. Initially, the study addressed the sparsity of $PM_{2.5}$ monitors by imputing $PM_{2.5}$ concentrations at PM_{10} monitoring sites using a RF model. Gaps in satellite AOD were also imputed with another RF trained on modelled AOD from atmospheric reanalysis with 80 km spatial and 3-hourly temporal resolution.

The framework built consisted of the following stages:

- In stage 1, $PM_{2.5}$ monitors' values were imputed with PM_{10} at monitors using a RF model alleviating the sparsity of PM2.5 observations. Predictors included PM_{10} at nearby monitors, urban/rural/suburban classification, latitude, longitude, weekday, day of the year, and other relevant variables.
- In stage 2, gaps in MAIAC AOD were filled using monthly RF models trained on atmospheric reanalysis and spatial variables such as latitude, longitude, and calendar-

based dummy variables (day of year, weekday, date, etc.), ensuring complete spatiotemporal coverage.

- In stage 3, three base learners—linear mixed models (LMM), Gaussian Markov random fields (GMRF), and RF—were trained to predict daily PM, with GMRFs capturing smooth spatiotemporal variation and outperforming the other learners. The LMM and GMRF models were trained on the logarithm of PM, while the RF model was trained directly on PM. Predictor variables included reanalysis AOD from different spectral bands, gap-filled MODIS AOD, and other spatiotemporal variables such as population, elevation, latitude, longitude, and calendar-based dummy variables.
- In stage 4, predictions from the base learners were combined using a generalized additive model (GAM), which allowed the contribution of each model to vary depending on location and time, leveraging the strengths of each method across regions and periods.

To validate their final results, they used a rigorous multi-stage cross-validation scheme. In stage 1, a 5-fold CV with monitor blocking was used to impute $PM_{2.5}$ at PM_{10} monitors, ensuring that test monitors were never included in the training set. In stage 2, a 5-fold CV with spatiotemporal blocking was applied to fill gaps in MAIAC AOD, dividing the study area into 50 regions per day to prevent spatial leakage. In stage 3, the base learners (LMM, GMRF, RF) were evaluated using multi-stage monitor-blocked cross-validation, assigning each monitor to one of five folds, resulting in 10 combinations in total; models were trained on three folds and tested on the two held-out folds to evaluate predictions at unmonitored locations. In stage 4, the GAM ensemble was trained on cross-validated predictions from held-out monitors and evaluated on monitors excluded from both the base learners and the ensemble training, preventing any possible data leakage, and producing realistic performance metrics.

This approach produced highly accurate predictions ($PM_{2.5}$: $R^2 = 0.76$, MAE = 2.72 μg/m³; PM_{10}: $R^2 = 0.71$, MAE = 4.26 μg/m³), and highlighted GMRFs as the dominant model in the ensemble. This study demonstrated that multi-stage ensembles incorporating GMRFs can predict daily PM concentrations more accurately than both mixed models and RF, producing a national, high-resolution PM dataset that is valuable for epidemiological studies and policy-making.

High-Resolution Estimation of Daily PM2.5 Levels in the Contiguous US Using Bi-LSTM with Attention (Wang et al., 2025)

With the rise of artificial intelligence, neural networks have become one of the most common models used to solve problems across different fields. This trend is also true for exposure modeling. However, no single structure has yet emerged as the clear leader in terms of performance.

The study by Wang et al. (2025) applied a deep learning approach, specifically a type of recurrent neural network (RNN), to estimate daily $PM_{2.5}$ at 1 km resolution across the contiguous United States from 2005 to 2021, focusing on improving prediction accuracy during high-concentration days and wildfire events. The study addressed gaps in existing datasets and improved prior models that failed to capture temporal dynamics and extreme pollution events. Unlike other PM models, this study incorporated predictors such as precipitation, water vapor pressure, air temperature (Ta), shortwave radiation, day length, wind direction, wind speed, wildfire smoke density, inverse-distance-weighted $PM_{2.5}$, encoded date variables, and spatial variables such as elevation, latitude, and longitude. MODIS AOD data was also included; however, CAMS AOD data was notably absent.

This study introduced several innovations and highlights key challenges in predicting PM, particularly the difficulty of modeling high concentrations caused by random environmental events. The authors applied feature engineering to create a new predictor based on the nearest-neighbor monitors to boost their models. Besides, they explicitly demonstrated the benefit of including the variable called "wildfire smoke density", which estimates the thickness of smoke in the atmosphere from wildfires and is expressed qualitatively (e.g., light, medium, heavy). This variable does not provide full spatial coverage, it is supplied by NOAA's Hazard Mapping System (HMS) and is only available for North America, Hawaii, and the Caribbean. It is captured from satellite instruments such as GOES-R/ABI and JPSS/VIIRS.

Their modeling approach does not follow a fixed set of stages, unlike previous studies, for producing $PM_{2.5}$ estimates. Instead, it uses an architecture made up of several layers that process the data. Specifically, it consists of three stacked Bi-LSTM layers (a type of recurrent neural network) designed to capture temporal dependencies, along with an attention mechanism (an approach originally developed in natural language processing that assigns different weights to inputs, allowing the model to focus on most important parts of the

training data) to identify cyclical temporal patterns. The main limitations of the study include temporal gaps in the HMS smoke density data and sparse or irregular ground-station data, which reduce the effectiveness of the KNN-IDW feature they developed. The authors also noted that the current Bi-LSTM approach is less effective at capturing local spatial patterns. They suggested that hybrid CNN–LSTM models or graph neural networks could improve spatial modeling and noted that the current attention mechanism offers limited interpretability, arguing that further research could enhance its explainability.

To end, this study used a 10-fold random cross-validation on 90% of the data to evaluate model robustness, alongside a 10% held-out test set for final performance assessment. The cross-validation results were averaged across folds, while the best-performing model from cross-validation was applied to the independent test set. For the KNN feature they derived, they commented that they had excluded the target location's own measurement during training to prevent data leakage.

XIS-PM2.5: A daily spatiotemporal machine-learning model for PM2.5 in the contiguous United States (Just et al., 2025)

Similar to what Just et al. (2025) achieved for Ta in the U.S, they developed a model to estimate daily $PM_{2.5}$ concentrations across the contiguous territory (CONUS) from 2003 to 2023. The model combined MODIS AOD with a parsimonious set of meteorological, land-use, and topographic predictors, enabling high-resolution estimates at arbitrary locations. Performance was evaluated using a 10-fold spatial cross-validation, which accounted for the spatial distribution of monitoring sites and applied weighted metrics to fairly represent sparsely monitored areas. Over the study period, the model achieved a weighted mean absolute error (MAE) of 2.09 μg/m³, and mean absolute deviation from the median (MAD = 4.15 μg/m³). When compared with other leading U.S. products, the model outperformed the EPA's Fused Air Quality Surface with a 17% reduction in MAE.

To train the model, the study used a combination of spatial, temporal, and environmental predictors: latitude and longitude, day of the year, an engineered feature interpolating nearby monitor $PM_{2.5}$ values within a 500 km radius of site weighted by the cube of the distance, MODIS AOD, daily surface $PM_{2.5}$ from MERRA-2 at 50 km resolution, the Enhanced Vegetation Index from Aqua and Terra satellites, distance to primary roads, land cover,

population density, elevation, and six meteorological variables from NLDAS-2 (temperature, specific humidity, air pressure, zonal and meridional wind speeds, and precipitation).

The authors describe their models as a computationally tractable workflow: unlike recent models that rely on many predictors and complex algorithms, their approach is easy to implement and reproduce, using variables that are straightforward to update. Although $PM_{2.5}$ is the ultimate target, the model is trained on the difference between $PM_{2.5}$ and the IDW-interpolated $PM_{2.5}$ feature, which is also included as a predictor; for predictions, this feature is added back to obtain the final estimates. They evaluate model performance using MAE and variability with MAD, and the model's optimization function was the log-cosh error.

Their cross-validation scheme was robust, using 10-fold spatial cross-validation with equally sized folds. Importantly, for each sample, the IDW feature was computed while excluding monitors of the validation fold, ensuring that information from the target did not leak into the training data. This follows the standard rule in machine learning that any feature derived from the target (like neighboring PM values) must be computed using only the training data (Mucci, 2024). However, for final predictions or deployment, the features should be recalculated using all available data, allowing the model to learn from the maximum amount of information.

A last comment, one of their key innovations compared to current approaches is that the model is not tied to a fixed grid, allowing predictions at arbitrary locations, which is especially useful for epidemiological studies that require estimates at specific points. The model leverages a derived predictor based on ground-based PM measurements from nearby monitors, calculated using a buffer around the prediction site and weighted by an inverse distance weighting (IDW) function.

Key Takeaways from PM Modeling with AOD

Collectively, these previous studies demonstrated that integrating satellite-derived AOD into PM estimation represents a significant and progressive advancement for the field. Key strengths included consistent global coverage, high spatial resolution (≤1 km), and temporal components (resolution ≤1 day), all of which are particularly valuable for epidemiological applications. Building on this foundation, current studies now model PM using satellite data alongside complementary variables, often employing machine-learning approaches. Looking

ahead, further improvements in satellite technology, coupled with the use of more advanced algorithms such as neural networks, are expected to enable PM estimations with minimal error, thereby reducing bias and increasing consistency in health-related studies in complex geographical terrains, such as those of Israel and France. Table 2 highlight innovations and limitations of these studies.

Summary of reviewed studies in this section

Table 2. Key innovations, contributions, and limitations of the studies linking PM with satellite data.

Study	Innovations-contributions	Limitations
1. Kloog et al. (2011)	• One of the first models using remote sensing data to predict $PM_{2.5}$. • Even though the spatial resolution of AOD was 10km, they managed to improve the resolution to 100m.	• None.
2. Emili et al. (2011)	• Compared satellite-derived models of that epoch with interpolations. • Applied a vertical correction method that can be used in actual models. • Support the idea of including IDW features to the model.	• Models of the epoch struggled to capture spatiotemporal variability, and satellite data had low resolution.
3. Kloog et al. (2015)	• Improved resolution of PM models locally to 1km and daily.	• Limited monitor coverage required spatial cross-validation to assess model performance. • Given Israel's small size, further improving of spatial resolution

Study	Innovations-contributions	Limitations
		would have benefited epidemiological studies.
4. Shtein et al. (2018)	• Through their detailed scientific analysis, they identified that the deterioration in model performance was driven by events such as dust storms. • Previous studies did not apply a dust-storm classification, which helps isolate PM predictions attributable solely to anthropogenic sources. • Innovation with predictions at 200m and hourly.	• Limited monitor coverage in some regions, leading to reduced model generalization. • Cross-validation was not spatial, potentially leading to overoptimistic performance estimates in regions with sparse monitoring. • Used models are old compared to the ones available at the epoch.
5. Shtein et al. (2020)	• First Italian study predicting $PM_{2.5}$. • Imputed $PM_{2.5}$ with PM_{10} to increase the coverage of monitors. • Their cross-validation scheme better simulates the true performance of the models compared to traditional CV approaches. • Demonstrated that their models outperform the models used by local institutions.	• The study ensembles models in multiple stages but did not evaluate whether such assembling will provide reasonable benefits.
6. (Hough et al., 2021)	• Explored new modelling alternatives like the GMRF. • Used ensemble learning, which started to become common in the field.	• None.

Study	Innovations-contributions	Limitations
	• Used spatial cross-validation, which better simulates the performance of models. • Their estimates were highly accurate.	
7. (Wang et al., 2025)	• Used deep learning (neural networks) to predict PM, following recent trends in exposure modeling. • Used different predictors than usual, including wildfire smoke density. • Created new features and ensured no data leakage during cross-validation. • Capture temporal patterns in different ways with the attention mechanism in neural nets.	• Did not use CAMS AOD data. • Limited interpretability of the model (e.g., attention mechanism and feature importance).
8. (Just et al., 2025)	• Their new way to train the models (PM vs IDW-PM) is extremely creative. • Applied feature Engineering. • Predictions at arbitrary locations will benefit epidemiological studies. • Easy to reproduce for the variables used. • The costumed performance metric they used (weighing the density of monitors). • Applied spatial cross-validation.	• High dependence on a dense network of monitors, because it relies a lot on the interpolation feature. • Does not explain exactly how it interpolates the 1km predictions, making them become arbitrary. • Do not exactly explain if the models they compared with used the same cross-validation approach, and the same metrics.

2. 4 Conclusions from the Literature Review on Satellite-Based Models for Ta and PM

2. 4. 1 General Conclusion

The literature review clearly shows that the modeling of environmental factors such as Ta and PM has advanced significantly with the inclusion of satellite-derived data. This integration has enhanced both the spatial and temporal resolution of models, which is particularly important for epidemiological studies, as higher resolution reduces bias and allows for more precise hypothesis testing with narrower confidence intervals. Nonetheless, previous studies highlight that these investigations are relevant in every country, given the growing concern about climate change and the adverse effects of environmental exposures worldwide, especially in populations that are more sensitive to even minimal climatic changes, such as in France.

Beyond epidemiological outcomes, these models also provide valuable insights into long-term changes in temperature and pollution, enabling a detailed understanding of temporal trends necessary for establishing new policies or public health warnings. In countries such as Switzerland, these studies have promoted the expansion of meteorological networks, demonstrating the importance of deploying monitors densely even in rural areas to accurately assess exposure risks, for example, to better understand urban heat effects (Flückiger et al., 2022). In France, these models have improved understanding of temperature trends and supported the development of heatwave warning systems over extended periods (Laaidi et al., 2012). In Israel, the models have facilitated the subdivision of PM exposure studies, distinguishing areas where anthropogenic sources dominate from those affected primarily by natural or hazardous events, such as dust storms (Shtein et al., 2020).

Although several complex modeling approaches, such as tree-based algorithms, were already known in the early 2000s, it was only around 20 years later that studies began fully exploiting their benefits, facilitated by the increasing availability of open-source software supporting research. At the same time, newer models can incorporate ever-larger datasets, which enhances their robustness, spatial coverage, and predictive accuracy. In this new era, artificial intelligence demonstrates remarkable capabilities, and it is expected that in the coming years

these technologies will become commonplace in environmental studies. Currently, exposure modeling still relies primarily on tree-based models, because they offer both accuracy and interpretability, a key limitation of neural networks. However, the future focus is on deep learning models. Several studies are now exploring ways to "open the black box" of deep learning, attempting to interpret the relationships between covariates and the learned weights. This suggests that in the near future, it will be possible to develop highly precise models while also understanding their outputs. Furthermore, AI experts such as Andrew Ng have noted that as datasets continue to grow, the learning curves of neural networks continue to improve, indicating that neural networks are likely to become the dominant approach in future exposure modeling approaches.

The current phase of Ta and PM exposure modeling involves the gradual incorporation of more advanced machine-learning strategies, such as feature engineering and ensemble learning. It is surprising that many models developed in the second decade of the century did not utilize these approaches, even though they are common, and often considered essential steps in standard industrial machine-learning workflows. This omission is understandable, as implementing such strategies can be time-consuming and difficult to evaluate.

In this context, the main weaknesses we identified in current exposure modeling of environmental variables such as Ta and PM, as highlighted by our literature review, include a lack of reproducibility, limited use of feature engineering, insufficient feature selection—even though feature importance is often considered in most models—and inconsistencies in the applied validation approach.

2. 4. 2 Explanation of Drawbacks

1. The lack of reproducibility means that, even if part of the data is already available, models must be rerun for new years and data recollected. Additionally, some models are so complex that reproducing them is difficult. Recent studies are attempting to tackle this issue (Just et al., 2025) by identifying a set of spatiotemporal variables that are easy to collect and capture much of the variability of the target variables, thereby improving the speed and reproducibility of the models.
2. Lack of feature engineering is minimal in the majority of studies, while generally implementing basic-to-medium transformation such as encoding categorical variables or sinusoidizing temporal variables. However, many additional strategies could be

applied, for example, transformation of target variables using a log scale (e.g., in Hough et al., 2021), or combining interpolation of the targets in training data to capture spatial relationships (Just et al., 2024; Just et al., 2025; Wang et al., 2025; Emili et al., 2011). Other types of feature engineering commonly used in industry are rarely seen in exposure modeling, such as parameter tuning of derived features (e.g., the number of monitors to use within a 100 km or 500 km buffesr from the site of interest), categorical target encoding, or techniques often applied in neural networks that can also be used with tabular data, such as convolutions, poolings, embeddings, and principal component analysis (transforming predictors into a lower-dimensional vector space).

3. Even though feature importance is calculated in most studies using SHAP plots or top-importance plots, sensitivity analyses are rarely applied. For instance, models are not retrained after excluding the least important predictors or using a permutation feature selection approach. Implementing such strategies could help identify the most relevant variables and avoid collecting those that are noise, potentially improving the final model performance. We suspect this is often not done because studies involve a substantial amount of work, which can become even more challenging and hard if performing additional experiments.
4. Models rely heavily on dense networks of monitors to measure pollutants; in their absence, performance is drastically affected. In countries like France, such infrastructure is well-developed, and models achieve strong results. However, in countries like Israel, where PM monitors are much sparser, performance is considerably poorer. This issue is particularly relevant when considering the transfer of these modeling approaches to countries without similar infrastructure, highlighting the need for strategies that depend less on meteorological networks. One promising opportunity for improvement is transfer learning, which could enable models to adapt to different contexts based on topography without relying on local monitors. If this is not feasible, at a minimum, more localized validation strategies should be employed to accurately evaluate model performance. These could range from traditional random cross-validation to leave-one-out monitor approaches (spatial blocking), potentially combined with weighting by monitor density.

2. 4. 3 Types of Goals in exposure modelling

From this literature review, we understood how exposure modeling evolved over time. Early models (2000s) focused on identifying approaches that could accurately replicate the behavior of the target variables, using methods such as linear regression, geographically weighted regression (GWR), linear mixed effects models (LME), and generalized additive models (GAM). Once this foundation was established, newer strategies emerged (2010s–2020s) aimed at generating more precise predictions, using tree-based models and ensemble learning to compare models and exploit their strengths under different conditions. During this period, better modeling was also anticipated from improvements driven by satellite instruments updates, achieving better spatial and temporal resolutions each time. From the 2020s to present, models have not only focused on precise estimates but also on evaluating the uncertainty and reliability of their predictions through cross-validation. Advanced validation schemes are common, to simulate scenarios such as predictions at further areas or in zones with low allocated density (e.g., leave-one-monitor-out cross-validation, spatial blocking by regions).

This evolution has helped us characterize the specific goals addressed in environmental modeling studies, enabling future research to target one or more of these goals directly and be more precise in the types of contributions sought. Additionally, it provides a solid foundation for the objectives pursued in this thesis.

1. Performance: Studies aiming to improve model performance (i.e., achieve more accurate predictions) compared with the existing literature. It typically implies robust cross-validation schemes.
2. Modeling alternatives: Although performance may vary, some approaches can achieve similar results as the ones of previous leading studies while using simpler methodologies. This also includes exploring techniques that have not been widely applied, such as the use of GMRF and RNN models (Hough et al., 2021; Wang et al., 2025).
3. Validation (simulation of predictions): Various cross-validation approaches have been developed, including hybrid evaluation methods, to better assess model performance. Examples include spatial blocking, leave-one-monitor-out spatial blocking, and stratified cross-validation based on environmental conditions or attributes other than distance, such as elevation, land use, or population density (Shtein et al., 2020; Jin et al., 2022; Flückiger et al., 2022).

4. Improvement of spatial and/or temporal resolution: Models that are designed to enhance spatial and temporal detail, tailored to meet the specific needs of epidemiological studies (Shtein et al., 2018; Zhou et al., 2018; Hough et al., 2020)

2. 4. 4 In-Depth Analysis of Ensemble Learning

We decided to include this dedicated section in our conclusions from the literature review to discuss ensemble learning, as it is increasingly integrated into current exposure modeling. Although it was noticeable that previous studies addressed it well, we consider that some important implications are not considered and might be omitted while applying it. Thus, we try to explain briefly our findings and propose a simple, evidence-based good practice to follow before constructing ensembles, based on insights from previous studies

Implications

1. When constructing ensemble models, labeled data must be divided into at least training and test sets (e.g., 80%-20%). For the training portion of the ensemble model, out-of-fold (OOF) predictions from the base learners are generated. An OOF prediction is the estimate for a label made by a model that has not seen that label during training, ensuring it represents truly unseen data. These predictions are obtained using cross-validation: in each iteration, the model is trained on all folds except one held-out fold, and predictions are made for the held-out fold. Repeating this process for all folds ensures that every observation in the training set has an OOF prediction. These OOF predictions are then used as the input features for the ensemble model, as described in most studies (Aghbalou et al., 2022). Because OOF are "unseen", this approach then avoids data leakage. However, generating these predictions requires multiple rounds of training and prediction, making cross-validation both mandatory and computationally intensive.
2. Second, after evaluating ensemble performance on the test set, all base learners and the ensemble should be retrained on the entire dataset, and the corresponding predictions regenerated. This ensures that the final ensemble benefits from all available information (Ng, 2011). Notably, this final step was often omitted in the

studies we reviewed within exposure modeling (Ta-PM), presumably because it is viewed as obvious or an standard procedure.

These disclosed implications mean that training ensembles and evaluating their performance with cross-validation can lead to a large number of model fits.

For example, with a 5-fold cross-validation scheme, the number of runs required would be:

- 5× the number of folds for generating OOF predictions.
- 1 meta-learner trained on the OOF predictions.
- The number of base models for final predictions (retraining base models with all available data).
- 1 meta-learner for the final ensemble predictions (retrained using the OOF predictions from all base learners on the full data).

As an example, with 5 base models and K-fold-CV=5, this results in 5 × 5 + 1 + 5 + 1 = 32 models, which is clearly computationally demanding. This likely explains why this final step, retraining all models on the full dataset to leverage all available information, is often not reported in published studies and we suspect frequently not executed.

Equation 6 shows the total number of model fits in a real-life situation using ensemble learning. On the other hand, Equation 7 exhibits the total number of fits required if only the selected model obtained through cross-validation is used for producing the final estimates (# of model fits decreases).

For instance, with 5 base models and K-fold-CV=5, the total number of fits becomes: 5 × 5 + 1 = 26. Even in this simplified but nor real scenario, the computational burden remains substantial.

$$Total\ model\ fits = (F * B) + B + 2 \qquad \text{Eq. 6.}$$

$$Total\ model\ fits = (F * B) + 1 \qquad \text{Eq. 7.}$$

Where F is the number of folds in cross-validation, B is the number of base learners used.

Having outlined these important implications of ensemble learning, it is important to note that, based on our literature review, several studies implemented ensemble methods at various

stages (Shtein et al., 2020; Zhou et al., 2020; Hough et al., 2021) without first assessing whether the ensemble approach would provide meaningful improvements. Previous research supports the notion that simply stacking outputs from multiple models does not automatically guarantee more reliable predictions (Breiman, 1994; Sándor et al., 2023). These studies emphasize the importance of analyzing the residuals of base learners beforehand to determine whether ensemble outputs will reduce uncertainty or merely propagate correlated errors.

Correlation Analysis of Model Residuals Prior to Ensemble Learning

Simple statistical methods, such as correlation analysis, can be applied prior to ensembling phases (Santos Júnior et al., 2023; Wang et al., 2024). This involves calculating residuals for each base learner on validation data and constructing a correlation matrix. Using clustering or similar approaches, pairs of highly correlated models can be identified, and only one model from each pair should be selected for the final ensemble. This ensures that the ensemble combines models with complementary error patterns, maximizing predictive benefit. If correlations remain high among the selected models, the potential improvement from the ensemble will be limited. Although it is not possible to quantify in advance the exact performance gain of an ensemble model, this procedure at least provides statistical justification for including or excluding specific models in the ensemble.

2. 4. 5 In-Depth Analysis of Cross-validation

The literature review also provided an opportunity to delve into the role of cross-validation. Cross-validation helps reduce uncertainty regarding a model's ability to generalize, quantify its predictive accuracy, and detect potential overfitting. Notably, in many studies that assess performance through cross-validation, results on the training sets are often omitted; however, reporting training performance is generally considered a valuable step for detecting overfitting, regardless of the chosen validation strategy.

For the case of our reviewed literature, cross-validation is widely applied in the different studies; however, these results alone cannot be used to determine whether one model truly outperforms another, as some of them suggest. Since the final models are trained on the full data (as assumed), which is the same for all models being compared, the only valid way to compare performance is by applying an identical cross-validation approach consistently across models, a limitation we perceived, for instance, in Just et al. (2025).

At the same time, it should be emphasized that cross-validation is principally used to assess the prediction error (as a measure of risk), and the choice of a cross-validation methodology can lead to different error estimations (Mushagalusa et al., 2024; Frontiers, 2025; Bates et al., 2021).

The choice of cross-validation method influences the performance metrics obtained, and the reviewed studies provided insights into the reasons behind these differences:

1. Random K-fold cross-validation: This approach can underestimate prediction error because spatial locations in the training set may also appear in the test set, creating substantial data leakage. As a result, performance metrics may appear better than in scenarios that more closely mimic reality, such as when predicting in unknown and distant locations, as in Kloog et al. (2015).
2. Leave-one-monitor-out: This approach sets each monitor as a hold-out test, one at a time. It avoids overfitting and data leakage. The main limitation is that it requires multiple training folds, one for each monitor. When monitors are densely distributed, the metrics provide more realistic error estimates; however, in sparsely monitored areas, bias can occur. A better approach is to weight errors by monitor density, giving higher weight to rural or sparser zones, as implemented in Just et al. (2025).

3. Spatial block cross-validation: The study region is divided into spatial blocks, with each block assigned to a fold, and predictions are made on unseen areas. This approach is particularly useful for assessing performance in new regions. However, the resulting metrics tend to be pessimistic, often overestimating the prediction error.
4. Station-level random cross-validation: In this approach, folds are created by randomly selecting a percentage of monitoring stations, with one fold held out for validation and not used in training. Because this method does not explicitly account for spatial separation, some stations in the training set may be spatially correlated with stations in the test set. Consequently, predictive performance may appear slightly better than it would using a spatial block CV, as in Shtein et al. (2020).

Indeed, the optimal cross-validation strategy depends on the specific goals of the modeling task:

- For predictions at specific locations, leave-one-monitor-out (LOMO) cross-validation is effective.
- For evaluating how well the model generalizes across regions, spatial block CV is preferable.
- For an unbiased assessment of overall performance, using an independent validation set is essential (a held-out set not used in training).

Andrew Ng, an expert data scientist and founder of DeepLearning.AI, supports the idea that comparing the performance of training and test sets can provide insights into model potential issues (Ng, 2017). This is particularly important, as none of the reviewed papers reported training set results. Ng emphasizes that the goal in machine learning is to achieve both low bias and low variance. According to Ng (2017), poor performance on the training set indicates high bias, while a large gap between training and test set performance indicates high variance. Good performance on the training set suggests low bias, and good generalization to the test set suggests low variance. Based on his explanation, 4 main scenarios (considering error as the metric) can illustrate the trade-off between bias and variance. To illustrate this concept, he represented each scenario with arrows aimed at a target center (see Figure 2):

1. High bias, low variance: Arrows are clustered together but miss the target center (The model underfits; it needs more complexity).
2. High bias, high variance: Arrows are widely spread and far from the target center (The model underfits and is unstable; it needs more complexity and/or more data).
3. Low bias, high variance: Arrows are spread out but centered around the target (the model needs more data or regularization; it overfits).
4. Low bias, low variance: Arrows are tightly clustered near the target center (the model performs well and is accepted).

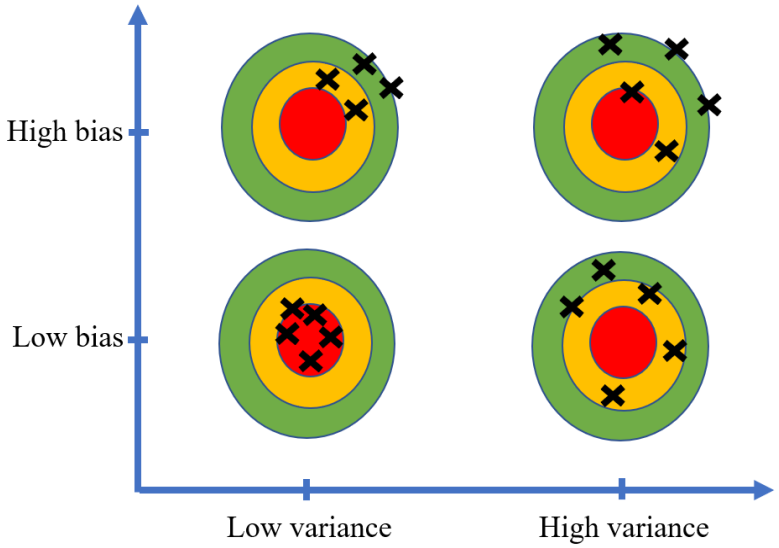


Fig. 2. Explanation of the bias-variance trade-off in Machine Learning (Ng, 2017).

Considering all factors involved in comparing different modeling approaches, competitions like Kaggle evaluate models based on predictions for specific labels rather than reported metrics. These predictions are withheld from the training data, ensuring a fair assessment of each model's true performance and providing a basis to judge whether one model outperforms another (Kaggle, n.d.).

Chapter 3: Methodology

3.1 Overview of Modelling Methodology

In this thesis, we model Ta and PM concentrations using a multistage framework inspired by the approaches reviewed in the previous chapters. First, we assemble a set of spatial and spatiotemporal predictors, including the variables of interest, for all years covered in each case study scenario. Data is retrieved from various publicly accessible datasets and institutional sources (e.g., LSA SAF, ECMWF reanalysis, CAMS reanalysis, NASA TIRS, the Israel Meteorological Service IMS, Météo-France, etc.). Where required, formal data requests are submitted, although most of the information is publicly accessible through APIs.

After data collection, all datasets are preprocessed to ensure consistency. This involves harmonizing coordinate reference systems (CRS) for satellite images, rasterizing data, and averaging variables when needed to match the study's temporal and spatial resolutions (≤1 km, ≤daily). Next, we build data tables that combine the predictors (independent variables) with the target variables (dependent variables) by merging the processed datasets.

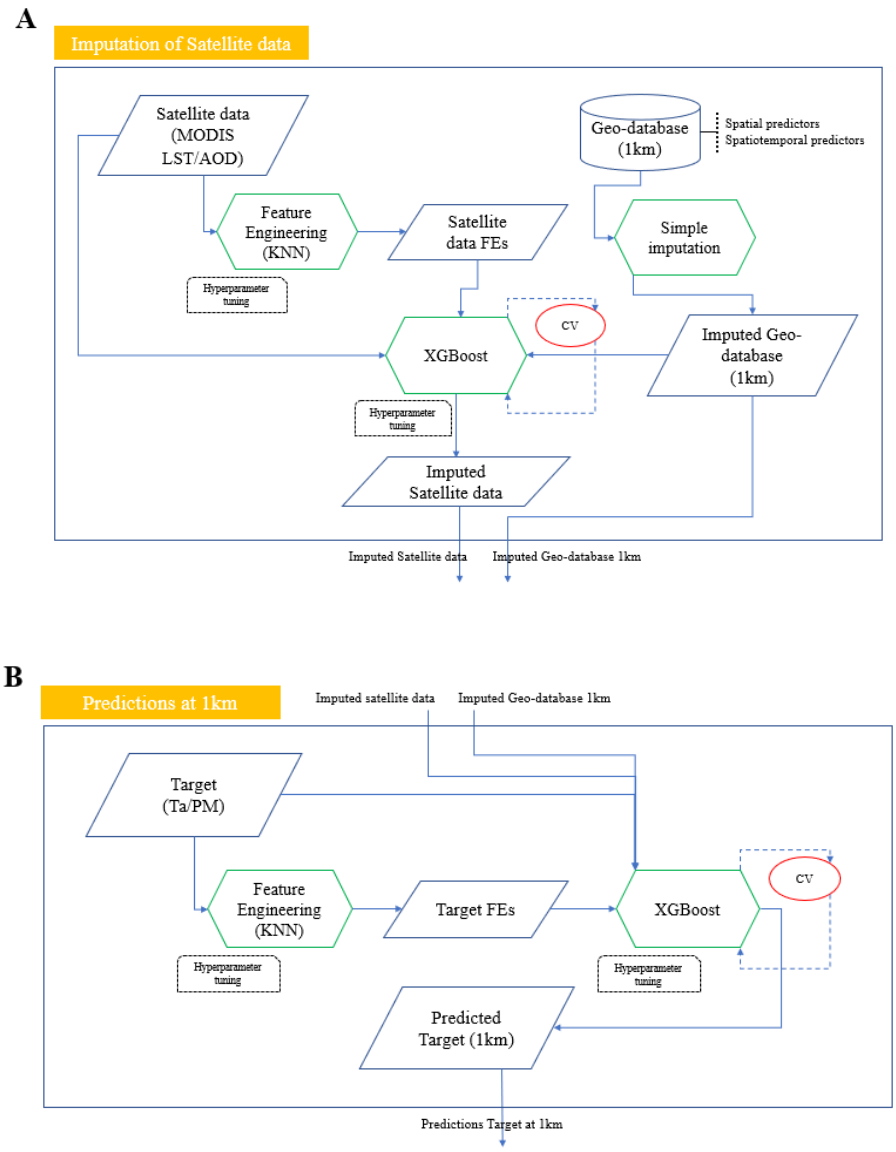
Where gaps appear in predictors (e.g., MODIS LST, MODIS AOD, SEVIRI LST) due to cloud contamination, surface brightness issues, snow cover, or other factors, we apply imputation strategies ranging from simple techniques to model-based approaches, depending on the context. Once the full set of predictors and labeled data is assembled, we proceed with model development, with slight variations depending on the target variable. Multiple experiments are carried out to identify the models with the best predictive performance.

Model validation is performed iteratively during this process using different cross-validation schemes. After validation, the final models are retrained and applied to predict across the complete spatial and temporal domains of interest. Finally, a local modeling stage is introduced, relying on a dedicated set of high-resolution predictors and techniques designed to address residual errors.

To further enhance performance, we incorporate feature engineering (from basic transformations to more advanced methods) into selected parts of the imputation, prediction, and validation stages, with the most sophisticated features generated via the KNN model, creating new spatiotemporal predictors that preserve relevant spatial relationships. These new

predictors are referred to as FEs by us, a concept that will be explained in more detail later in the chapter.

By extending the traditional frameworks used in exposure modelling, we improve the predictive performance of the models through the integration of feature engineering. Our innovation addresses the "type 1" and "type 2" objectives outlined in Chapter 2. Specifically, we develop a strategy focused entirely on improving the performance of previous similar literature (type 1 objective) while also partially adopting techniques that, although less common in environmental exposure studies, are widely applied in other industrial fields that use machine learning and data science (type 2 objective). Figure 3 presents the general workflow we used to model Ta and PM:



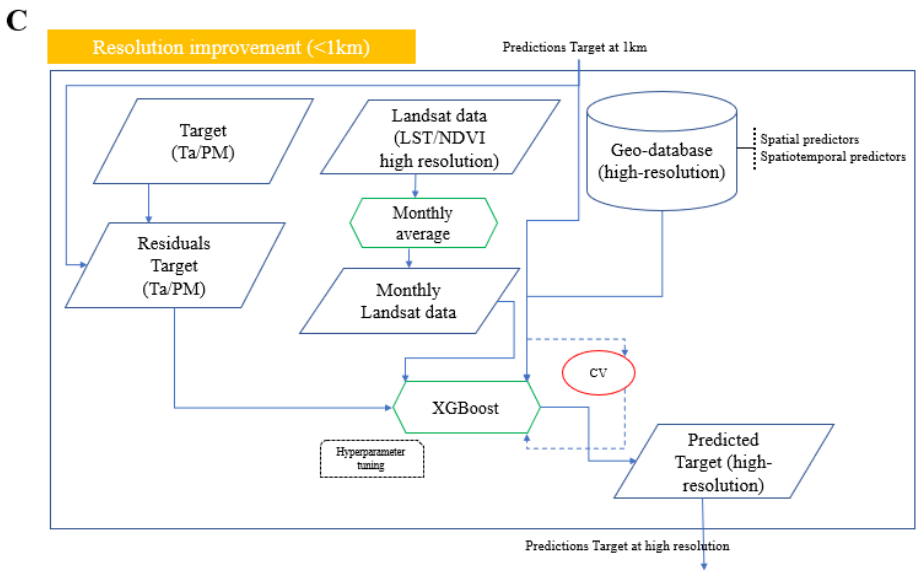


Fig. 3. General diagram workflow for modelling Ta and PM.

A. Modelling part that imputes satellite data.

B. Modelling part that predicts Ta/PM at 1km.

C. Modelling part that predicts Ta/PM at higher resolution (e.g., 200 m).

3.2 Data

The models developed in this study rely on a diverse set of spatial and spatiotemporal predictors to estimate Ta and PM. Key details of the implicated variables, such as their names, associated papers within this thesis, and their relevance, are provided here (Table 3), while more comprehensive descriptions are presented in the following chapter. These predictors include coarse regional variables that provide contextual information at medium-to-high spatial and temporal resolution (25 m–1 km; yearly–daily), other high-resolution local predictors (30–200 m; montly) are designed to capture fine-scale spatial variability. For some predictors, feature engineering was applied to enhance their contribution to predictive performance; these cases are indicated accordingly. Detailed information on each dataset, including sources, temporal coverage, retrieval, and preprocessing steps, is provided in Chapter 4 within the sections corresponding to each paper. This overview is intended to summarize the predictors used and highlight their relevance within the general modeling framework.

Table 3. Brief description of the collected variables used in this dissertation**.**

Predictor / Variable	Spatial Resolution	Temporal Resolution	Relevance / Purpose	Applied Feature Engineering	Associated Article
Latitude / Longitude	200m-1km	N/A	Captures spatial location and trends in the models.	No	Article 1, 2, 3
Elevation	25m-1km	N/A	Provides topographic context.	No	Article 1, 2, 3
Calendar dates	N/A	Hourly-Daily	Encodes temporal information such as day of year, month, week of year, weekday, etc.	Basic	Article 1, 2, 3
Population density	200m-1km	Static (yearly)	Represents the distribution of people.	No	Article 2, 3
Land cover	100m	N/A	Describes surface types (urban, vegetation, water, etc.) that influence local temperature and pollutant dispersion.	No	Article 2
French climatic regions	N/A (categorical)	N/A	Classifies areas into established climatic zones, providing contextual information for modeling Ta in France.	No	Article 2
Distance to nearest city	200m-1km	N/A	Indicates proximity to urban areas, which can influence local air quality patterns.	No	Article 3
Distance to highways	200m-1km	N/A	Indicates proximity to major highways, which can affect pollutant concentration patterns.	No	Article 3

Predictor / Variable	Spatial Resolution	Temporal Resolution	Relevance / Purpose	Applied Feature Engineering	Associated Article
Distance to main roads	200m-1km	N/A	Measures proximity to major roads, which may affect pollutant concentration patterns.	No	Article 3
Length of highways	200m-1km	N/A	Quantifies the density of highway infrastructure in a given area, serving as a proxy for traffic-related emissions.	No	Article 3
Length of all roads	200m-1km	N/A	Represents overall road network density, which influences emissions and land-use characteristics.	No	Article 3
Israel's natural areas	N/A (categorical)	N/A	Identifies protected or natural zones that may influence local microclimate and pollutant dispersion.	No	Article 3
Land use	200m-1km	N/A	Describes functional characteristics of the surface (e.g., residential, industrial, agricultural), which can influence pollutant dispersion.	No	Article 3
Israel region	N/A (categorical)	N/A	Classifies areas within Israel into administrative regions, providing contextual information for	No	Article 3

Predictor / Variable	Spatial Resolution	Temporal Resolution	Relevance / Purpose	Applied Feature Engineering	Associated Article
			modeling Ta and PM		
MODIS NDVI (Normalized Difference Vegetation Index)	1km	2 weeks	Represents vegetation cover and greenness, which can influence surface temperature and pollutant dispersion.	No	Article 2, 3
MODIS EVI (Enhanced Vegetation Index)	1km	2 weeks	Captures vegetation density with improved sensitivity in high-biomass areas and reduced atmospheric effects, complementing NDVI.		Article 3
MODIS LST (Land Surface Temperature)	1km	Daily	Measures surface temperature, which is directly related to Ta and can influence local PM dynamics.	Advanced (using the KNN model).	Article 1, 2
2 meters air temperature	10km	Hourly	Provides near-surface air temperature, directly related to Ta and influencing PM behavior	No	Article 1, 2
Skin Temperature	10km	Hourly	Represents the temperature of the Earth's surface "skin," complementing air temperature to improve Ta and PM modeling.	No	Article 1, 2
SEVIRI LST	3km	15min	Provides high-frequency surface	No	Article 1, 2

Predictor / Variable	Spatial Resolution	Temporal Resolution	Relevance / Purpose	Applied Feature Engineering	Associated Article
			temperature measurements, capturing diurnal variations that can improve Ta and PM prediction.		
Landsat LST	100m	16days	Provides high-resolution surface temperature or reflectance data, useful for capturing fine-scale spatial variability	No	Article 1, 2
Landsat NDVI	100m	16days	Measures vegetation greenness at high spatial resolution, complementing coarser NDVI products for fine-scale variability.	No	Article 2, 3
MODIS AOD 470 nm	1km	Daily	Provides information on aerosol concentration in the atmosphere from satellites, which is directly related to PM levels.	Advanced (using the KNN model).	Article 3
CAMS AOD 469 nm	80km	Daily	Provides modeled aerosol concentration at 469 nm, sensitive primarily to ultrafine and fine particles (PM0.1–PM2.5), complementing satellite AOD for PM estimation.	Basic (combined with other spectral bands to calculate indexes and ratios).	Article 3
CAMS AOD 550 nm	80km	Daily	Provides modeled aerosol	Basic (combined	Article 3

Predictor / Variable	Spatial Resolution	Temporal Resolution	Relevance / Purpose	Applied Feature Engineering	Associated Article
			concentration at 550 nm, sensitive primarily to fine particles (PM2.5), complementing satellite AOD for PM estimation.	with other spectral bands to calculate indexes and ratios).	
CAMS AOD 670 nm	80km	Daily	Provides modeled aerosol concentration at 670 nm, sensitive mainly to coarse particles and dust, complementing other spectral AOD bands for PM estimation.	Basic (combined with other spectral bands to calculate indexes and ratios).	Article 3
CAMS AOD 865 nm	80km	Daily	Provides modeled aerosol concentration at 865 nm, sensitive mainly to large particles, such as dust or smoke, complementing other spectral AOD bands for PM estimation.	Basic (combined with other spectral bands to calculate indexes and ratios).	Article 3
CAMS AOD 1240 nm	80km	Daily	sensitive primarily to very coarse particles and biomass burning aerosols, complementing other spectral AOD bands for PM estimation.	Basic (combined with other spectral bands to calculate indexes and ratios).	Article 3
CAMS DUST AOD	80km	Daily	Provides modeled aerosol optical depth specifically for mineral dust	No	Article 3

Predictor / Variable	Spatial Resolution	Temporal Resolution	Relevance / Purpose	Applied Feature Engineering	Associated Article
			particles (like sand), helping to distinguish dust from other aerosol types in PM estimation.		
Wind Speed	50km	Daily	Represents near-surface wind velocity, influencing pollutant dispersion and local temperature patterns	No	Article 3
Wind direction	50km	Daily	Indicates the prevailing wind direction, which affects the transport and dispersion of pollutants and influences local temperature patterns	No	Article 3
Air temperature	N/A (point data)	Hourly		Advanced (using the KNN model).	Article 1, 2
Particle matter	N/A (point data)	hourly		Advanced (using the KNN model).	Article 3

3.3 Data preprocessing

Because our models can cover long temporal horizons (e.g., 15 years of more), data collection and preprocessing required significant effort, taking anywhere from one month to several months of work for some predictors. This is a careful and meticulous process that must be performed systematically and validated at multiple stages to ensure data remains coherent and homogeneous.

Data Collection and Structuring

For environmental modeling applications, which often require diverse types of predictors, multiple data sources must be accessed. Each dataset comes with its own structural

characteristics, reflecting the origin, resolution, and intended use. Globally, the datasets used in this study can be categorized into the main formats described below, along with their specific advantages for environmental modeling and data processing.

HDF (Hierarchical Data Format): MODIS satellite products are typically provided in HDF format. These files contain multiple images or layers corresponding to specific tiles that cover different regions of the globe. HDF is particularly advantageous because it allows the storage of large, multidimensional arrays and metadata within a single file, facilitating the organization of complex datasets. Moreover, each HDF file includes the native coordinate reference system (CRS), often sinusoidal for MODIS, which simplifies georeferencing during subsequent analysis.

NetCDF (Network Common Data Form): Many datasets from the Copernicus Climate Data Store (e.g., reanalysis skin temperature) and CAMS (e.g., AOD 469 nm) are provided in NetCDF format. NetCDF is ideal for medium-resolution gridded data (10–80 km) with high temporal frequency (hourly), covering global scales. Its self-describing structure and compatibility with multiple programming languages (Python, R, MATLAB) make it suitable for storing large multidimensional arrays with associated metadata, supporting efficient access and analysis of climate datasets.

NC4 (NetCDF-4): High-resolution satellite datasets such as SEVIRI LST from EUMETSAT are typically distributed in the NC4 format, a NetCDF-4 variant. NC4 supports advanced data compression and chunking, which is critical for handling high-resolution (e.g., 3 km) and high-frequency (hourly) imagery. This format facilitates faster reading and writing while preserving the detailed temporal and spatial characteristics essential for precise environmental monitoring.

JSON (JavaScript Object Notation): Point-based meteorological observations, such as those collected from monitoring stations in Israel, are often distributed in JSON format. JSON is lightweight and structured, making it suitable for web-based API requests. Its human-readable and nested structure allows efficient data parsing and manipulation, reducing download time and enabling flexible integration into a data processing workflows.

Parquet: Observational datasets from meteorological stations in France, managed by local public institutions, are commonly provided in Parquet format. Parquet is a columnar storage

format derived from Apache Hadoop and Java structures, which provides significant advantages for processing large tables. Its columnar nature allows for efficient compression and query performance, particularly when working with specific variables or subsets of the data, reducing computational costs and storage requirements.

Once a clear understanding of the different data formats for collected datasets was established, the process of retrieving and transforming the data into a unified structure (data frame) can be summarized in the following steps.

1. Datasets are retrieved from different sources, each with its own native format (HDF, NETCDF, JSON, etc.). Once collected, each dataset is transformed into a consistent tabular structure, such as CSV or Snappy Parquet, to facilitate subsequent processing, aggregation, and joining.
2. Collected datasets are projected, if necessary, to ensure all datasets share the same coordinate reference system (CRS). It is important to retain the geometric columns of each dataset, including latitude-longitude coordinate pairs and the original CRS information.
3. Datasets are organized on a grid. Usually, the grid of one reference dataset is used; if needed, a grid can be created manually while respecting the required spatial resolution. Each grid contains at least a unique ID, a geometric column with coordinate pairs, and a temporal column (e.g., date).
4. For datasets with higher spatial or temporal resolution than required by the models, values are aggregated to match the target resolution. Spatial aggregation involves averaging values within buffer areas, while temporal aggregation can include computing daily averages, minima, or maxima. Point data are assigned to the nearest grid cell.
5. For datasets with lower resolution than required, values are mapped to the corresponding higher-resolution grid cells. Temporal downscaling follows a similar approach, matching each date to the relevant week, month, or year.
6. Units are verified and standardized across all variables of the same type. All variables are then joined into a single comprehensive table based on the predefined grid, containing the ID, geometric columns (latitude-longitude coordinates), a time column, predictors, and dependent variables. Variables with coordinates that do not exactly match the predefined grid are assigned to the nearest grid coordinates.

Factors That May Lead to Misleading Results

We experienced several factors that can increase the risk of producing inaccurate results in the models if not carefully addressed during the preprocessing and structuring. One such factor is the presence of different time zones in the retrieved data, as development was conducted across different continents, resulting in variations in recorded hours (e.g., retrieving Israel data from France).

Another important consideration is the reprojection of data between coordinate reference systems (CRS). Converting from geographic systems using degrees to projected systems using distance units requires careful handling, as multiple CRS exist and transforming between them is akin to changing data types. It is essential to define a target CRS and to understand the effects of transforming the native CRS to the chosen system. All spatial data must be strictly aligned to maintain consistency.

The data retrieval process itself can also introduce risks. Geospatial data are typically accessed through APIs in various programming languages, often requiring custom code to retrieve the data automatically and exhaustively. This code must be carefully verified and regularly updated, as data repositories are constantly evolving with updates, new datasets, changes in naming conventions, and modifications in resolution.

Finally, miscalibration of the instruments used to measure target variables (e.g., air temperature and particulate matter) can result in inconsistent or unreliable values. It is therefore important to examine key descriptive statistics to identify and filter out anomalies or incoherent measurements.

3.4 Feature Engineering

In the following paragraphs, we present the main innovation presented by this thesis, which is grounded in feature engineering. We focused on this aspect, as most of the literature does not explicitly address exposure modelling using data-driven feature methods for improving model performance. Although feature engineering is rarely mentioned explicitly, many studies have already applied it to enhance their models (Shtein et al., 2020; Hough et al., 2021; Wang et al., 2025; Just et al., 2025).

In this section, we introduce our main motivation and describe what feature engineering entails. Even though Chapter 4 will present each thesis article with clear and straightforward definitions and explanations of how feature engineering is used, we take advantage of this section to provide a deeper explanation, illustrating how we developed our new features from basics to advanced. We also present a set of best practices for implementing feature engineering with various techniques, and how to integrate them into the model validation process. Finally, we compare our proposed approach with similar methods, such inverse distance weighting interpolation and universal kriging, highlighting its advantages, weaknesses, and remaining limitations. We do not explicitly show results in this section, but theoretically, we summarize everything to provide a comprehensive overview.

What's feature engineering?

Feature engineering encompasses a suite of techniques designed to transform raw data into informative and meaningful predictors. This process may include a range of tasks such as applying mathematical transformations, filtering out noise, combining or deriving variables, selecting the most relevant inputs, and performing other forms of tailored data manipulation, all with the goal of producing predictors that are well suited to statistical or machine-learning models.

An engineered feature, such as a new attribute, explanatory variable, or column, might for instance be created by transforming the day of the year (DoY) using sine and cosine functions. This approach encodes both the variable and its cyclical nature (1-2-3 … 365-1-2-3) preserving the seasonal ordering. Representing the day of the year as a simple binary variable would not allow the model to recognize this cyclical pattern. Other simple real-world examples of feature engineering include the word count in a document (Suhaidi et al., 2021) or an ordinal number representing the age of a property (Aksoezen et al., 2015), illustrating the broad applicability of feature engineering across multiple fields.

Everything that introduces new information from the original dataset can be considered an engineered feature. Good features capture relevant information, eliminate noise, reduce irrelevant variation, help the model detect patterns, avoid overfitting, and sometimes improve interpretability. Feature engineering encompasses several steps, including addressing missing or inconsistent data, encoding categorical variables, scaling or normalizing numerical values,

and building features based on domain expertise. It may also involve aggregating or transforming existing features and selecting the subset most relevant to the predictive task.

Feature engineering is not limited to predictors; it can also involve modifying dependent variables. This may include applying transformations, such as log or Box-Cox, to skewed numerical variables like AOD or PM to approximate normality and reduce the influence of outliers (Cheddad, 2020), or by redefining a target, for example, as the difference between PM and the values of neighboring monitors (Just et al., 2025).

A quick Recap into Beginnings

The idea of handcrafting variables to improve modeling is much older than the point at which the formal term "feature engineering" began to be used. In classical statistics, particularly in linear regression, transformations such as logarithms, square roots, power transformations (e.g., Box-Cox), ranking, or normalization have been used for decades to satisfy key model assumptions such as linearity, homoscedasticity, and normality of residuals (Montgomery, Peck & Vining, 2021). Other direct application was in Exploratory Data Analysis. By the 1970s, Tukey emphasized the importance of transforming data and exploring relationships through variable transformations to better understand underlying structures (Tukey, 1977). In financial forecasting, using dummy variables to represent calendar dates and capture daily, weekly, monthly, and yearly patterns has long been a standard practice (Gujarati & Porter, 2009).

More recently, with the rise of large-scale machine learning, large and varied datasets, heterogeneous data types, and more complex model algorithms, the explicit practice of feature engineering became more central. Around the 2000s, the concept broadened within the data science community and literature, particularly as researchers recognized that model performance often depends more on how data are represented than on small tweaks to algorithm parameters (Domingos, 2012; Bengio et al., 2013). Empirical studies have compared how different kinds of engineered features affect different models: for example, Heaton (2017) conducted an empirical analysis showing how models like decision trees, random forests, neural networks, and gradient boosting respond differently to engineered features, and how certain features may or may not need to be provided explicitly depending on what the model is able to "learn" internally (Heaton, 2017). In industrial settings, frameworks like SAFE (Scalable Automatic Feature Engineering) have been developed to

automate feature engineering processes. SAFE is a framework that automatically processes features to improve efficiency, scalability, and interpretability in large-scale tasks such as real-time fraud detection (Shi et al., 2020).

Finally, recent advances such as feature stores, automated feature generation (commonly known as wide learning), and representation learning, which allows models to automatically discover and learn new features, are increasingly supplementing or even replacing manual feature engineering (Banerjee et al., 2016; Shi et al., 2020). Despite these automation efforts, domain knowledge remains crucial; many of the most significant improvements in successful applications still come from carefully selecting, designing, and handcrafting features.

In the past, scientists relied mostly on formal mathematical approaches to derive solutions. Today, data scientists can exhaustively explore parameters and representations, an evolution that has contributed to the growing popularity of the term "feature engineering" since the early 2010s. In particular, modern neural networks demonstrate this shift: operations like convolutions automatically generate relevant features, enabling models to achieve near-perfect accuracy in many classification problems (Zhao et al., 2024).

Driving Factors for the Innovation

The primary goal of this thesis was pre-defined as investigating innovative methods to enhance the modeling of Ta and PM. Beyond this scientific objective, our work was additionally motivated by personal experiences. In particular, while being exposed to an educational assignment based on a Kaggle competition, I gained firsthand experience with the widespread use of feature engineering in predictive modeling. Kaggle is an online platform where data scientists compete to develop the most accurate models for public, private, or industrial challenges posted by organizations. It can be seen as an open form of consultancy, where problems are addressed using analytical reasoning, computational approaches, or a combination of both (kaggle).

Several years ago, Kaggle launched the Home Credit Default Risk competition, with the objective of developing predictive models to assess the likelihood of loan applicants defaulting. Participants were provided with a comprehensive dataset containing features related to applicants' demographics and credit history, including repayment behavior on previous loans. This competition attracted thousands of data scientists and machine learning

practitioners from around the world (Kaggle, 2018). Detailed information about the competition is available on Kaggle: https://www.kaggle.com/competitions/home-credit-default-risk (last accessed February 2022).

After the competition concluded, its solution was disclosed, revealing the systematic strategies used by the winner team. The awarded solution explained that success lay in the application of extensive feature engineering, particularly leveraging a KNN-based approach. Using the dataset of clients with detailed credit records, including both approved and rejected applications, they created a new feature for each applicant based on the 500 most similar clients (Kaggle, 2018). This feature was not part of the original dataset, and its careful engineering was pivotal in achieving the highest model performance.

After a thorough review of the environmental exposure literature, it became evident that this type of logic is both applicable and extendable to regression problems. However, for tasks such as those addressed in environmental exposure, we found that this kind of approach has been applied very little, and when it has, only superficially. We posed a series of simple, focused questions that allowed us to identify concrete and non-negligible benefits. Our reasoning can be likened to navigating a landscape: for air temperature (Ta), a point at intermediate elevation would have an intermediate temperature, while for particulate matter, distant points require neighbors defined by demographic or anthropogenic characteristics rather than geographic coordinates. We applied this logic in designing our feature-engineered variables.

The second motivation for emphasizing feature engineering arose from the inherent structure of environmental exposure models and their similarity to the convolutional neural network model (CNNs). Exposure models are typically designed to estimate environmental emissions across space, focusing on reconstructing past patterns rather than predicting future values. Consequently, information from neighboring locations is readily available through measurements from fixed monitors, allowing the models to fully exploit spatial relationships (patterns). This structure is analogous to CNNs, where convolutions leverage spatial relationships between neighboring data points to improve predictions (LeCun et al., 2015). Inspired by this analogy, we designed feature-engineered variables that explicitly capture spatial dependencies in a similar manner.

3.4.1 From Basic to Advanced Feature Engineering

We define basic feature engineering as straightforward transformations that require minimal development effort, often involving direct or simple calculations, yet still capable of producing substantial improvements in model performance. Examples include computing averages, applying binarization, performing clustering, merging classes, deriving class-wise averages, transforming target variables, and applying operations between variables. These techniques can be used on both predictor and target variables.

3.4.1.1 Basic Feature Engineering

The following examples illustrate some of the features developed during the course of each thesis article. Some of these features had already been identified in the reviewed studies, demonstrating their established relevance. We also discuss the specific benefits of each feature, emphasizing how they contribute to improved model performance and the capture of key patterns in the data. Equations 8 to 15 provide a summary of these features.

Time-Based Features

These features include the binarization of calendar date columns, such as weekday, day of year, week of year, month, and year, as well as sine and cosine transformations to capture cyclical patterns (e.g., month binarization with sine-cosine encoding). Without these transformations, important temporal information may be lost.

Month binarization (one-hot encoding):

For month $m \in \{1, 2, \ldots, 12\}$, create 12 dummy variables:

$$M_i = \begin{cases} 1, & if\ m = i \\ 0, & otherwise \end{cases} \quad for\ i = 1, 2, \ldots, 12 \qquad \text{Eq. 8}$$

Benefit: Separates each month explicitly, allowing the model to learn month-specific effects.

Sine-cosine transformation to capture cyclical patterns:

$$Month_{sin} = sin\left(2\pi * \frac{m}{12}\right),\ Month_{cos} = cos\left(2\pi * \frac{m}{12}\right) \qquad \text{Eq. 9}$$

Benefit: Captures the cyclical nature of temporal variables, for example, showing that December and January are adjacent. Encoding each month using sine and cosine functions

preserves this periodic structure while substantially reducing dimensionality compared with one-hot encoding (2 columns instead of 12 for months).

Categorical Target Encoding

Categorical target encoding converts categorical predictors into numerical representations based on the target variable. This approach captures the relationship between each category and the outcome while avoiding the high-dimensionality that can result from one-hot encoding. For example, a categorical variable representing regions (such as cities, natural areas, or similar) can be encoded by computing the mean of the target variable within each category.

For a categorical predictor region $\in \{R_1, R_2, \ldots, R_R\}$, the encoded value for each observation is:

$$region_{encoded} = \frac{\sum_{j \in R_i} y_j}{n_i} \; \forall \; i = 1,2,\ldots,R \quad \text{Eq. 10}$$

Where y_j is the target value for observation j in region R_i, and n_i is the total number of observations in category R_i.

Benefit: This encoding captures the relationship between each region and the target, while keeping the feature dimension stable, unlike one-hot encoding which would create R separate columns.

Spatial Spline Features

Spatial splines allow regression models to capture smooth spatial trends in the target variable by transforming geographic coordinates into a set of smooth basis functions. This approach captures local spatial variation across the study area in a compact way, avoiding the need to manually integrate numerous location-based predictors in the model (e.g., interaction terms, multiple coordinates, or polynomial expansions), which can quickly lead to high-dimensional and complex datasets. For example, given geographic coordinates (latitude x, longitude y), a two-dimensional spline can be fitted to represent the spatial structure of the target variable. Spatial splines have been successfully applied in environmental modeling, as demonstrated by Kloog et al. (2018).

For a regression model $Y = f(x, y)$, the application of a spatial spline transforms the model into:

$$Y = \sum_{i=1}^{k} \beta_i B_i(x, y) + \varepsilon \quad \text{Eq. 11}$$

where $B_i(x,y)$ are the spline basis functions derived from the coordinates, β_i are the corresponding coefficients, k is the number of basis functions, Y is the target, and ε is the residual error.

Benefit*:* Using raw geographic coordinates alone may fail to capture complex spatial patterns. Spatial splines, by contrast, summarize smooth spatial trends and account for spatial autocorrelation, the tendency for nearby locations to have similar values, thereby improving predictive performance. Including the original coordinates (longitude x, latitude y) in addition to the spline is generally unnecessary, as the spline already encodes the location effects (Naimi et al., 2011).

Aggregates by temporal units (e.g., Week of Year or Day of Year) and by spatio-temporal pairs (latitude–longitude-time)

To improve model performance (e.g., when imputing MODIS AOD), we incorporated aggregated features that summarize the target variable across both temporal and spatial dimensions. These features help the model capture seasonal cycles and location-specific trends, which are often missed by raw observations alone.

Aggregates per Day-of-Year (DoY):

For each calendar day (DoY = 1–365), we compute summary statistics of the target variable across all locations. Typical statistics include:

- Mean**:** average value of the target on that day.
- Minimum / Maximum**:** extreme values of the target on that day.
- Standard deviation**:** variability of the target on that day.

- Example: For DoY = 150 (late May), the mean temperature across all point-sites represents the expected seasonal condition for that day, while the min/max capture potential extremes.

- Benefit: These aggregates allow the model to account for general seasonal patterns and expected variations without explicitly modeling time trends.

Aggregates per (Latitude–Longitude, Week of year) pairs:

- Similar to aggregates per DoY, for each spatial location (lat, lon) and each week of year, we calculate aggregated statistics of the target variable. Typical summaries include:
    - Mean: expected target value at a location in the corresponding week of the year.
    - Minimum / Maximum**:** local target extremes of at a location in the corresponding week of the year.
    - Standard deviation**:** local variability at a location in the corresponding week of the year.
- Example: For a grid cell representing a site in Paris (CRS Sinusoidal) where x= 174000 m, y = 5437000 m, and week of year = 5 (beginning Feb), the mean temperature in the 5th week of the year at the point (174000, 5437000) captures location-specific seasonal trends.
- Benefit**:** These features incorporate historical spatio-temporal patterns, helping the model distinguish between general seasonal effects (e.g., from DoY-day of year aggregates) and local deviations (e.g., from location-specific WoY-week of year aggregates).

Overall Benefit*:*

- Both types of aggregates act as informative priors that guide the model toward realistic seasonal and spatial patterns.
- They complement spatial splines and other predictors, enhancing predictive accuracy without substantially increasing model complexity. Although the model may exhibit some bias toward these aggregated values, when combined with other predictors, these engineered features effectively contribute to improving overall predictive performance.

Logarithmic Transformation of Skewed Target Variables

For target variables with strongly skewed distributions, such as PM concentrations or Aerosol Optical Depth (AOD), applying a logarithmic transformation can enhance model performance by stabilizing variance and moderating the impact of extreme observations, thereby allowing the model to more effectively capture underlying patterns in the data (Patel et al., 2025). A key practical consideration is that logarithmic transformations are only defined for strictly positive values. For datasets containing zeros, it is common to add a small constant $c > 0$ prior to transformation to ensure all values are positive:

Logarithmic transformation:

For a target variable Y>0, define:

$$Y_{log} = \log(\mathrm{Y}) \quad \text{Eq. } 12$$

$$Y_{log} = \log(\mathrm{Y} + c) \quad \text{Eq. } 13$$

Benefit: logarithmic transformations stabilize variance and reduce skewness, allowing the model to capture patterns in the bulk of the data rather than being dominated by extreme values. Predictions can then be converted back to the original scale using the exponential function:

$$\hat{Y} = \exp\left(\hat{Y}_{log}\right) \quad \text{Eq. } 14$$

$$\hat{Y} = \exp\left(\hat{Y}_{log}\right) - c \quad \text{Eq. } 15$$

The five strategies described above represent only a subset of the many possible feature engineering approaches that can be applied to generate new predictors from the original variables. These engineered features can subsequently enhance model performance and provide a deeper understanding of the dependent variable's behavior. In the context of this thesis, substantial improvements in the final evaluation metrics were observed, particularly for satellite data imputation, where even minimal feature engineering yielded measurable benefits. In a later section, we will discuss the implications of incorporating these variables into the models, highlighting the importance of handling them carefully to prevent data leakage and to ensure rigorous validation.

3.4.1.2 Advanced Feature Engineering: The KNN Model

The nearest-neighbors model proved to be a powerful method for grouping target observations and creating aggregated features (Kaggle, 2018). This is particularly valuable in exposure modeling, where predictors in their initial raw form are rasters or subsequently rasterized, with spatial relationships that are explicit and cannot be ignored (neighboring pixels are dependent). A key concern, however, is that these spatial relationships can be lost during the model's training process, due to the multiple stages and inherent randomness involved in model fitting and data handling. For this reason, creating features that preserve these relationships is particularly valuable and can be accomplished through feature engineering.

In the case of Ta, it was already reasonable to expect potential improvements before applying this approach. Our aim was therefore to evaluate to what extent predictions could be enhanced. Previous studies have shown that Ta can be adjusted by the combined effects of other spatial predictors, such as elevation, which motivated us to introduce greater flexibility to include such adjustments in our modeling (Yao & Zhang, 2013; Kloog et al., 2017).

For particulate matter (PM), the application was more uncertain, as, unlike Ta, PM is influenced by a wider range of complex factors that vary substantially with geography and human activity. Taking these considerations into account, a key aspect of the proposed innovation is its flexibility in incorporating spatial dimensions: it allows the selection of the spatial variables to include when creating features based on nearest-neighbor values, rather than being restricted solely to the geographic coordinate pairs. This flexibility addresses both the need to capture known spatial effects, such as latitude, longitude, and elevation, and the capacity to incorporate additional, potentially influential spatial factors for more complex targets like PM.

This approach differs fundamentally from traditional interpolation methods, allowing for an exhaustive search of optimal nearest-neighbor configurations tailored to a specific problem. Studies such as Just et al. (2025) and Wang et al. (2025) have already explored similar modeling based on feature engineering and KNN. However, our innovation goes further by applying this technique not only in calibration (prediction stages) but also in imputation. Additionally, we performed intensive hyperparameter tuning of k (the number of nearest neighbors), offering significant potential for improvement over conventional methods.

The following section explains how we derived our new set of engineered predictors, referred to as FEs. The logic for calculating these features using the KNN model is consistent across all stages. However, for imputation processes, the neighborhood of search is not limited to sites occupied by Ta or PM monitors. In this context, the first application involves identifying neighbors for each site from all available points to impute missing satellite data. The second application consists of finding the nearest neighboring monitors for each site to generate the new predictors used to feed the models. The following steps detail the calculation of these features for both scenarios. Equations 16 and 17 group both complete formulations.

FEs for Imputation

1. Represent a territory (such as France, Israel, or another country) with a grid of points, each site is denoted by a point $p \in \{1, 2, \ldots, P\}$ where P is the total number of points used to represent the territory.
2. Each point p is characterized by a set of spatial predictors ($I_1, I_2, I_3, \ldots, I_l$), such as I_i = longitude of p, I_{i+1} = latitude of p, I_{i+2} = elevation of p, or any other relevant attributes. The distance between two points is then defined as a function of the selected spatial predictors $f(I_1, I_2, I_3 \ldots, I_l)$ allowing any number of attributes to be incorporated.
3. For instance, if the desired spatial predictors for calculating distance are I_1 = longitude and I_2 =latitude, and the chosen function is the Euclidean distance, then the distance from a point $i \in \{1, 2, \ldots, P\}$ to a point $j \in \{1, 2, \ldots, P\}$ can be defined as:

$$d_{ij} = \begin{cases} \sqrt{\left(I_{1i} - I_{1j}\right)^2 + \left(I_{2i} - I_{2j}\right)^2}, & if\ i \neq j \\ 0, & if\ i = j \end{cases}$$

4. Then, we calculate the distance matrix D using Euclidean distance as a function for every pair of points, and given l predictors following the same logic:

$$X = \begin{bmatrix} I_{11} & I_{21} & \cdots & I_{l1} \\ I_{21} & I_{22} & \cdots & I_{l2} \\ \vdots & \vdots & \ddots & \vdots \\ I_{1P} & I_{2P} & \cdots & I_{lP} \end{bmatrix}, \text{ where } I_{ai} \text{ is the a-th spatial predictor of point } i.$$

- We let X_a denote the column vector of size P containing the values of spatial predictor a for all points. 1_P is a column vector of ones.

$$X_a = \begin{bmatrix} I_{a1} \\ I_{a2} \\ \vdots \\ I_{aP} \end{bmatrix}, 1_P = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix};\ then\ X_a \otimes {1_P}^T = \begin{bmatrix} I_{a1} & I_{a1} & \cdots & I_{a1} \\ I_{a2} & I_{a2} & \cdots & I_{a2} \\ \vdots & \vdots & \ddots & \vdots \\ I_{aP} & I_{aP} & \cdots & I_{aP} \end{bmatrix}$$

$$1_P \otimes X_a = \begin{bmatrix} I_{a1} & I_{a2} & \cdots & I_{aP} \\ I_{a1} & I_{a2} & \cdots & I_{aP} \\ \vdots & \vdots & \ddots & \vdots \\ I_{a1} & I_{a2} & \cdots & I_{aP} \end{bmatrix}$$

$$\Delta X_a = X_a \otimes {1_P}^T - 1_P \otimes X_a = \begin{bmatrix} I_{a1} - I_{a1} & I_{a1} - I_{a2} & \cdots & I_{a1} - I_{aP} \\ I_{a2} - I_{a1} & I_{a2} - I_{a2} & \cdots & I_{a2} - I_{aP} \\ \vdots & \vdots & \ddots & \vdots \\ I_{aP} - I_{a1} & I_{aP} - I_{a2} & \cdots & I_{aP} - I_{aP} \end{bmatrix}$$

$$\Delta X_a \odot \Delta X_a = \begin{bmatrix} (I_{a1} - I_{a1})^2 & (I_{a1} - I_{a2})^2 & \cdots & (I_{a1} - I_{aP})^2 \\ (I_{a2} - I_{a1})^2 & (I_{a1} - I_{a2})^2 & \cdots & (I_{a1} - I_{aP})^2 \\ \vdots & \vdots & \ddots & \vdots \\ (I_{aP} - I_{a1})^2 & (I_{aP} - I_{a2})^2 & \cdots & (I_{aP} - I_{aP})^2 \end{bmatrix}$$

- Repeating this process for all predictors $a = 1,\ldots, l$ and summing the squared differences (element-wise) across all predictors allows for the computation of the full distance matrix D (e.g., using the Euclidean distance).

$$D = \sqrt{(\Delta X_1 \odot \Delta X_1) + (\Delta X_2 \odot \Delta X_2) + \cdots + (\Delta X_l \odot \Delta X_l)}$$

$$D = \sqrt{\sum_{a=1}^{l} (\Delta X_a \odot \Delta X_a)}$$

- Where each element of the matrix D, is a d_{ij} value.
- D is a matrix of dimensions $P \times P$.

5. Self-exclusion for points.

- Distance of a point p to itself is zero, so the diagonal of D is zero. We set these values to infinity so a point won't be chosen as its own neighbor when using the argsort function.

$$D_{p,p} = \infty$$

- This ensures a location does not select itself as a neighbor.

6. Given the distance matrix *D*, we use the *argsort* function to identify the *k* nearest neighbors for each point *p*. For every point, this results in a set of *k* neighbors, which we store in a new matrix, denoted as N.

   - For each point p = 1, 2, …, P, use the argsort function on the p-th row of *D*.

$$sorted_indices_p = argsort(D_{p,:})$$

   - Select the first *k* neighbors:

$$N_p = sorted_indices_p[1:k];\ \forall p \in \{1,2,3,\dots,P\}$$

   - Form the neighbor matrix N:

$$N = \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_P \end{bmatrix}$$

   - N is a matrix of dimensions $P \times k$.

7. Calculation of FEs for a dependent variable *Y* (LST or AOD). FEs are the average of the *k* nearest neighbors' values at time *t*. Only non-missing neighbor values are used, and if all values of neighbors are missing for a point p, then the average at that site is NA.

   - $Y^t = [Y_1^t, Y_2^t, \dots, Y_P^t]^T \in \mathrm{R}^{\mathrm{P}}$ be the target variable at time *t*.
   - $N \in \mathrm{R}^{\mathrm{P}\times\mathrm{k}}$ be the neighbor matrix, fixed across times (t=1,2,3…). Each row N_p contains the indices of the *k* nearest neighbors of point *p*.
   - A binary indicator should be defined to distinguish whether *Y* is missing or observed:

$$Z_{pj}^t = \begin{cases} 1, & if\ Y_{N_{(p,j)}}^t is\ observed\ (not\ NA). \\ 0, & if\ Y_{N_{(p,j)}}^t is\ missing\ (NA). \end{cases}$$

   - Select neighbor targets:

$$Y_{neighbors,p}^t = \left[Y_{N_{(p,1)}}^t, Y_{N_{(p,2)}}^t, \dots, Y_{N_{(p,k)}}^t\right];\ \ \forall p \in \{1,2,3,\dots,P\}$$

- From the selected neighbor target of a point p, we calculate how many are non-missing:

$$S_p^t = \sum_{j=1}^{k} Z_{pj}^t$$

- Compute the average value of Y over the neighbors of each point p at time t:

$$\bar{Y}_p^t = \begin{cases} \left(\frac{1}{S_p^t}\right) * \sum_{\substack{j=1 \\ Z_{pj}^t \neq 0}}^{k} Z_{pj}^t * Y_{N(p,j)}^t \ , if\ S_p^t > 0 \\ NA, \quad if\ S_p^t = 0 \end{cases}$$

- Compute for all points the averages of their neighbors.

$$\bar{Y}^t = \begin{bmatrix} \bar{Y}_1^t \\ \bar{Y}_2^t \\ \vdots \\ \bar{Y}_P^t \end{bmatrix} = \begin{bmatrix} \frac{\sum_{\substack{j=1 \\ Z_{pj}^t \neq 0}}^{k} Z_{1j}^t * Y_{N(1,j)}^t}{S_1^t} \\ \frac{\sum_{\substack{j=1 \\ Z_{pj}^t \neq 0}}^{k} Z_{2j}^t * Y_{N(2,j)}^t}{S_2^t} \\ \vdots \\ \frac{\sum_{\substack{j=1 \\ Z_{pj}^t \neq 0}}^{k} Z_{Pj}^t * Y_{N(P,j)}^t}{S_P^t} \end{bmatrix}, \quad with\ \bar{Y}_p^t = NA\ if\ S_p^t = 0$$

- Where $\bar{Y}^t \in \mathrm{R}^{P \times k}$ is the vector of average non-missing neighbor targets for all points at time t.

Note: The neighbor matrix N is constant across all times; only $\bar{Y}^t$ changes.

- We apply the same neighbor-averaging procedure for each time step t in the modeled period (e.g., a year). The resulting FE variable is denoted as:

$$FE_{attributes\ I,k} = [\bar{Y}^1, \bar{Y}^2, \dots, \bar{Y}^T] \in \mathrm{R}^{\mathrm{P} \times \mathrm{T}}$$

- Example: If the spatial predictors used to define the features are latitude and longitude, and the number of neighbors is k = 5 then:

$$FE_{(latitude,longitude),5} = [\bar{Y}^1, \bar{Y}^2, \dots, \bar{Y}^T] \in \mathrm{R}^{\mathrm{P} \times \mathrm{T}} \qquad \text{(Eq. 16)}$$

However, for simplicity we just called the new feature as FE in latitude and longitude coordinates.

FEs for prediction

Now we consider the case where feature generation is applied at the prediction stage. The overall logic remains the same, with three important modifications: (1) the distance matrix is subset only with columns corresponding to the locations of monitors. (2) for points that are themselves monitors, their distance to themselves is adjusted so that they are not included as neighbors when computing the average, and (3) we integrate IDW to give more emphasis to nearby stations. The following steps outline these modifications:

1. Distances from points to monitors.
    - Monitors are located at S={s_1,…,s_m}, with $m \leq P$.
    - Each point and site of monitor has l spatial predictors I_1,…,I_l.
    - The following matrixes are useful:

$$X = \begin{bmatrix} I_{11} & I_{21} & \cdots & I_{l1} \\ I_{12} & I_{22} & \cdots & I_{l2} \\ \vdots & \vdots & \ddots & \vdots \\ I_{1P} & I_{2P} & \cdots & I_{lP} \end{bmatrix} \in R^{\boldsymbol{P}\times\boldsymbol{l}}, \qquad X^s = \begin{bmatrix} I_{1,S1} & \cdots & I_{l,S1} \\ \vdots & \ddots & \vdots \\ I_{1,Sm} & \cdots & I_{l,Sm} \end{bmatrix}$$

    - X has dimensions (P, P), and X^s has dimensions (m, l).

2. Distances matrix from points to monitors
    - Compute distances between every point and each monitor:

$$\Delta X_a^S = X_a \otimes 1_m{}^T - 1_P \otimes (X_a^S)^T$$

$$D_{p,s} = \sqrt{\sum_{a=1}^{l} (\Delta X_a^S \odot \Delta X_a^S)}, \quad p = 1..P, s = 1..m$$

    - $X_a \in R^{P\times 1}$ is the a-th predictor for all points.
    - $X_a^S \in R^{m\times 1}$ is the a-th predictor for all monitors.
    - 1_P and 1_m are column vectors of ones of sizes P and m, respectively.
    - This yields a new distance matrix D$\in R^{P\times m}$ from all points to all monitors.

3. Self-exclusion for monitors.

   - If a point p is itself a monitor ($p \in S$), change the value of its distance to infinity so that it is excluded from the average when using the *argsort* function.

$$D_{p,s=p} = \infty$$

   - This ensures a location with a monitor does not select itself as a neighbor.

4. Select k nearest monitors.

   - For each point p, select indices of the k nearest monitors:

$$N_p = argsort_p\left(D_{p,:}\right)[1{:}k];\ \forall p \in \{1,2,3,\dots,P\}$$

   - Neighbor index matrix:

$$N = \begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_P \end{bmatrix} \in R^{P \times k}$$

   - Distance to the selected neighbors is retrieved:

$$D_P^{(k)} = [D_{p,N_{(p,1)}}, \dots, D_{p,N_{(p,k)}}]$$

5. We include also the binary indicator for missing Y, where Y is a target such as Ta or PM.

$$Z_{pj}^t = \begin{cases} 1, & if\ Y_{N_{(p,j)}}^t is\ observed\ (not\ NA). \\ 0, & if\ Y_{N_{(p,j)}}^t is\ missing\ (NA). \end{cases}$$

$$S_p^t = \sum_{j=1}^{k} Z_{pj}^t$$

6. Inverse Distance Weighting (IDW): If a neighbor's value is NA, the weighted average is computed using only the available neighbors.

   - Raw weights:

$$\omega_{pj} = \frac{1}{D_{pj}^{(k)} + \varepsilon} \; ; \; \forall p \in \{1,2,3, \dots, P\}$$

- Normalized row weights per point p and date t:

$$\bar{\omega}_{pj}^{t} = \frac{\omega_{pj} * Z_{pj}^{t}}{\sum_{\substack{j=1 \\ z_{pj}^{t} \neq 0}}^{k} \omega_{pj} * Z_{pj}^{t}} \; ; \; \forall p \in \{1,2,3, \dots, P\}, \; t \in \{1,2,3, \dots, T\} \; ; \; if \; S_p^t > 0$$

7. Weighted NA-aware for FE computation for variable generation at prediction stages.

$$\bar{Y}_p^t = \begin{cases} \sum_{j=1}^{k} \bar{\omega}_{pj}^{t} * Y_{N_{(p,j)}}^{t} \; , if \; S_p^t > 0 \\ NA, \quad if \; S_p^t = 0 \end{cases}$$

- For all points.

$$\bar{Y}^t = \begin{bmatrix} \bar{Y}_1^t \\ \bar{Y}_2^t \\ \vdots \\ \bar{Y}_P^t \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{k} \bar{\omega}_{1j}^{t} * Y_{N_{(1,j)}}^{t} \\ \sum_{j=1}^{k} \bar{\omega}_{2j}^{t} * Y_{N_{(2,j)}}^{t} \\ \vdots \\ \sum_{j=1}^{k} \bar{\omega}_{Pj}^{t} * Y_{N_{(P,j)}}^{t} \end{bmatrix}, \quad with \, \bar{Y}_p^t = NA \; if \; S_p^t = 0$$

- Where $\bar{Y}^t \in \mathrm{R}^{P \times k}$ is the vector of average non-missing neighbor targets for all points at time t.

8. Calculation of FE for each time step t in the modeled period (e.g., a year).

$$FE = [\bar{Y}^1, \bar{Y}^2, \dots, \bar{Y}^T] \in \mathrm{R}^{\mathrm{P} \times \mathrm{T}} \quad \text{(Eq. 17)}$$

Improving Model Performance by Tuning the *k* Parameter

The generation of new predictors through the KNN model relies heavily on the choice of the parameter *k*, which determines the number of neighboring observations used when computing the engineered features (FEs). Understanding and optimizing this parameter is crucial because it directly affects the quality, completeness, and informativeness of the resulting features.

When *k* is small, the constructed feature reflects primarily the values of the closest neighbors. This typically yields high local accuracy, as the feature closely represents observed values in the immediate vicinity. However, a small *k* also increases the likelihood of missing values in the FE, because some of the nearest neighbors may themselves have missing target values. Consequently, the resulting feature may be highly precise in areas with dense data but incomplete in regions with sparse observations.

Conversely, increasing *k* incorporates information from more distant neighbors. This reduces the proportion of missing values in the FEs and improves completeness but introduces potential bias, as the feature increasingly reflects the characteristics of more distant locations that may not be representative of the local context. Therefore, selecting *k* involves balancing two competing objectives: (1) maximizing precision by using local neighbors and (2) minimizing missingness by including more observations. For this reason, the *k* parameter must be optimized systematically.

Dealing with Missing Data in FEs

Because FEs are computed from observed variables, missing values are common, especially for target variables such as LST and AOD, which often exhibit substantial gaps. By contrast, monitoring data (e.g., station-based targets) is nearly complete in most cases. This difference makes the imputation stage particularly relevant for the engineered features.

Two main strategies can be used to impute missing values:

1. Simple imputation methods: From our experiments, when the target variable exhibits relatively low missingness (e.g., less than 25%), simple techniques such as using the mean or median (calculated for each season) can give reasonable results.
2. Model-based imputation methods**:** When missingness is higher, or when greater precision is required, model-based imputation techniques consistently produce better results. These approaches use information from other fully observed predictors to estimate missing values in a more informed and data-driven manner (Stekhovena and Bühlmann, 2011). Equation 18 illustrates this situation.

Let $I_1, I_2, I_3..., I_l$ denote spatial predictors (such as latitude, longitude, or elevation) and x_{1t}, $x_{2t}, ..., x_{nt}$ denote spatiotemporal predictors that are fully complete. Each engineered feature can be imputed by fitting a predictive model of the form:

$$FE = FE_{pt} = f(I_1, I_2, I_3 \dots I_l, x_{1t}, x_{2t}, x_{3t} \dots, x_{nt}) + \varepsilon_{pt} \quad \text{(Eq. 18)}$$

where ε_{pt} is the residual error and $f(\cdot)$ can be implemented using various machine-learning regression learners (e.g., linear regression, random forests, or gradient boosting methods). This model-based approach ensures that missing values in the engineered features are estimated using all relevant information rather than relying on simple imputation or heuristic approach.

Cut-Off Point Rule for Selecting the Optimal *k in* FEs

To systematically identify the optimal number of nearest neighbors, multiple engineered features can be generated by varying the parameter *k* over a pre-defined range, e.g. k=1, 2, …, H. Each resulting FE is then subjected to an imputation step using the methods described above, either simple or model-based. This process yields a comprehensive set of candidate features for subsequent experiments.

Next, a rigorous feature selection procedure is applied to determine which FE, and therefore which *k*, delivers the most informative and predictive value for the target variable *Y*. Each candidate FE is sequentially added to a target regression model (see Equation 19), and its predictive performance is evaluated using the root mean square error (RMSE). We used traditional *k*-fold cross-validation to check model performance and reduce overfitting, but more advanced validation methods can also be used, as described in Chapter 2.

The optimal value of *k* is then selected based on a **cut-off point rule** in model performance. Specifically, *k* is considered optimal when further increases do not lead to meaningful improvements in RMSE. The FE constructed with this selected *k* is then used in all subsequent predictive models for *Y*.

Mathematically, for each *k* value we fit:

$$Y_{pt} = f(I_1, I_2, I_3 \dots I_l, x_{1t}, x_{2t}, x_{3t} \dots, x_{nt}, FE_k) + \varepsilon_{pt} \quad \forall k \in \{1,2,3 \dots H\} \quad \text{(Eq. 19)}$$

This iterative process ensures that the final chosen feature is both locally accurate and sufficiently complete.

Extension to Multiple Attribute Sets

The procedure described before can be repeated using different combinations of input attributes (spatial predictors) $I_1, I_2, I_3 \ldots, I_l$, such as latitude, longitude, elevation, etc., to generate multiple sets of engineered features. Each set then undergoes the same tuning and selection procedure to determine its own optimal k.

The final predictive model integrates the selected features across the defined spatial dimensions (Equation 20 illustrates this), ensuring that the engineered features are both locally accurate and globally informative. This can be expressed as:

$$Y_{pt} = f\left(I_1, I_2, I_3 \ldots I_l, x_{1t}, x_{2t}, x_{3t} \ldots, x_{nt}, FE_{\text{attr set 1,k1}}, FE_{\text{attr set 2,k2}} \ldots, FE_{\text{attr set 3,k3}}\right) + \varepsilon_{pt} \quad \text{(Eq. 20)}$$

Each FE is associated with its own optimized number of neighbors k_i, determined through the cut-off point rule. By incorporating FEs derived from multiple attribute sets, the model captures both local and broader spatial–temporal dependencies, enhancing predictive accuracy while maintaining robustness against missing values.

Summary of the Tuning Workflow

1. Generate engineered features for a sequence of k values.
2. We filled in missing values using simple or model-based methods. To make this easier, we used random K-fold cross-validation.
3. Evaluate the predictive contribution of each feature (FE) using the root mean square error (RMSE) within a cross-validation framework, either with random CV or spatial blocking CV (described in more detail later in this chapter).
4. Apply the cut-off point rule to select the optimal k for each feature set.
5. Integrate the selected FEs (with their optimal k) into the final predictive model.

This workflow produces engineered features that strike an effective balance between local specificity and completeness, ensuring robust and accurate predictions and potentially improving model performance

3.4.2 How to integrate Feature Engineering During Cross-Validation

Integrating feature engineering within cross-validation requires careful planning, as it can introduce computational challenges and time constraints. The goal here is not to identify the

best cross-validation strategy but to adopt an approach that suits the model's purpose, evaluates performance realistically, and ensures efficient implementation.

Feature engineering must be handled carefully to avoid data leakage. Based on our experiments (results from articles about Ta and PM), we propose the following generic protocol. Minor adaptations are made in individual studies, but this methodology ensures robust evaluation.

The first key principle is that cross-validation should simulate the prediction scenario of the modeled problem (from chapter 2). This allows us to choose a CV strategy aligned with the problem's characteristics. Similar to the Feature generation, for CV we can broadly distinguish two types of models:

1. Imputation models (e.g., LST, AOD)
2. Prediction models (e.g., Ta, PM)

3.4.2.1 Feature Engineering for Satellite-Derived Data Imputation (LST-AOD)

The core principle of integrating feature engineering within cross-validation lies in the premise that features must be computed using only the available training data. In each cross-validation fold, features are derived exclusively from the training set, excluding any held-out observations from the computations. Additionally, the nature of the problem should guide the choice of cross-validation strategy. For MODIS LST and AOD, we observed gaps occurring at different locations and times, with no single location missing data for an entire year. This behavior suggested that random K-fold cross-validation was a suitable approach, even though it is well known that spatial blocking cross-validation produces more robust evaluations.

While it may slightly overestimate performance, random K-fold cross-validation is generally faster to apply, and results are still evaluated across several scenarios. Spatial blocking cross-validation, on the other hand, provides a more realistic performance estimate but imposes limitations on which features can be included, and the experiment should be carefully designed to simulate the spatial behavior of the target. Both strategies are valid as long as their respective limitations are acknowledged (Roberts et al., 2017).

At this point, it is also important to distinguish engineered features as target-based or non-target-based**:**

- Non-target-based features can be computed directly from the training data and then applied to the validation or test set without risk of leakage.
- Target-based features (e.g., averages of LST/AOD, interpolations from points, KNN features) require careful consideration. In exposure modeling, both space and time are critical:

    - If a feature depends on the actual spatial location (e.g., a weekly average of a site), it can be used with random K-fold CV, but cannot be used in purely spatial blocking CV, as held-out points would not have the feature available.
    - If a feature does not depend on the actual spatial location (e.g., neighbors' aggregations or buffer averages excluding the location self), it can be used in both random and spatial blocking CV, and its computation has to be solely from the training fold. Random K-fold CV may introduce slight leakage due to overlapping locations between training and test sets (not times), whereas spatial blocking CV avoids this issue entirely.

For our thesis studies, we preferred random K-fold cross-validation for several reasons. First, satellite imputation was not our primary target, and spending excessive effort on fine-tuning it could have diverted attention from the main goals. As discussed previously, this cross-validation strategy is valid if properly applied, including careful integration of feature engineering.

For MODIS LST and AOD, missing values occurred at different points in time and space, with a generally random pattern during most of the year but markedly higher and more clustered missingness in winter. Given this mixed structure, we employed random K-fold cross-validation as our main approach. This strategy reflects our intended use case, imputing missing values at the same locations across different times, and provides a straightforward implementation of feature engineering within each fold. We acknowledge, however, that random K-fold CV can slightly overestimate performance in the presence of spatially or temporally clustered missingness compared to spatial or spatio-temporal blocking CV (Roberts et al., 2017). In scenarios where predictions are required at new locations or during periods of strong missingness, spatial or spatio-temporal CV would provide a more conservative and realistic performance estimate.

A valid alternative approach using spatial blocking cross-validation would involve creating multiple small subsamples (e.g., five different subsamples representing 5% of points across the territory, obtained through clustering so that each subsample adequately represents all areas). Clustering can be done for instance with the k-means model over the geographical coordinates (Valavi et al., 2019). In each iteration, one subsample is held out, features are computed from the remaining training data, and predictions are made for the held-out points. Only target-based features that do not depend on the actual spatial location can be included in this scenario. Feature engineering is performed once per training fold, and model performance is evaluated across all held-out subsamples.

3.4.2.2 Feature Engineering within Cross-Validation for Prediction (Ta-PM)

As discussed previously, the cross-validation approach should reflect the actual prediction scenario and consider whether the features to be included depend on the target or not. For prediction, the problem is more straightforward because features based on averages per location-time (point-specific features) are not valid.

Spatial correlation is particularly important here, as the monitoring locations are nearly fixed and predictions are required at new specific locations. Therefore, a spatial blocking cross-validation approach is necessary. Feature engineering is performed once per fold, resulting in a number of implementations equal to the number of folds.

The choice of the spatial blocking strategy depends on the density and distribution of the monitoring network:

- If the network is dense and covers most of the territory, clusters of monitors can be created to form held-out folds. In each iteration, one cluster is held out, features are computed from the remaining monitors, and the model is trained. Feature engineering is performed using the training clusters, excluding the held-out monitors, and performance is evaluated. This process is repeated for each cluster, and overall performance is averaged with feature engineering performed one time per cluster (Valavi et al., 2019). Figure 4 illustrates this scenario.

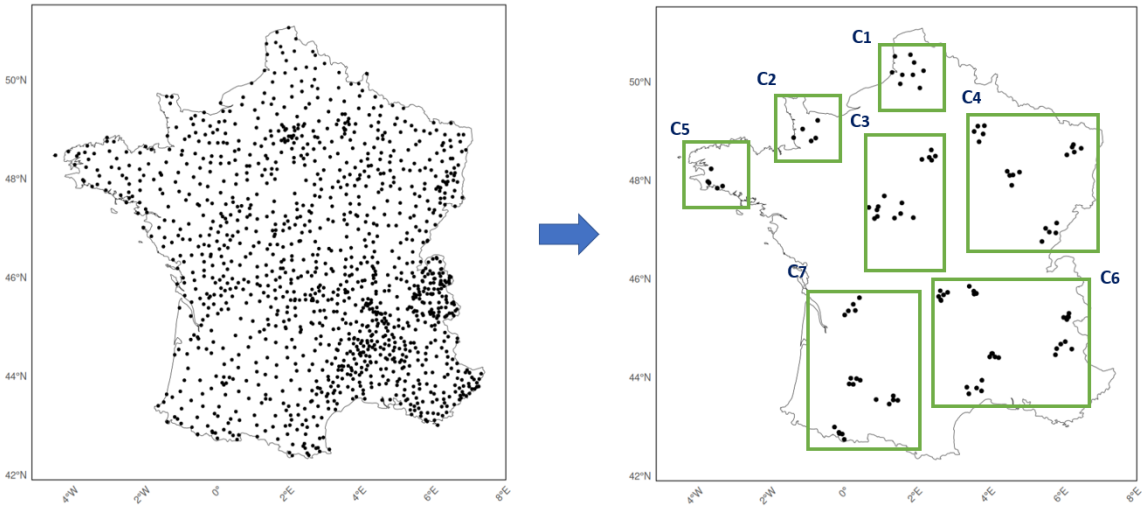


Fig. 4. Illustrative example of a dense network of monitors. Held-out folds (C_1-C_7) are taken per clusters of monitors in different areas.

- If the network is sparse, creating clusters is not feasible because it would reduce training data too much. In this case, individual monitors are held out sequentially. Models are trained on all other monitors, features are computed using only the training data, and predictions are made for the held-out monitor. This process is repeated for each monitor, and results are averaged. Feature engineering is performed once per held-out monitor (see Figure 5).

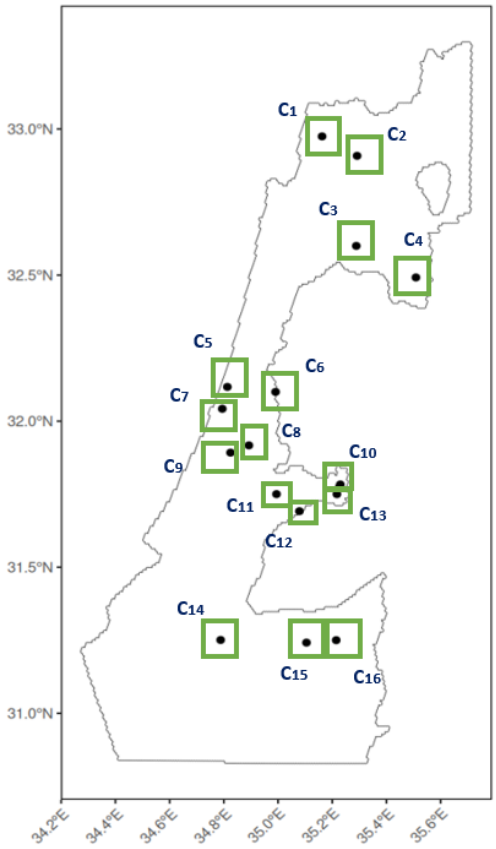


Fig. 5. Illustrative example of a sparse network of monitors. Held-out folds (C_1-C_{16}) are taken per monitor location.

3.4.2.3 Summary of integrating feature engineering within cross-validation

Table 4 provides examples of engineered features discussed in this section, detailing their types, the cross-validation methods they can be used with, and the prediction tasks they support. Note that binary dummies, for instance, can be applied to both historical and forecasting models, even though the approaches examined in this dissertation assume prediction based only on historical data (i.e., not future steps).

Table 4. Brief description of engineered features and their integration in cross-validation.

Feature Type	Example Feature	CV Method Allowed	Stage of Computation	Prediction Type
Non-target-based	Spatial spline, Daily average all points	Random K-fold, Spatial blocking CV	Compute on training fold, apply to validation/test fold	Spatial only
Non-target-based	Binary dummy	Random K-fold, Spatial blocking CV	Compute on training fold, apply to validation/test fold	Spatial, temporal
Target-based, spatial point dependent	Weekly average per site (LST/AOD)	Random K-fold only (substantial leakage)	Compute on training fold, apply to validation/test fold	Spatial only
Target-based, spatial point independent	KNN FEs, buffer averages (excluding point itself)	Random K-fold (substantial leakage), Spatial blocking CV (zero leakage)	Compute on training fold, apply to validation/test fold	Spatial only

3.4.2.4 Major Considerations in Feature Engineering Integration

The need for Big Data alternatives

Previous explanations demonstrate that integrating feature engineering into cross-validation demands careful attention to multiple aspects in order to guarantee a robust model validation. One fundamental principle is that engineered features must be calculated using only training data. This applies to both imputation and prediction phases. Calculating features using held-out locations would introduce data leakage, artificially inflating model performance and leading to incorrect conclusions. Maintaining this strict separation between training and validation sets is therefore essential to obtain unbiased estimates of model generalization.

As a result, feature computations need to be repeated for each fold of the cross-validation process. The total number of calculations depends on the number of folds, regardless of the specific cross-validation used. This becomes particularly important when tuning the number of nearest neighbors k. If k is varied across a predefined range (e.g., 1 to 20) and the number of cross-validation folds is, for example, 5, then the total number of feature computations is

$20 \times 5 = 100$. This illustrates the computational complexity of the approach, which requires efficient management of several experiments.

Distributed computing frameworks, such as Spark or H2O, can facilitate this process by enabling parallel execution and distributed processing, significantly reducing runtime for large datasets.

Distance Matrix Update per Cross-Validation Fold

Another important consideration is that feature engineering calculations must be dynamic and recalculated per iteration during cross-validation. This means that features are not static or precomputed once for all, but instead adapt to the training data available. In general terms, our formulation for advanced feature engineering follows this principle. Each new predictor FE depends on three arguments (see Equation 21): the distance matrix *D*, the number of nearest neighbors *k*, and the target variable *Y*. Formally, a feature can be expressed as:

$$FE = f(D, k, N(D, k), Y) \qquad Eq.21$$

This functional formulation is applicable to both imputation and prediction stages, ensuring reproducibility and flexibility. Besides, our vectorized implementation allows for fast computation even when working with large spatial grids.

Through our experiments, we observed that the main computational bottleneck for applying feature engineering based on the KNN model lies in calculating the distance and neighbor matrices (*D*, *N*). For large territories, such as an entire country, computing these matrices can take hours (e.g., France). However, once the distance matrix has been calculated, subsequent computations of features for different cross-validation folds can be performed efficiently without recomputing the entire matrices.

To handle feature calculations at intermediate iterations during cross-validation, the approach is straightforward: setting the columns corresponding to the held-out locations in the distance matrix *D* to infinity. This switch ensures that held-out locations are never selected as neighbors during the feature calculation, while keeping the precomputed distances to the remaining training points intact. This strategy allows feature engineering to be performed at each CV iteration simply by modifying columns in the distance matrix, rather than recomputing *D* entirely.

Moreover, this approach also facilitates handling missing values in a consistent way. By combining the neighbor matrix with a binary indicator of observed values, the feature calculation can ignore missing neighbors and perform averaging only over available values. For prediction tasks, inverse distance weighting can be applied in combination with this masking strategy, ensuring that closer neighbors contribute more to the weighted average while still excluding held-out points.

3.4.3 Similarities and Differences between FEs and Interpolation Techniques

Our FE approaches, applied to environmental exposure modeling of variables such as Ta and PM, can be compared to traditional geostatistical methods like inverse distance weighting (IDW) and universal kriging.

These classical techniques are well-established for predicting values at unobserved locations using information from nearby observations (Isaaks & Srivastava, 1989; Oliver & Webster, 1990). Both rely on the spatial configuration of the monitoring network and weight observations based on their distances, assigning greater importance to nearby measurements when estimating values at unsampled locations. Their dependence on spatial distance and the gradual weighting decrease makes them conceptually similar to the ideas underlying our feature engineering proposal.

In this section, we provide a brief comparison of our feature generation method with IDW interpolation and universal kriging, discussing how these classical methods can be interpreted as strategies for generating spatial features summarizing local neighborhood information at each grid point. We highlight their similarities, differences, and limitations in comparison with our FE approach.

IDW interpolation

According to Burrough and McDonnell (1998), inverse distance weighting (IDW) is a deterministic interpolation method that provides a single, fixed estimate for unobserved locations by using the values and distances of neighboring points, without explicitly modeling uncertainty or spatial correlation. The main parameters of IDW include the number of neighbors considered (k), the distance weighting function, and a small regularization term (ε) added to prevent division by zero when the estimation point coincides with an observed location. The influence of each neighbor decreases with distance, meaning that closer points

contribute more strongly to the predicted value. Equation 22 illustrates this underlying principle.

$$\hat{z}(x_0) = \frac{\sum_{i=1}^{k=} \frac{z(x_i)}{(d(x_0, x_i) + \varepsilon)^p}}{\sum_{i=1}^{k=} \frac{1}{(d(x_0, x_i) + \varepsilon)^p}} \quad \text{Eq. 22}$$

where $\hat{Z}(x_0)$ is the prediction at location x_0, $z(x_i)$ is the observed value of the neighbor i, $d(x_0, x_i)$ is the distance between the prediction location x_0 and neighbor x_i, k is the number of neighbors considered, p is the power parameter controlling the weights decay with distance, and ε a regularization term.

Universal Kriging

On the other hand, universal kriging is a more advanced geostatistical method that is not fully deterministic, as it models the target variable as a spatially correlated random field combining a deterministic trend with a stochastic residual component. This formulation allows universal kriging to provide both predictions and associated uncertainty.

Modeling the target as a spatially correlated random field implies that nearby locations tend to exhibit similar values, while also accounting for large-scale spatial trends and the dependence structure between observations (Cressie, 1993). The deterministic trend is typically represented as a linear function of spatial coordinates, whereas the stochastic residual captures spatially correlated deviations from this trend.

Universal kriging requires specifying a variogram model, which characterizes how similarity between observations decreases with distance. The main parameters of the variogram are the nugget, representing measurement error or microscale variability; the sill, corresponding to the overall variance of the field; and the range, defining the distance beyond which observations become effectively uncorrelated. Predictions are obtained by solving a system of linear equations that optimally weights neighboring observations according to both their distance and the spatial covariance structure. While we do not explore further technical details here, this general formulation serves to clarify the underlying principles of the method. Equation 23 summarizes its logic.

$$\hat{z}(x_0) = \sum_{i=1}^{k} \lambda_i \, z(x_i) + trend(x_0) \quad \text{Eq. 23}$$

In this formula, $\hat{z}(x_0)$ represents the prediction at location x_0, $z(x_i)$ the observed value at location x_i, and λ_i are the kriging weights, which are determined based on the spatial correlation described by the variogram. The stochastic residual is represented by the product $\lambda_i z(x_i)$. The term *trend(x_0)* represents the deterministic component of the model (often assumed as a linear function of spatial coordinates), and *k* is the total number of neighbors included in the prediction. The kriging weights are computed by solving a system of linear equations that minimizes the prediction variance while ensuring unbiased estimates using the nugget, sill, and range parameters in the variogram.

A Comparative Analysis of Neighborhood-Based Approaches

When comparing our feature engineering approach with IDW interpolation and universal kriging, we observed several similarities and differences. All three methods rely on the spatial configuration of points and aggregate information from neighboring locations to inform predictions. In our approach, neighbor selection is explicit and can be tuned through the parameter *k*, while IDW also depends on the number of neighbors and distance-based weights. Universal kriging, by contrast, introduces a model-based weighting through the variogram, which captures more complex spatial patterns but requires stronger assumptions and additional parameter estimation.

Both IDW and universal kriging could, in principle, be integrated as methods for feature generation. For example, predicted surfaces, kriging weights, or even spatial residuals could be computed at each grid point and then used as additional predictors in a machine-learning model. It can be achieved following a similar protocol to the one described by us when applying FE within cross-validation.

In practice, however, incorporating these methods within a feature engineering workflow can be both debatable and more complex than initially anticipated. IDW is relatively straightforward to implement and bears strong similarity to our approach. Since our framework already relies on precomputed distance matrices and vectorized operations, it could be easily adapted to replicate the IDW behavior by adjusting parameters such as the weighting exponent (to simulate decay) or the distance function used for computing *D*. Universal kriging, in contrast, is considerably more demanding computationally. It requires fitting a variogram and solving a kriging system separately for each CV iteration in order to prevent data leakage, making it highly sensitive to parameter choices and much more time-

consuming, which may be unfeasible given current computational capacity. Despite these challenges, the strengths of universal kriging, particularly the inclusion of stochastic components in predictions, represent a promising avenue for future research. Such an approach could potentially enhance FE methods like ours by integrating measures of uncertainty and spatial dependence that go beyond purely deterministic weighting schemes.

In Table 5, we present a brief comparison between our approach and the two other interpolation methods, emphasizing and discussing several key aspects.

Table 5. Brief description of engineered features and their integration in cross-validation.

Aspect	**Feature Engineering (FEs)**	**Inverse Distance Weighting (IDW)**	**Universal Kriging (UK)**
Core principle	Aggregates information from nearest neighbors to create new predictors	Weighted average of k nearest observed points	Weighted combination of neighbors based on variogram model
Neighbor selection	Explicit k-neighbor parameter, flexible	Number of neighbors (k), distance metric, epsilon	Implicit via variogram parameters: nugget, sill, range; trend model
Distance-based weighting	Yes, typically inverse distance or user-defined	Yes, inverse distance weighting	Yes, through spatial covariance function
Captures local spatial structure	Yes, via k-neighbor aggregation	Yes, distance-weighted	Yes, weighted via variogram, includes global trend

Aspect	Feature Engineering (FEs)	Inverse Distance Weighting (IDW)	Universal Kriging (UK)
Assumptions about spatial process	Minimal	Minimal	Requires assumptions: spatial stationarity, variogram model, trend function
Deterministic / Stochastic	Deterministic	Deterministic	Not fully deterministic; includes stochastic residuals to capture spatial correlation
Computational efficiency	High, especially with precomputed distance/neighbor matrices.	Moderate; scales poorly for large numbers of points if recalculated repeatedly. If developed with precomputed distance/neighbor matrices, then high computational effiency can be reach.	Low; variogram fitting and solving kriging system are computationally intensive for large datasets or multiple folds in cross validation.
Flexibility across CV folds	Easy to exclude held-out points and reuse matrices.	Moderate; recalculation needed for each fold.	Challenging; variogram and kriging system must be recomputed to prevent leakage.

Aspect	Feature Engineering (FEs)	Inverse Distance Weighting (IDW)	Universal Kriging (UK)
Advantages	Fast, reproducible, flexible, allows tuning and parameter experimentation.	Simple, interpretable, deterministic, easy to implement.	Optimal predictions. Models autocorrelation and global trends, provides uncertainty estimates. Higher accuracy.
Limitations	Does not model spatial autocorrelation beyond local neighborhood.	Sensitive to choice of k, distance metric, and epsilon; does not model global trend or stochastic variability	High computational cost, sensitive to parameter selection, complex to implement in cross-validation or feature generation workflows

3.4.4 Conclusions from feature Engineering and its integration within validation techniques

In this section, we explained how to apply both basic and advanced feature engineering techniques and how these can be adapted to the context of exposure modeling. The application of these features helps models capture temporal and spatial relationships that are not explicitly represented in the original predictor sets, establishing new relationships while benefiting from dimensionality reduction.

In our main dedicated chapter, we introduced and described our advanced feature engineering approach based on the KNN model for feature generation, offering flexibility in the selection

of spatial attributes to maximize the spatial information extracted from predictors. We detailed the methodology for developing features using the KNN approach and discussed how to tune the parameter *k* effectively.

We also clarified what feature engineering entails and how it should be correctly integrated within cross-validation schemes, ensuring a strict separation between training and held-out data during feature computation. Additionally, we discussed which cross-validation techniques are more suitable for specific types of features and highlighted cases where certain features can or cannot be used depending on the validation approach. We emphasized the computational limitations and challenges in terms of the number of run experiments for FE computations, underlining the need for more advanced frameworks capable of handling large-scale problems (Big Data).

Finally, we presented a straightforward comparison between our innovative FE approach and other well-known methods in environmental modeling, such as IDW interpolation and universal kriging. We discussed their respective strengths and weaknesses and explained why our approach represents a suitable choice in terms of flexibility, feasibility, and complexity.

.

Chapter 4

In this chapter, we present the three research projects carried out as part of this PhD thesis. Each project has been developed in the form of a scientific article prepared according to the formatting and submission requirements of peer-reviewed journals (such as the *International Journal of Climatology* and *Environmental Research*).

The articles are reproduced here either in the versions that were submitted to the journals or in their final complete versions prior to submission. For each article, we indicate its current status (e.g., in preparation, submitted, under review, or accepted) and specify the PhD candidate's contribution to the research and manuscript preparation.

4.1 Article 1

Full reference:

Briceno, J. D., Dorman, M., Barbalat, G., Colicino, E., Stafoggia, M., Lepeule, J., & Kloog, I. (2024). A new hybrid approach applying feature engineering to model high-resolution air temperature across Israel.

Status: The manuscript is currently under review.

PhD candidate's contribution:
The PhD candidate designed the study, downloaded and processed the data, performed the statistical analyses, interpreted the results, and wrote the manuscript. The candidate was the main author of the paper.

(The article in its complete version follows on the next pages. Tables, figures, and equations are numbered as in the submitted article and have not been modified.)

A new hybrid approach applying Feature Engineering to model high resolution air temperature across Israel

Juan David Briceno, Michael Dorman, Guillame Barbalat, Elena Colicino, Massimo Stafoggia, Johanna Lepeule, Itai Kloog

A new hybrid approach applying Feature Engineering to model high resolution air temperature across Israel

Juan David Briceno[a,b*], Michael Dorman[b], Guillame Barbalat[a], Elena Colicino[d], Massimo Stafoggia[c], Johanna Lepeule[a1], Itai Kloog[b,d1*]

[a] University Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Development and Respiratory Health, Institute for Advanced Biosciences (IAB), La Tronche, France.
[b] Department of Environmental, Geoinformatics and Urban Planning Sciences, Ben-Gurion University of the Negev, P.O.B. 653, Beer Sheva, Israel
[c] Lazio Regional Health Service, Department of Epidemiology, Unit of Environmental and Occupational Epidemiology and Cancer Registry, Roma, Italy
[d] Department of Environmental Medicine and Climate Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA

* Corresponding author. Institute for Advanced Biosciences (IAB), Site Santé – Allée des Alpes, 38700 La Tronche, France. E-mail: juan.briceno-guerrero@univ-grenoble-alpes.fr, Tel: +33768296374

[1] these authors contributed equally

Abstract

The increasing global planetary temperature, combined with rapid urbanization, underscores the urgent need to evaluate human exposure to air temperature. This study proposes an innovative methodology to reconstruct air temperature (Ta) from 2004 to 2022 at both 1 km and 250 m resolutions across Israel, leveraging a novel feature engineering (FE) approach. Our method enables the daily prediction of three key Ta measures: minimum (T_{min}), average (T_{mean}), and maximum (T_{max}) temperatures. First, we employed FE techniques to develop geographically weighted predictors and introduced a gap-filling procedure for satellite-derived land surface temperature (LST) data, producing continuous surface temperature (Ts) fields. Next, we applied Extreme Gradient Boosting (XGBoost) models to calibrate and predict Ta using spatial and spatio-temporal predictors, such as elevation, LST data from MODIS, ECMWF Reanalysis, and EUMETSAT SEVIRI. Finally, we downscaled the 1km Ta estimations to a 250 m resolution by calibrating the residuals of the 1 km model with variables at higher spatial detail such as Landsat LST data. The resulting models demonstrate excellent cross-validated performance, with RMSE ranging from 0.53°C to 1.01°C. This methodology offers significant advancements in exposure assessment methods for various environmental modeling applications.

Keywords

Air temperature; land surface temperature; MODIS; Feature Engineering; K-Nearest Neighbors; Machine Learning algorithms,

1. Introduction

The changing climate and the global rise in air temperatures (Ta) represent some of the most pressing challenges of our time. As such, it has become increasingly critical to thoroughly evaluate the adverse human health effects associated with this global increase in Ta. Extensive research has shown that temperature fluctuations can significantly impact health, exacerbating conditions such as adverse birth outcomes (Basu et al., 2010; Kloog et al., 2015), increased heart rate, changes in peripheral blood circulation, higher blood viscosity, and elevated cholesterol levels (Keatinge, 1997), as well as leading to dehydration and renal failure (Schwartz et al., 2004). In Israel, there has been a notable increase in the number of warm days and nights, with summers becoming significantly hotter and the frequency and intensity of heatwaves rising (Green et al., 2013).

To accurately assess the impact of this environmental factor, it is crucial to obtain high-resolution estimates of Ta that capture both spatial and temporal variability. Accounting for spatio-temporal variations enables the identification of critical exposure windows, which are particularly relevant for vulnerable populations, such as the elderly, individuals with chronic illnesses, and those living alone (Zhou et al., 2020). Furthermore, ensuring the completeness of estimated Ta surfaces significantly reduces bias in environmental studies by providing more accurate exposure profiles across different regions and time periods. In contrast, relying solely on remotely sensed measurements at specific points can yield similar results for cities and surrounding rural areas. However, this approach is misleading, as urban areas typically experience higher temperatures due to greater heat accumulation and slower heat dissipation compared to rural landscapes (Hough et al., 2020).

To achieve this, several strategies have been proposed, most of which involve numerical prediction models based on physical laws. However, due to computational limitations, these calculations have often been restricted to medium spatial resolutions (4 km). The most accurate approximations, as measured by Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), have been achieved through the use of Machine Learning (ML) models combined with satellite data. In this context, MODIS Land Surface Temperature (LST) has proven to be effective indicators of spatiotemporal variations in Ta. ML-based strategies have demonstrated high levels of performance across large and heterogeneous

areas, with reported RMSEs of 2.2 °C in the USA (Kloog et al., 2014), 1.06°C in Warsaw, Poland (Hassani et al., 2024), 1.7 °C in France (Kloog et al., 2017), and 1.2 °C in Israel (Rosenfeld et al., 2017). More recent studies have further enhanced prediction accuracy by employing advanced techniques such as ensemble learning, which effectively minimizes generalization error (e.g., RMSE of 1.3 °C in France; Hough et al., 2020).

This study introduces several innovations to improve Ta estimation, including novel feature engineering (FE) techniques applied to LST and Ta data enabling a more robust reconstruction of diurnal air temperatures. A key innovation is the implementation of FE approaches based on the k-nearest neighbors (KNN) model, which employs distance-based calculations to neighboring values to improve regression predictions.

2. Data and methods

2.1 Study area

This study focuses on the State of Israel, which covers an area of approximately 21,670 km² and is characterized by large variations in elevation throughout its territory, with its highest point (Mount Hermon) reaching an elevation of 2,807 meters above sea level, while its lowest point is the Dead Sea, situated about 430 meters below sea level in the Jordan Rift Valley. In terms of dimensions, Israel spans approximately 470 km from north to south, with its widest point measuring around 135 km. The country is divided into four main geographical regions: the Mediterranean coastal plain, which experiences long, hot, rainless summers and short, rainy winters; the Central Hills, which are predominantly semi-arid with occasional rainfall, particularly in the northern areas; the Jordan Rift Valley, which has an arid climate and receives less than 100-300 mm of rainfall per year; and the Negev Desert, characterized by a hot desert climate with minimal rainfall (Ochsenwald et al., 2019).

Annual precipitation in Israel varies considerably, ranging from 1,000 mm in the northern Golan Heights to approximately 30 mm in the southern regions. Rainy and cold periods generally occur between November and March, while dry and hot conditions dominate from June to September (see Figure 1). During the summer months, temperatures in desert areas (southern regions) can reach between 35°C and 40°C.

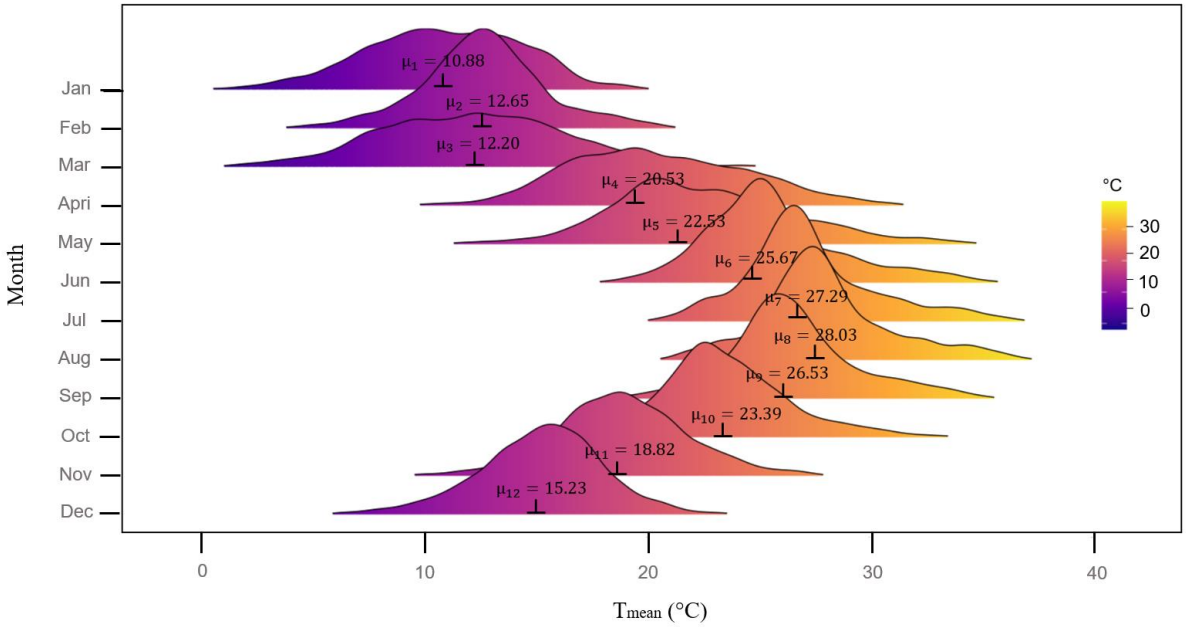


Fig. 1. Observed mean air temperature in Israel during the year 2022. Data retrieved from the Israel Meteorological Service (IMS). At the time, Israel had 85 weather stations.

The population in Israel is unevenly distributed, with more than half concentrated in the coastal plain, while the Negev Desert in the south remains sparsely populated. The total population is approximately 8.9 million inhabitants (Central Bureau of Statistics, 2018).

2.2 Meteorological data

2.2.1 Monitor data

As shown in Figure 2, 85 weather stations were operating across Israel in 2022, their control and maintenance are handled by the Israel Meteorological Service (IMS). Hourly temperature data (Ta) has been collected from stations operating between 2004 and 2022. These stations are distributed across the study area, with a higher concentration near urban regions in the northern part of the country. Fewer stations are located in the southern desert areas. Since IMS data is available every quarter-hour, we aggregated it by calculating daily mean, minimum, and maximum values (T_{mean}, T_{min}, and T_{max}). Additional details on the number of stations deployed over the years can be found in Appendix 1.

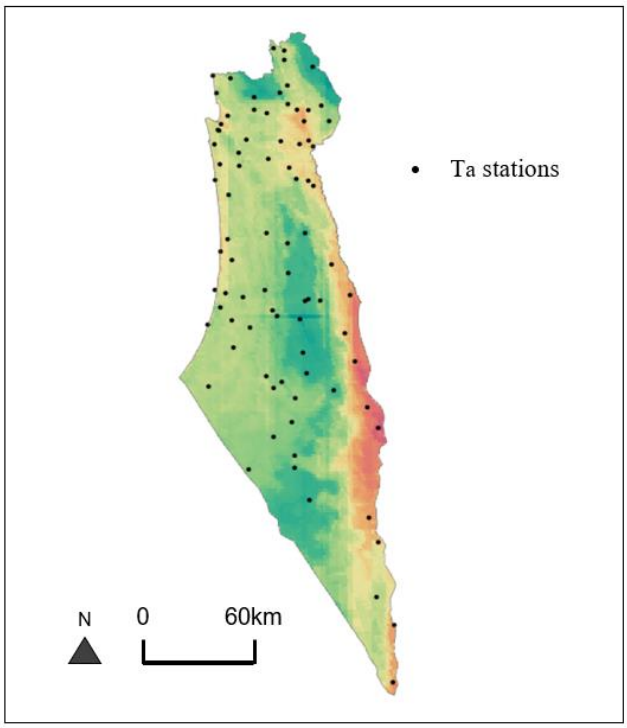


Fig 2. As of 2022 year, 85 Ta monitoring stations are distributed across Israel, with data provided by the Israel Meteorological Service (IMS).

2.2.2 Remotely sensed surface skin temperature

Land Surface Temperature (LST) data was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA's Aqua and Terra satellites, with a spatial resolution of 1 km². These satellites continuously monitor the Earth's surface, each following a distinct polar orbit that provides one daytime and one nighttime pass over a given location per day. Combining observations from both Terra and Aqua results in up to four measurements per location each day. Although differences in overpass times may cause slight temporal variations, the two satellites provide complementary information. For Israel, LST geospatial rasters correspond to the MODIS tiles h20v05, h20v06, h21v05, and h21v06. Cloud cover makes MODIS collections susceptible to data gaps. Additionally, LST is captured by the Spinning Enhanced Visible and InfraRed Imager (SEVIRI), which offers high temporal resolution (15 minutes) but lower spatial resolution (3 km). SEVIRI data for this study was downloaded from LSA SAF collections (accessed December 21, 2021) and aggregated on a daily basis following the methodology outlined by Zhou et al. (2020). To produce estimates at higher spatial resolution (250 m), we used Landsat 8-9 Collection 2 LST data, which were averaged monthly over 18-year period composites (2004–2022) to capture fine-scale spatial variations of Ta (see Table 1).

2.2.3 ERA5 Reanalysis data

We used atmospheric reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF) to retrieve other temperature datasets. This data is generated using methods similar to those in traditional forecasting systems, incorporating data assimilation techniques through Integrated Forecasting System (IFS). In brief, available spaceborne data and historical in-situ measurements are combined and modeled to produce climate estimates. Rigorous quality control is applied to ensure reliable results. In terms of spatial and temporal resolution, ERA5 offers hourly estimates with a spatial resolution of approximately 10 km, spanning from 1979 to the present. We extracted covariates such as skin temperature and 2-meter temperature from this repository. ERA5 data for the period 2004–2022 at a 0.125° (~10 km) resolution was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store (https://cds.climate.copernicus.eu, accessed December 21, 2021). The downloaded datasets contained less than 5% missing values, which we addressed using simple mean imputation for gap filling.

Table 1. A summary of the time-invariant and time-variant covariates employed in this study

Name	Source	Temporal resolution	Spatial resolution	Units	Nature	Stages use
Elevation	EU-DEM (Copernicus)	-	25m	m	Time invariant	1-5
Population density	ICBS	-	Point-data	people /km²	Time invariant	1-5
MODIS LST Aqua	Nasa	12 hours	1km	Kelvin	Time variant	1-4
MODIS LST Terra	Nasa	12 hours	1km	Kelvin	Time variant	1-4
2 meters air temperature	Reanalysis (ECMWF)	Hourly	~10km	Kelvin	Time variant	1-4
Skin temperature	Reanalysis (ECMWF)	Hourly	~10km	Kelvin	Time variant	1-4
LST SEVIRI	MVIRI - SEVIRI	Every 15mins	3km	°C	Time variant	1-4
Landsat 8,9 collection 2 LST	TIRS Nasa	16 days	100m	Kelvin	Time variant	5

2.2.4 Elevation

Elevation data was obtained from version 1.1 of the European Digital Elevation Model (EU-DEM) provided by the Copernicus Land Monitoring Service. This dataset has a spatial resolution of 25 meters and a vertical root mean square error (RMSE) of ±7 meters (Tøttrup, 2014).

2.2.5 Population density

We included population density data from the Israeli Central Bureau of Statistics (ICBS, 2014). To align with the modeling grids, population values were aggregated by calculating the population-weighted average for each 1 km grid cell.

2.3 Statistical methods

To model Ta, we developed a five-stage machine learning (ML) approach that integrates methodologies from previous studies (Zhou et al., 2020; Hough et al., 2020; Kloog et al., 2017; Kloog et al., 2014) with a novel feature engineering (FE) technique (see Figure 3). This approach generates geographically weighted predictors for MODIS LST and monitored Ta, allowing us to capture spatial and temporal patterns while preserving the positional integrity of specific geographic points and their nearest neighbors—relationships often lost when transforming satellite images into data tables for training and randomization.

Incorporating these new predictors significantly improves the accuracy of 1 km Ta models across Israel. To summarize, stages 1 through 4 focused on predicting Ta at 1 km resolution. Once this was achieved, we proceeded to stage 5, where we downscaled Ta to 250 m resolution. In this stage, as described by Hough et al. (2020), residuals from the 1 km Ta models—as the difference between observed Ta values at monitoring stations and the 1 km model predictions—are mapped onto a 250 m grid. These residuals are then used to train an XGBoost model, capturing their full spatio-temporal variation at a finer resolution. Finally, we added the 250 m predicted residuals to the 1 km Ta estimates to produce the final 250 m Ta surfaces.

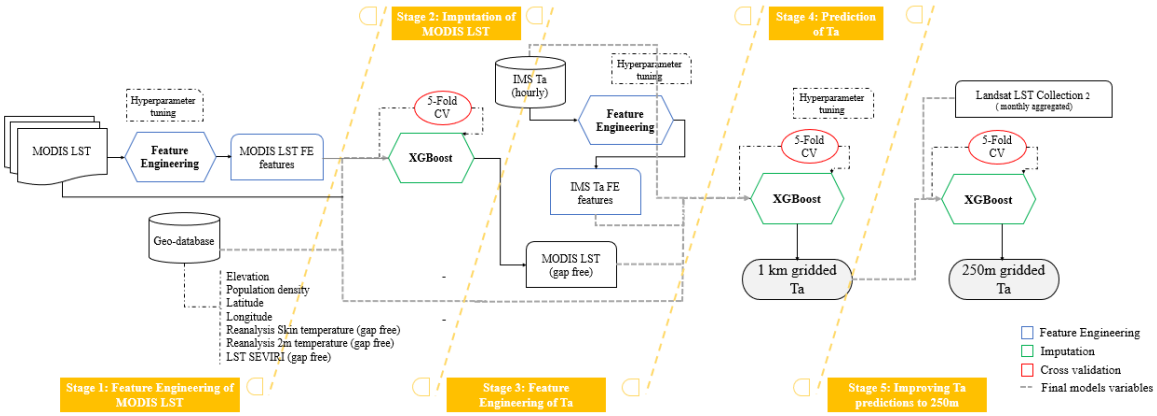


Fig 3. Schematic diagram of the five-stage model for estimating air temperature (Ta)

Modelling Ta at 1km

Stage 1 model: Feature Engineering of MODIS LST

In Stage 1, we perform a gap-filling process for MODIS LST aqua variables (MODIS LST Night Aqua and MODIS LST Day Aqua) and MODIS LST mean (the average of all four MODIS overpasses). To accomplish this, we create new features using a strategy based on the K-Nearest Neighbors (KNN) model, a non-parametric supervised machine learning algorithm. The KNN model stores all available data and predicts new cases by comparing them to existing data points using a selected similarity metric, such as Euclidean distance, Manhattan distance, or Mahalanobis distance (Swathi et al., 2022). We start with a matrix containing P cells that represent Israel. For each cell (p) in P, we identify a set of neighboring cells (i, j, g, …l in P | i, j, g, …l ≠ p) that share for instance similar latitude and longitude values. Similarity in this algorithm is determined using distance functions based solely on geographical attributes, such as longitude (x_p), latitude (y_p), and elevation (E_p) (see Equation 1). Distances between cells are given by a distance matrix D. Once we identify a set of k neighbors (kN_p) for a given point p, we calculate the average LST values for these neighbors at each date t, creating a new feature that estimates LST for different locations (p) and periods of time (t).

This process was conducted iteratively by varying the number of neighbors (k) to generate several new features (FEs) (see Appendix 2 for further details on the FE generation). To finalize our selection, we identified the optimal feature by tuning the parameter k. The

selection criterion focused on minimizing RMSE by adding each engineered variable (FE) into a regression model of the MODIS LST variable, one at a time. We then identified the feature with the optimal k (denoted as FE(b), where $k = b$) as the one where no significant improvement in RMSE was observed when increasing k from b to $b + a$ (with $a = 1, 2, 3, 4, \ldots$). This feature, FE($b + a$), was then used to regress the MODIS LST variable, and no significant improvement in the RMSE metric was observed compared to the results given with FE(b).

We defined this methodology for tuning the parameter k as the "cut-off point rule".

x_p : Longitude of cell p ∈ P, y_p : Latitude of cell p ∈ P, E_p : Elevation of cell p ∈ P

D_{pj} : Distance from cell p ∈ P to cell j ∈ P.

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2 + (E_p - E_j)^2} \quad \forall p, j \in P \qquad \text{Eq. 1}$$

$$D = \begin{vmatrix} 0 & D_{1,2} & \ldots & D_{1,|P|} \\ D_{2,1} & 0 & \ldots & \ldots \\ \ldots & \ldots & 0 & D_{|P|-1,|P|} \\ D_{|P|,1} & \ldots & D_{|P|,|P|-1} & 0 \end{vmatrix}$$

Hyperparameter tuning was conducted to select the FEs to include in our models. For each MODIS LST variable, we extracted a total of 20 features associated with different values of the parameter k (see equation 2). We considered only latitude and longitude coordinates as dimensions for the distance matrix D (see Figure 4). We adopted a parsimonious approach, using XGBoost models. Date variables were binarized, including the calendar month (e.g., Jan, Feb, Mar) and weekdays (e.g., Mon, Tues, Wed).

$$\text{LST nearest neighbors } k \text{ mean}_{p,t} = \sum_{j \in kN_p}^{k=20} \text{LST}_{j,t}/k \quad \forall p \in P, t \in T \qquad \text{Eq. 2}$$

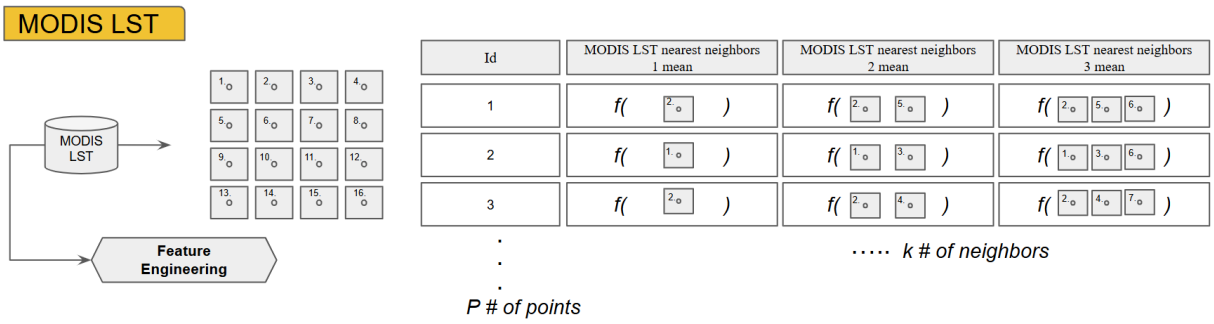


Fig 4. A simplified example for a specific day is provided, illustrating how distance is quantified using latitude and longitude geographical coordinates when calculating FEs for MODIS LST variables.

Stage 2 model: Imputation of MODIS LST

In Stage 2, we filled gaps in MODIS LST variables by training XGBoost models using the following predictors: latitude, longitude, elevation, population density, binarized date variables, LST SEVIRI, reanalysis 2-meter temperature, skin temperature, and the selected FEs from Stage 1. Equation 3 summarizes our final Stage 2 model, which includes the optimized selected FE feature related to the latitude and longitude geographical coordinates used for the distance matrix D.

$$\text{LST} \sim f(\text{lat, long, elevation, population density, Reanalysis skin temperature, Reanalysis 2m temperature, LST SEVIRI, Monday, Tuesday, Wednesday ..., January, Febrary, March, ... FE MODIS LST}) \quad \text{Eq. 3}$$

Stage 3 model: Feature Engineering of T_a

To generate daily estimations of Ta, we implemented a new FE process that restricted FE creation to monitor locations. It is worth noting the change in dimensions of the distance matrix D compared to the FE process in Stage 1. In this stage, the distance matrix D undergoes a dimension reduction from (P, P) to (P, s) where P represents the total number of grid cells describing Israel, and s denotes the total number of active allocated monitors (see Appendix 3 for a detailed explanation of the modifications in the distance matrix D resulting from the application of FE to Ta). During this stage, we calculated FEs regarding latitude and longitude geographical coordinates. Additionally, we extended the FE process to the elevation dimension to generate another set of FE features and further enhance model performance. Finally, we applied our cut-off point rule strategy to select the best FE features per dimension.

We incorporated inverse distance weighting (IDW) within the FE approach at this stage, assigning more weight to monitor values closer to points of interest when deriving FEs (see Figure 5). The cut-off point rule was also applied during this step. Figure 6 illustrates the logic behind the IDW approach. The weight W_{1,S_2} places greater emphasis on the Ta measurements recorded by station S_2, as this monitor is the closest to point $p = 1$. This logic changes for each cell p in P because the distances between cells and monitors differ.

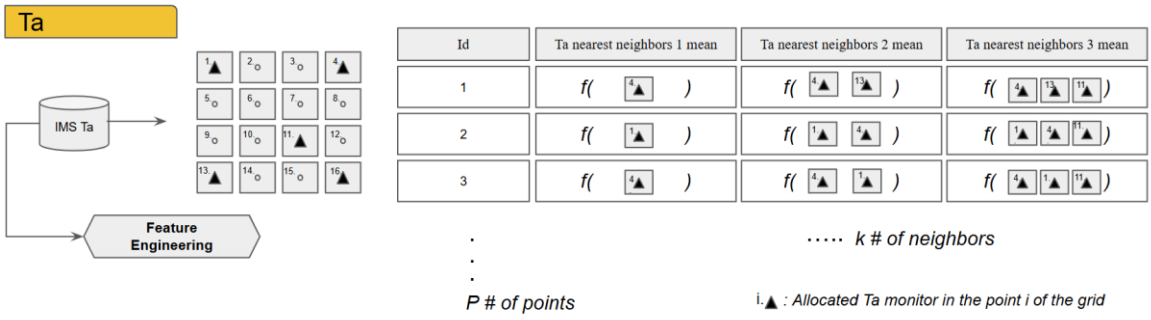


Fig 5. Explanation of the Ta FE calculation. This minimalistic example illustrates the process for a specific day, using latitude and longitude geographical coordinates to quantify distance in the distance matrix *D*.

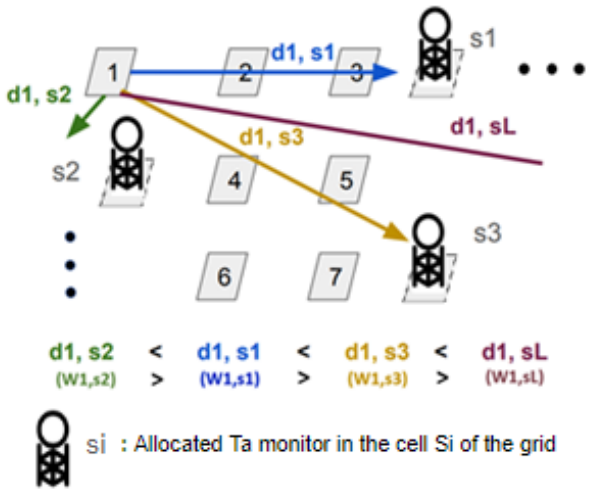


Fig 6. Logic behind the implementation of inverse distance weighting (IDW) for the FE calculation of Ta.

We trained an XGBoost model for each *i-th* value of k. Our final model for this stage is summarized by the following expression (Equation 4):

$$Ta \sim f(\text{lat, long, elevation, population density, Reanalysis skin temperature, Reanalysis 2m temperature,}$$
$$\text{LST SEVIRI, Monday, Tuesday, Wednesday ..., January, Febrary, March, ...}$$
$$\text{gap} - \text{free MODIS LST, } Ta \text{ FE}) \qquad \text{Eq. 4}$$

Stage 4 model: prediction of T_a

In Stage 4, we predicted Ta (T_{min}, T_{mean}, and T_{max}) using latitude, longitude, elevation, population density, binarized date variables, LST SEVIRI, reanalysis 2-meter temperature, skin temperature, gap-free MODIS LST, and Ta FE variables selected in Stage 3. Equation 5 summarizes the mathematical formulation for T_{mean} as an example

$$Ta \sim f(\text{lat, long, elevation, population density, Reanalysis skin temperature, Reanalysis 2m temperature, LST SEVIRI, Monday, Tuesday, Wednesday ..., January, Febrary, March, ... MODIS LST , selected } Ta \text{ FE in lat and long dimensions, selected } Ta \text{ FE in elevation dimension}) \quad \text{Eq. 5}$$

Stage 5: improving Ta predictions to 250 m

In the final stage (Stage 5), we improved the resolution of Ta to 250 m to better capture intra-urban variability. We calculated residuals at monitor locations by subtracting the 1km predicted Ta from the observed values. These residuals were then used to train an XGBoost model to predict across the entire 250 m grid. Our models incorporated 250 m latitude and longitude coordinates, elevation, population density, month, monthly averaged Landsat LST Collection 2 (products 8 and 9) from 2004 to 2022, and the three predicted 1km Ta values (T_{min}, T_{mean}, T_{max}), as described in Equations 6 and 7.

$$\text{Residuals } Ta \text{ mean at 250 m} \sim \text{XGBoost(monthly averaged Landsat LST 250 m, latitude, longitude, elevation, population density, month, } Ta \text{ min at 1km , } Ta \text{ mean at 1km, } Ta \text{ max at 1 km)} \quad \text{Eq. 6}$$

$$Ta \text{ mean at 250 m} = Ta \text{ mean at 1 km} + \text{Residuals } Ta \text{ mean at 250 m} \quad \text{Eq. 7}$$

Validation

Our results were cross-validated using a standard 5-fold cross-validation approach for all stages of the model. To address potential spatial bias in the 1 km Ta estimates—particularly the risk of overfitting due to the proximity of monitoring stations—we implemented spatial blocking. Specifically, we excluded from training data eight sparsely distributed monitors with few or no neighboring stations (Figure 7). We then predicted values for these excluded locations and recalculated the RMSEs. This evaluation provided a more robust assessment of the model's performance in regions with limited monitoring coverage.

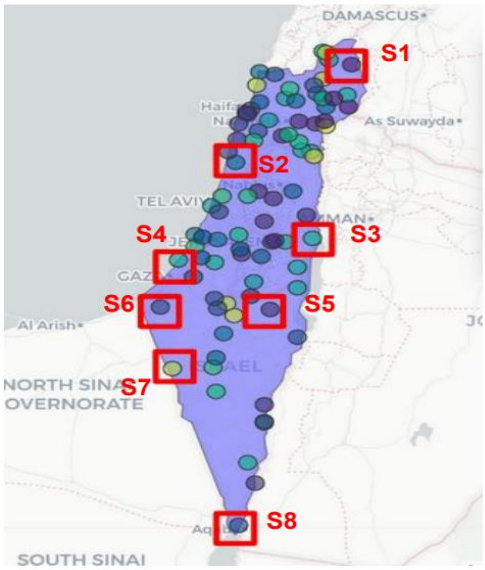


Fig 7. Monitors held out for the spatial blocking cross-validation. The selected monitors are enclosed within red squares.

3. Results

3.1 Correlation heatmap

Figure 8 displays the correlations calculated between all spatiotemporal covariates used in this study.

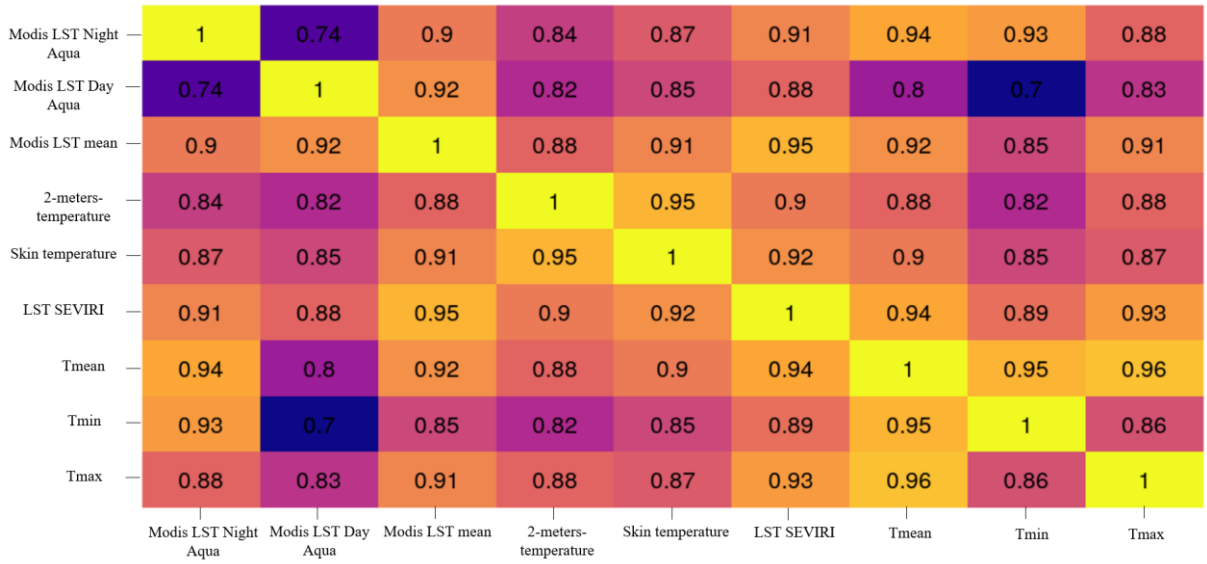


Fig 8. Correlation heatmap of all spatiotemporal predictors, and Ta for 2022 year

3.2 Tuning *k* parameter in stage 1 model

We began by working with the MODIS LST Night Aqua variable, with parameter k varying from 1 to 20 (Figure 9). At first, we observed an RMSE of approximately 2°C for a model trained without any FE feature. Subsequently, the RMSE decreased from 2°C to nearly 0.8°C after adding the first calculated FE (k = 1). After completing additional iterations, we identified the optimal cut-off point at k = 4, which further reduced the RMSE to 0.35°C. For the other two MODIS LST variables—MODIS LST Day Aqua and MODIS LST Mean—the optimal cut-off point was also k = 4, resulting in RMSEs of 0.51°C and 0.99°C, respectively. These results were for 2022. We applied a similar approach for the other years in the study.

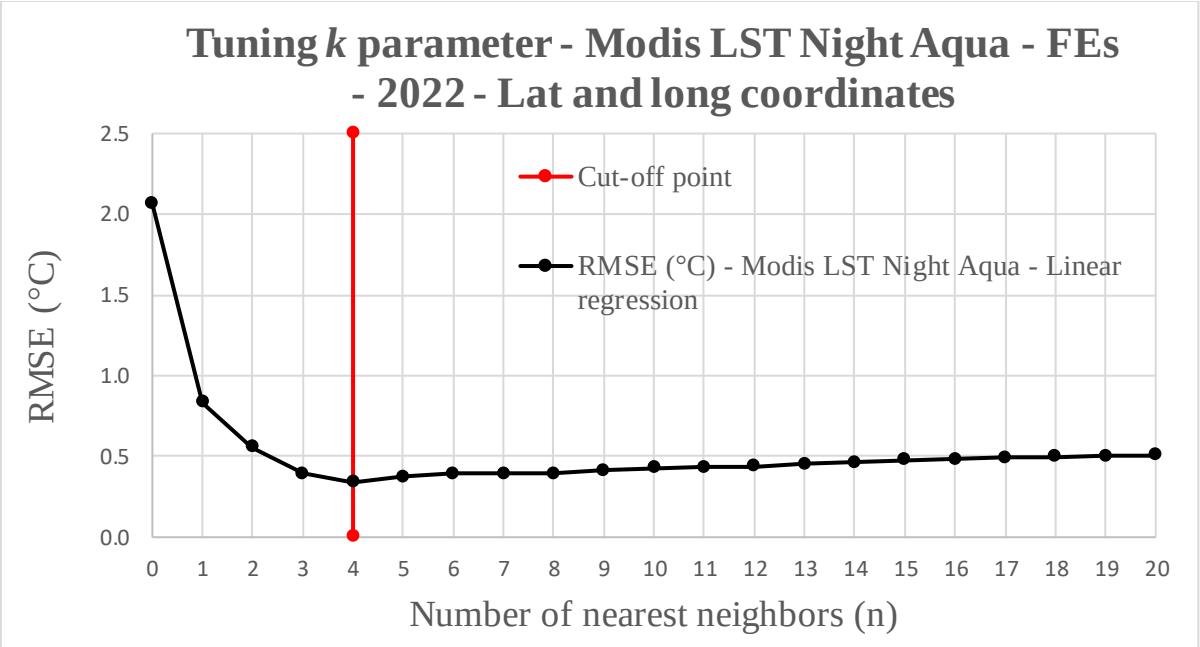


Fig 9. FE selection for the model that imputes the MODIS LST Night Aqua variable, year 2022.

3.3 Performance of the stage 2 model

Table 2 presents the 5-fold cross-validation performance of the stage 2 models across all years, incorporating the selected FE features from the previous stage. The average RMSEs over the years were 0.35°C for MODIS LST Night Aqua, 0.58°C for MODIS LST Day Aqua, and 1.02°C for MODIS LST Mean. As an example, we present the results for gap-filling of MODIS LST Night Aqua on May 05, 2016, where the percentage of missing values was approximately 20% (see Figure 10).

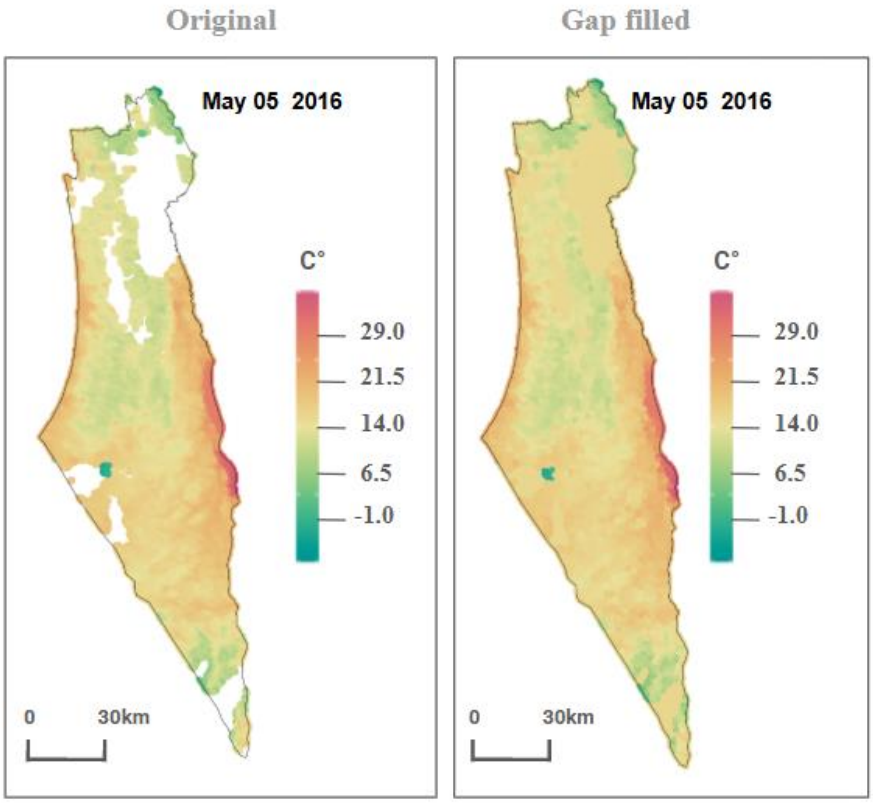


Fig 10. Gap Filling of MODIS LST Night Aqua for May 05, 2016.

Table 2. Stage 2 model performance of MODIS LST variables, 5-fold cross-validation have been used across all years (2004–2022). RMSE metrics were obtained from XGBoost models (trees = 85, learning rate = 0.01, max depth =7).

	MODIS LST Night Aqua	MODIS LST Day Aqua	MODIS LST mean
Year	RMSE (°C)	RMSE (°C)	RMSE (°C)
2004	0.214	0.587	1.020
2005	0.363	0.585	1.018
2006	0.372	0.584	1.015
2007	0.355	0.578	1.005
2008	0.371	0.590	1.048
2009	0.370	0.602	1.042
2010	0.377	0.657	1.009
2011	0.357	0.595	1.047
2012	0.353	0.591	0.966
2013	0.355	0.578	1.042

2014	0.341	0.605	1.024
2015	0.367	0.594	1.020
2016	0.351	0.581	0.965
2017	0.348	0.573	1.032
2018	0.333	0.608	1.015
2019	0.349	0.594	1.019
2020	0.351	0.585	1.035
2021	0.359	0.506	1.084
2022	0.364	0.574	0.987

3.2 Tuning *k* parameter in stage 3 model

To determine an optimal value for *k* and select the final FE features for Ta models, we generated 20 features with and without Inverse Distance Weighting (IDW). Our selection was performed based on our predefined cut-off point rule. We observed significant fluctuations in RMSE scores (increments and decrements) when IDW was not applied, which highlights the importance of its implementation to ease the feature selection.

The tuning process resulted in $k = 5$ for the FE features corresponding to Ta variables (T_{min}, T_{mean}, T_{max}). These results were consistent across the years 2022 and 2015 when using Latitude and Longitude as inputs for the distance matrix *D*. This consistency led us to conclude that annual optimization of this parameter is unnecessary. Figure 11 illustrates the tuning results for T_{mean} in 2022, where latitude and longitude were used as the geographical dimensions for computing *D*. Tuning results for the T_{mean} model in 2015 are provided in Annex A (Figure A.1)

Similarly, we found that $k = 10$ was the optimal value for the *k* parameter of the FE features in the Ta variables (T_{min}, T_{mean}, T_{max}) when the distance matrix *D* was computed using only the Elevation dimension (see Figure 12).

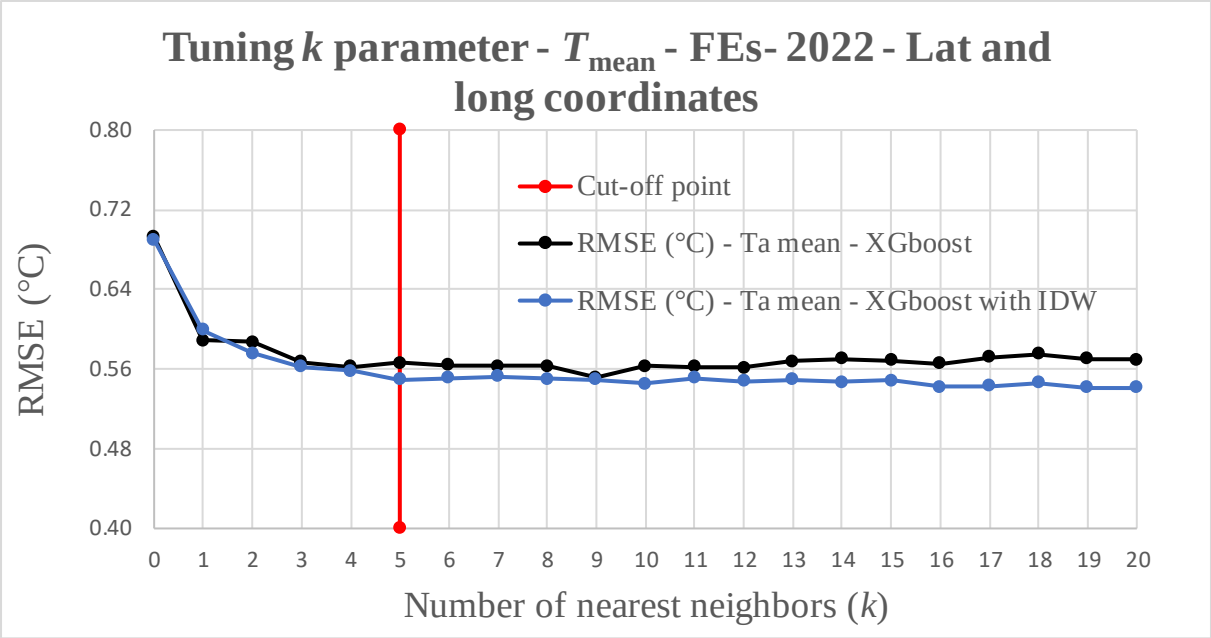


Fig 11. Selection of the best FE (using latitude and longitude dimensions in *D*) for the model that predicts $T_{mean,}$ year 2022.

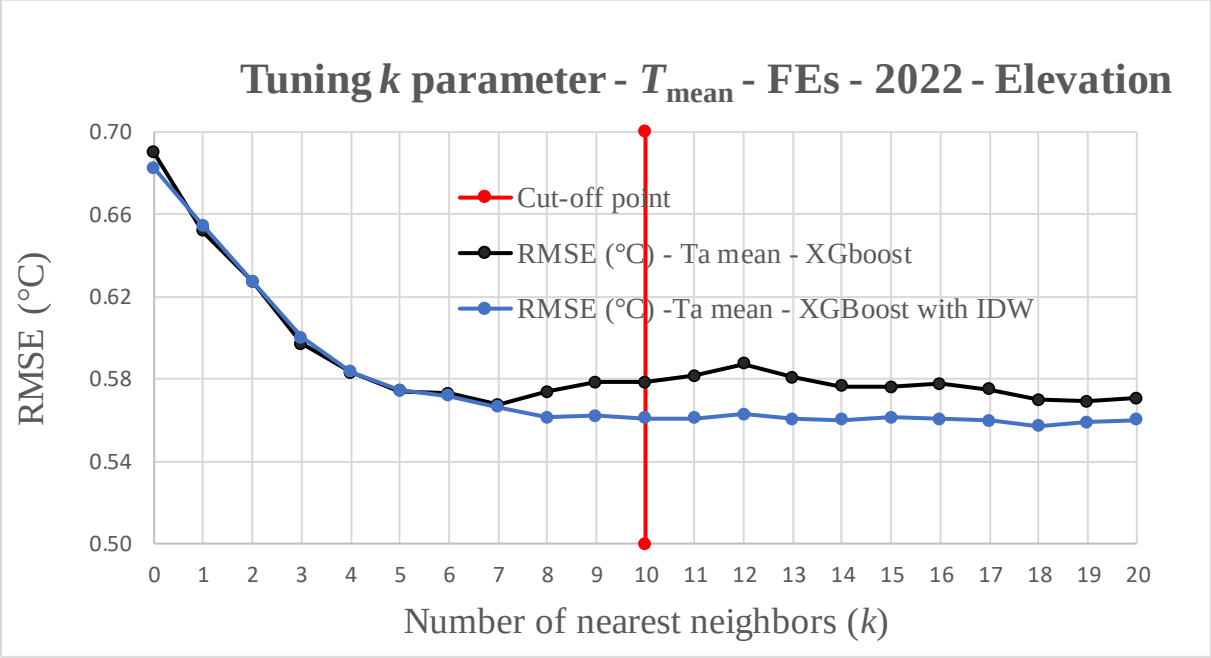


Fig 12. Selection of the best FE (using elevation dimension in *D*) for the model that predicts $T_{mean,}$ year 2022.

3.3 Performance of the stage 4 model

Table 3 presents the overall results for the 1 km Ta models across the years. In general, the models performed well, with average RMSE values of 0.61°C for T_{mean}, 0.92°C for T_{min}, and 0.73°C for T_{max} (based on results from 2004 to 2022). To assess spatial bias in the results, we spatially blocked the data of 8 monitors at training stages, applying this strategy on years 2005, 2015, and 2022 (see Table 4). The validation approach yielded good performance on the 8 held-out monitors. The RMSE of T_{mean} increased slightly under this validation approach, from 0.72°C (average for all years with normal cross-validation) to 1.07°C (average for 2022, 2015, and 2005 when spatially blocking). The following tables present our findings for T_{mean}, with held-out monitors shown in Figure 7.

Table 3. Stage 4 model (predicting daily 1 km Ta): 5-fold cross-validated performance across all years (2004–2022), with RMSE values from XGBoost models (trees = 1250, learning rate = 0.01, max depth = 12).

	T_{min}	T_{mean}	T_{max}
Year	RMSE (°C)	RMSE (°C)	RMSE (°C)
2004	1.011	0.680	0.812
2005	0.982	0.641	0.761
2006	0.977	0.594	0.766
2007	0.900	0.537	0.747
2008	0.966	0.639	0.740
2009	0.919	0.630	0.722
2010	0.962	0.689	0.764
2011	0.862	0.566	0.681
2012	0.925	0.618	0.719
2013	0.948	0.635	0.748
2014	0.906	0.607	0.715
2015	0.915	0.622	0.717
2016	0.941	0.625	0.702
2017	0.933	0.634	0.726
2018	0.912	0.629	0.711

2019	0.918	0.632	0.724
2020	0.916	0.603	0.746
2021	0.882	0.617	0.720
2022	0.884	0.539	0.711

Table 4. Cross-validated results for Tmean from the 8 spatially blocked monitoring stations in 2022, 2015, and 2005.

	2022	2015	2005
T_a monitor	Results held out monitors – RMSE (°C)		
S1	1.465	0.976	1.310
S2	0.921	1.014	0.780
S3	0.557	0.935	0.701
S4	1.449	2.050	1.705
S5	0.875	-	-
S6	0.799	1.235	1.110
S7	0.948	0.965	0.961
S8	0.782	1.050	0.904
Average RMSE (Held out monitors)	0.975	1.175	1.067

On December 12, 2022 (see Figure 13), estimates from the 1 km Ta models ranged from 5°C in both the northern and southern parts of the country—near Lake Tiberias in the north and the Negev Desert in the south—to nearly 26°C along the border with the Mediterranean Sea, and in other regions such as the far south, where Eilat and Yotvata are located. Regarding T_{min}, notable metropolitan areas such as Tel Aviv, Jerusalem, and Haifa were identified as hotspots in the west, while Beersheba, Hebron, the West Bank, and the Dead Sea remained among the warmest zones in the east. When examining T_{max} (warmest daily temperatures), the areas bordering Lebanon remained cooler, with temperatures ranging from 11°C to 17°C. In contrast, the hottest temperatures were observed around the Gaza Strip, Jerusalem in the center, Netanya in the northern-central region, and Eilat in the far south.

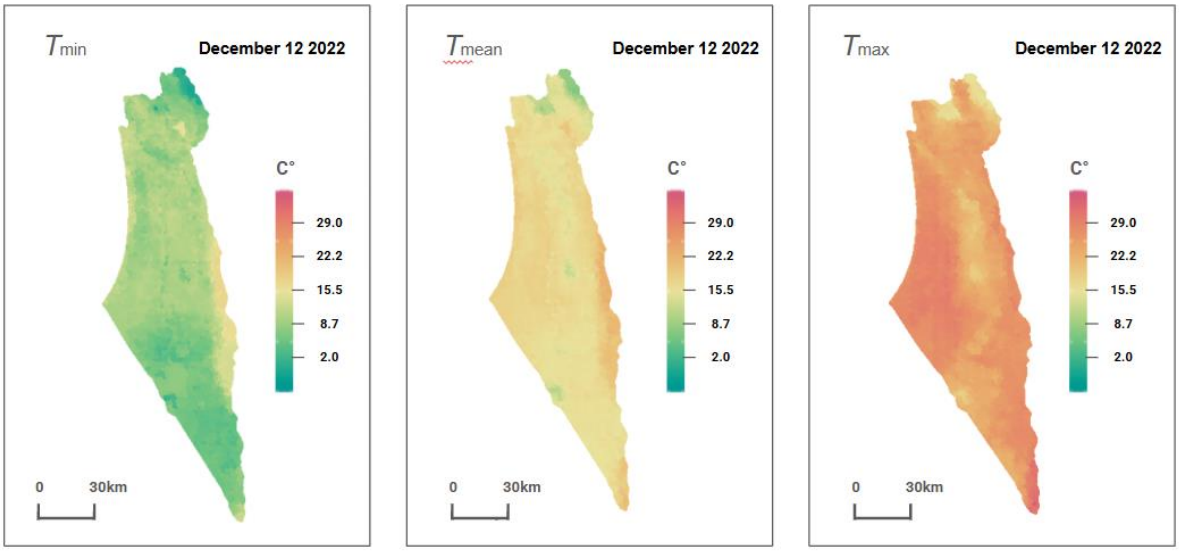


Fig 13. Predicted 1 km Ta from the Stage 4 model for December 12, 2022.

3.3 Performance of the stage 5 model

Table 5 presents the overall results for the 250 m models estimating Ta residuals across all years of the study. In summary, these models performed excellently, with average RMSE values of 0.32°C for R_{Tmean}, 0.40°C for R_{Tmax}, and 0.58°C for R_{Tmin} on the validation sets.

Table 5. RMSEs of Ta residuals at a 250 m resolution, calculated using 5-fold cross-validation across all years (2004–2022). The analysis employed XGBoost models with the following parameters (trees = 1250, learning rate = 0.01, max depth = 12).

	R_{Tmean}	R_{Tmin}	R_{Tmax}
Year	RMSE (°C)	RMSE (°C)	RMSE (°C)
2004	0.329	0.556	0.432
2005	0.338	0.558	0.420
2006	0.317	0.570	0.425
2007	0.306	0.571	0.411
2008	0.320	0.581	0.411
2009	0.330	0.583	0.401
2010	0.330	0.581	0.399
2011	0.325	0.592	0.398
2012	0.325	0.592	0.400
2013	0.325	0.588	0.412

2014	0.325	0.600	0.410
2015	0.329	0.589	0.399
2016	0.325	0.586	0.400
2017	0.317	0.591	0.400
2018	0.319	0.591	0.401
2019	0.319	0.589	0.392
2020	0.319	0.586	0.402
2021	0.313	0.579	0.390
2022	0.292	0.553	0.369

4. Discussion

Our modeling demonstrated robust performance at both 1 km and 250 m resolutions for estimating Ta. Over the period from 2004 to 2022, the models achieved average RMSE scores of 0.92°C, 0.61°C, and 0.73°C for daily T_{min}, T_{mean}, and T_{max}, respectively. A major advancement of this study is the achievement of a very fine spatial resolution of 250 m for Ta in Israel, compared to previous studies that have primarily focused on 1 km scales (Zhou et al., 2020; Kloog et al., 2012, 2014, 2017). At the 250 m resolution, our models achieved RMSE values of 0.58°C, 0.31°C, and 0.40°C for daily Ta residuals; R_{Tmin}, R_{Tmean}, and R_{Tmax}, respectively (2004–2022). This improvement in spatial resolution is critical for more accurate temperature exposure assessments, which are essential for better understanding the impacts of temperature on human health.

Using our new methodology to predict Ta at 1 km and 250 m resolutions in Israel, our results outperformed recent studies conducted in the same area (Rosenfeld et al., 2017; Zhou et al., 2020). For example, we achieved a 61.4% improvement in T_{mean} RMSE (0.61°C in our study compared to 1.58°C in Zhou et al., 2020). The performance of our models is comparable to those from other studies using coarser 1 km resolutions, such as in France (Ta RMSE of 1.3–1.9°C, Hough et al., 2020), in the Netherlands (Ta RMSE < 1°C reported by Dirksen et al., 2020 using linear regression and multiple adaptive regression splines), and in China (Wang et al., 2022) predicting daily T_{max} at a 0.1° resolution (approx. 11 km) with an RMSE of 1.52°C. Additionally, our results align with high-resolution models predicting Ta residuals at scales finer than 1 × 1 km. For instance, Hough et al. (2020) reported RMSE values of 0.4–0.6°C for R_{Tmin}, R_{Tmean}, and R_{Tmax} at 200 m resolution in France, while Flückiger et al. (2022) achieved RMSE values of 1.05–1.86°C for R_{Tmin}, R_{Tmean}, and R_{Tmax} at 100 m resolution in Switzerland.

To ensure the robustness of our findings, we extended the conventional cross-validation approach by spatially blocking the data of some monitors. This strategy excluded eight predefined monitors' areas during training, predicted Ta over them, and averaged the RMSE results to assess spatial bias. This method demonstrated excellent performance, with RMSE values ranging from 0.97°C to 1.17°C for T_{mean}, highlighting our model's ability to explain Ta variance. Comparable studies using similar validation strategies, such as those conducted in the U.S. (Li et al., 2018), reported higher RMSE values of 1.9°C and 2.1°C for T_{min} and T_{max}, respectively, when predicting over US urban surrounding areas.

To the best of our knowledge, this is the first study applying Feature Engineering to enhance the performance of exposure models for Ta estimation. In general, few studies within the fields of epidemiology and environmental exposure have employed this type of approach, leaving significant room for future advancements that integrate FE with emerging ML methods, such as artificial neural networks. For instance, Choe and Yom (2020) utilized FE to improve LST predictions in the Seoul Metropolitan area of Korea, achieving predictions at 30 m resolution with RMSE values ranging from 1.4 to 3.6°C using deep learning. Similarly, Kartal et al. (2022) aimed to predict one-step-forward MODIS LST in Turkey (specifically in the northern part of Adana province, in south-central Turkey) using deep-learning convolutions—a type of FE as defined by Tymoshenko et al. (2016)—combined with interpolation techniques, reporting RMSE values between 2.85 and 3.62°C.

This methodology provides an innovative and adaptable approach for constructing new independent predictors (FEs) by using the KNN model. Experts have the flexibility to apply their own criteria when selecting dimensions in the distance matrix D, at FE stages, to improve model performance. While our model supports the integration of multiple spatial dimensions to improve interpretability, for the Ta case, we limited our focus to a few key spatial factors, including latitude, longitude, and elevation. Nonetheless, many additional dimensions could be explored to further enhance the model's explainability. We encourage researchers studying other pollutants (e.g., PM, CO_2, or NO_2) to adopt our FE approach and develop new predictors based on factors such as traffic congestion, population density, NDVI, and proximity to watercourses. We anticipate that incorporating these predictors could result in significant improvements for their models.

5. Conclusion

This study presents a novel five-stage ML methodology for estimating Ta using satellite data from various sources. By leveraging advancements in computational power, we enhanced traditional artificial intelligence algorithms through the integration of FE techniques, specifically utilizing distance matrix functions and the KNN model. Our approach enables high-resolution Ta estimates in both spatial and temporal dimensions. Furthermore, we provide a dataset explaining Ta variability through minimal, mean, and maximal predictions covering all the territory of Israel. The methodology includes detailed steps for deriving (FEs) features from collected data, what could be adaptable to other research areas and objectives. Our models demonstrated strong performance based on RMSE evaluation metrics over a yearly timeframe, achieving predictions at an impressive spatial resolution of 250 m. This innovative approach harnesses globally accessible geospatial datasets, allowing for broad applicability. While our study focused on Israel, the methodology can be easily replicated in other regions with access to comparable geospatial datasets. However, we acknowledge that researchers may face challenges in reproducing these results without adequate computational resources, particularly in terms of hardware memory. We believe our findings will contribute to reducing exposure misclassification in epidemiological studies, thereby minimizing bias and strengthening the reliability of scientific hypotheses. We encourage the scientific community to adopt this approach for more efficient and robust exposure modeling.

Appendix 1

Table A1. (Appendix 1) Results for the number of Ta monitors allocated in Israel from 2004 to 2022. Data were retrieved from IMS.

Year	Number of T_a monitors
2004	50
2005	59
2006	61
2007	69
2008	76
2009	78
2010	84
2011	85

2012	85
2013	83
2014	84
2015	82
2016	81
2017	80
2018	79
2019	80
2020	80
2021	84
2022	85

Appendix 2

Distance is calculated with respect to stationary variables (l_1, l_2, l_3, etc.); for example, $l_i = x_p$ (longitude), $l_i = y_p$ (latitude), or $l_i = E_p$ (elevation). Once distances between points are computed, sets of nearest neighbors are determined by selecting the column arguments with the smallest scores.

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2} \text{ or } D_{pj} = \sqrt{(E_p - E_j)^2}, \text{or even}$$

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2 + (E_p - E_j)^2} \quad \forall p, j \in P \qquad \text{Eq. 1}$$

kN_p : Set of first k nearest neighbors points of the point $p \in P$.

$$kN_p = [N_1 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}), N_2 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}) \mid D_{p,N1} \notin \{D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}\}, N_3 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}) \mid D_{p,N1}, D_{p,N2} \notin \{D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}\} \ldots, N_k = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}) \mid D_{p,N1}, D_{p,N2}, D_{p,N3} \ldots, D_{p,NK-1} \notin \{D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}\}]$$

Eq. A2.1

To calculate one of the MODIS LST FEs, we used the following expression:

$$\text{LST nearest neighbors } k \text{ mean}_{p,t} = \sum_{j \in kN_p}^{k} LST_{j,t}/k \quad \forall p \in P, t \in T \qquad \text{Eq. A2.2}$$

Appendix 3

$P_{monitors}$: Set of points with monitors allocated in Israel.

S_i : Denotes a monitor i $\in P_{\text{monitors}}$; $P_{\text{monitors}} = \{S_1, S_2, S_3, .. S_s\}$; $P_{\text{monitors}} \in$ P

$$D = \begin{vmatrix} D_{1,S_1} & D_{1,S_2} & \dots & D_{1,S_s} \\ D_{2,S_1} & D_{2,S_2} & \dots & D_{2,S_s} \\ \dots & \dots & \dots & D_{|P|-1,S_s} \\ D_{|P|,S_1} & D_{|P|,S_2} & D_{|P|,S_s-1} & D_{|P|,S_s} \end{vmatrix}$$

kS_p : Set of first k monitors to the point p ∈ P.

$$k < s$$

$$Ta \text{ nearest neighbors } k \text{ mean}_{\text{p,t}} = \sum_{\text{j} \in kS_p}^{k} Ta_{\text{j,t}}/k \; \forall \text{p} \in \text{P}, \text{t} \in \text{T}$$

Eq. A3.1

Annex A

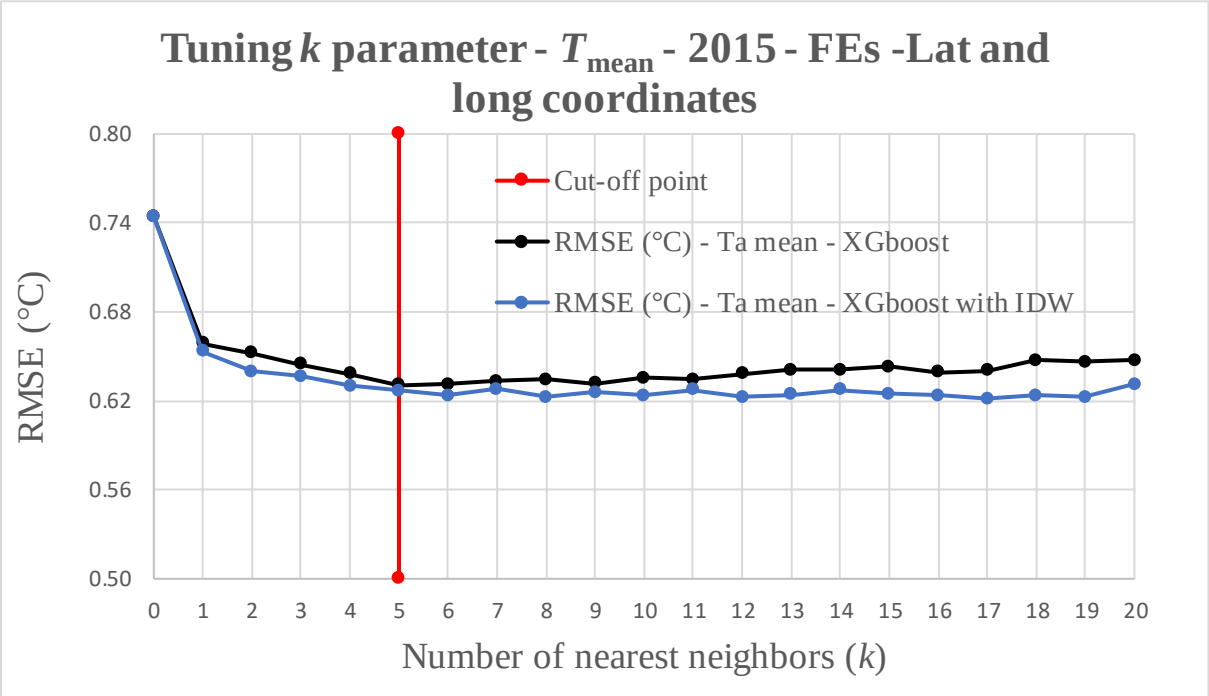


Fig A.1 (Annex A) Selection of the best T_{mean} FE feature (latitude and longitude coordinates in distance matrix *D*) based on 2015 data. Identical results (*k*=5) were obtained for the 2022 dataset.

ACKNOWLEDGEMENTS

This work was funded by the U.S.-Israel Binational Science Foundation (grant number 2021197) and the French National Research Agency (ANR-15-IDEX-02). Juan David Briceno is supported by a French public grant under the "Investissements d'avenir" program and a scholarship from the Kreitman School for Advanced Graduate Studies at Ben-Gurion University of the Negev. We would like to thank the EUMETSAT Satellite Application Facility on Land Surface Analysis (LSA SAF) for providing SEVIRI LST data, NASA LP DAAC for MODIS LST data, and IMS for remotely sensed weather data.

AUTHOR CONTRIBUTIONS

Juan David Briceno: Data curation; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing.

Itai Kloog: Formal analysis; Methodology; writing – review and editing.

Johanna Lepeule: Formal analysis; methodology; writing – review and editing.

Guillaume Barbalat: writing – Review and editing.

Elena Colicino: Writing – Review and editing.

Massimo Stafoggia: Methodology; conceptualization; writing – review and editing.

Michael Dorman: Data curation; visualisation.

Data Availability Statement

The data supporting the findings of this study are not publicly available due to confidentiality restrictions. Access to the data can be requested from the IAB (Institut pour l'Avancée des Biosciences), UGA, where the data is managed.

Funding information

U.S.-Israel Binational Science Foundation (BSF), Grant/Award Numbers: 2021197; and French National Research Agency in the framework of the "Investissements d'avenir" program, Grant/Award Numbers: ANR-15-IDEX-02 IRGA 2021 PAPex project.

ORCID

Juan David Briceno https://orcid.org/0009-0006-3771-528X

References

Hough, I., Just, A. C., Zhou, B., Dorman, M., Lepeule, J., & Kloog, I. (2020). A multi-resolution air temperature model for France from MODIS and Landsat thermal data. *Environmental Research, 183*, 1–11. https://doi.org/10.1016/j.envres.2020.109244.

Zhou, B., Erell, E., Hough, I., Rosenblatt, J., Just, A. C., Novack, V., & Kloog, I. (2020). Estimating near-surface air temperature across Israel using a machine learning-based hybrid approach. *International Journal of Climatology, 40*, 6106–6121. https://doi.org/10.1002/joc.6570.

Basu, R., Malig, B., & Ostro, B. (2010). High ambient temperature and the risk of preterm delivery. *American Journal of Epidemiology, 172*(10), 1108–1117. https://doi.org/10.1093/aje/kwq170.

Kloog, I., Chudnovsky, A., Koutrakis, P., & Schwartz, J. (2012). Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. *Science of the Total Environment, 432*, 85–92. https://doi.org/10.1016/j.scitotenv.2012.05.095.

Kloog, I., Melly, S. J., Coull, B. A., Nordio, F., & Schwartz, J. D. (2015). Using satellite-based spatiotemporal resolved air temperature exposure to study the association between ambient air temperature and birth outcomes in Massachusetts. *Environmental Health Perspectives, 123*(10), 1053–1058. https://doi.org/10.1289/ehp.1308075.

Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2014). Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. *Remote Sensing of Environment, 150*, 132–139. https://doi.org/10.1016/j.rse.2014.04.024.

Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., & Schwartz, J. (2017). Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data. *International Journal of Climatology, 37*(1), 296–304. https://doi.org/10.1002/joc.4705.

Hassani, A Hassani, A., Santos, G. S., Schneider, P., & others. (2024). Interpolation, satellite-based machine learning, or meteorological simulation? A comparison analysis for spatio-

temporal mapping of mesoscale urban air temperature. Environmental Modeling & Assessment, 29(2), 291–306. https://doi.org/10.1007/s10666-023-09943-9

Ochsenwald, W. L., Sicherman, H., et al. (2019). Israel. *Encyclopaedia Britannica*. Encyclopaedia Britannica, Inc.

Rosenfeld, A., Dorman, M., Schwartz, J., Novack, V., Just, A. C., & Kloog, I. (2017). Estimating daily minimum, maximum, and mean near-surface air temperature using hybrid satellite models across Israel. *Environmental Research, 159*, 297–312. https://doi.org/10.1016/j.envres.2017.08.017.

Schwartz, J., Samet, J. M., & Patz, J. A. (2004). Hospital admissions for heart disease. *Epidemiology, 15*(6), 755–761. https://doi.org/10.1097/01.ede.0000134875.15919.0f.

Keatinge, W. R. (1997). Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. *The Lancet, 349*(9062), 1341–1346. https://doi.org/10.1016/S0140-6736(96)12338-2.

Green, M., Pri-or, N., Capeluto, G., Epstein, Y., & Paz, S. (2013). Climate change and health in Israel: Adaptation policies for extreme weather events. *Israel Journal of Health Policy Research, 2*, 23. https://doi.org/10.1186/2045-4015-2-23.

Dudhia, J. (2008). Fundamental issues in numerical weather prediction. In *Large-Scale Disasters Prediction, Control, and Mitigation* (pp. 447-452). Cambridge University Press. https://doi.org/10.1017/CBO9780511535963.021.

Israel Meteorological Service (IMS). (n.d.). *Israel Meteorological Service website*. Israeli Ministry of Transport and Road Safety. https://ims.gov.il/en (accessed 20 February 2022).

Central Bureau of Statistics. (2018). *Statistical Abstract of Israel 2018*.

Tøttrup, C. (2014). *EU-DEM Statistical Validation Report*.

Swathi, N., Bhat, M., Reddy, N., & Rao, A. (2022). Study of distance metrics on k-nearest neighbor algorithm for star categorization. *Journal of Physics: Conference Series, 2161*, 012004. https://doi.org/10.1088/1742-6596/2161/1/012004.

Zhang, T., Zhou, Y., Zhao, K., Zhu, Z., Chen, G., Hu, J., & Wang, L. (2022). A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003–2020). *Earth System Science Data, 14*, 5637–5649. https://doi.org/10.5194/essd-14-5637-2022.

Zhou, K., Liu, H., Deng, X., Wang, H., Zhang, S., & Bonafoni, S. (2020). Comparison of machine-learning algorithms for near-surface air-temperature estimation from FY-4A AGRI data. *Advances in Meteorology*. https://doi.org/10.1155/2020/8887364.

Kohli, S., Godwin, G. T., & Urolagin, S. (2021). Sales prediction using linear and KNN regression. In S. Patnaik, X. S. Yang, & I. Sethi (Eds.), *Advances in Machine Learning and Computational Intelligence: Algorithms for Intelligent Systems*. Springer, Singapore. https://doi.org/10.1007/978-981-15-5243-4_29.

Kaggle. (2018). *Home Credit Default Risk*. https://www.kaggle.com/c/home-credit-default-risk (accessed 18 May 2022).

Kaggle. (2020). *Fashion MNIST-CNN Beginner (98% Accuracy)*. https://www.kaggle.com/general/207202 (accessed 19 May 2022).

Mukid, M. A., Widiharih, T., Rusgiyono, A., & Prahutama, A. (2018). Credit scoring analysis using weighted k-nearest neighbor. *Journal of Physics*. https://iopscience.iop.org/article/10.1088/1742-6596/1025/1/012114/pdf.

Kaggle. (2018). *Can you predict how capable each applicant is of repaying a loan?* https://www.kaggle.com/c/home-credit-default-risk (accessed 18 May 2022).

De Blesser, P., Hooghe, B., Vlieghe, D., & Van Roy, F. (2007). A distance difference matrix approach to identifying transcription factors that regulate differential gene expression. *Genome Biology, 8*(5), R83. https://doi.org/10.1186/gb-2007-8-5-r83.

Watson, G., Telesca, D., Reid, C., Pfister, G., & Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution, 254*, 112792. https://doi.org/10.1016/j.envpol.2019.06.088.

Dirksen, M., Knap, W., Steeneveld, G., Holtslag, A., & Klein Tank, A. (2020). Downscaling daily air-temperature measurements in the Netherlands. *Theoretical and Applied Climatology, 142*, 751–767. https://doi.org/10.1007/s00704-020-03313-1.

Wang, P., Mao, K., Meng, F., Qin, Z., Shu, F., & Bateni, S. (2022). A daily highest air temperature estimation method and spatial–temporal changes analysis of high temperature in China from 1979 to 2018. *Geoscientific Model Development, 15*, 6059–6083. https://doi.org/10.5194/gmd-15-6059-2022.

Li, X., Zhou, Y., Asrar, G., & Zhu, Z. (2018). Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.201

4.2 Article 2

Full reference:

Briceno, J. D., Dorman, M., Barbalat, G., Colicino, E., Stafoggia, M., Lepeule, J., & Kloog, I. (2025). A new hybrid approach applying feature engineering to model high-resolution air temperature across France.

Status: The article was completed in June 2025 and is currently finalized, awaiting agreement with the thesis directors on the target journal for submission.

PhD candidate's contribution:
The PhD candidate designed the study, downloaded and processed the data, performed the statistical analyses, interpreted the results, and wrote the manuscript. The candidate was the main author of the paper.

(The complete version of Article 2 follows on the next pages. Tables, figures, and equations retain their numbering from the latest version and have not been modified.)

A new hybrid approach applying Feature Engineering to model high resolution air temperature across France

Juan David Briceno, Michael Dorman, Guillame Barbalat, Elena Colicino, Massimo Stafoggia, Johanna Lepeule, Itai Kloog

A new hybrid approach applying Feature Engineering to model high resolution air temperature across France

Juan David Briceno[a,b*], Michael Dorman[b], Guillame Barbalat[a], Elena Colicino[d], Massimo Stafoggia[c], Johanna Lepeule[a1], Itai Kloog[b,d1*]

[a] University Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Development and Respiratory Health, Institute for Advanced Biosciences (IAB), La Tronche, France.
[b] Department of Environmental, Geoinformatics and Urban Planning Sciences, Ben-Gurion University of the Negev, P.O.B. 653, Beer Sheva, Israel
[c] Lazio Regional Health Service, Department of Epidemiology, Unit of Environmental and Occupational Epidemiology and Cancer Registry, Roma, Italy
[d] Department of Environmental Medicine and Climate Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA

* Corresponding author. Institute for Advanced Biosciences (IAB), Site Santé – Allée des Alpes, 38700 La Tronche, France. E-mail: juan.briceno-guerrero@univ-grenoble-alpes.fr, Tel: +33768296374

[1] these authors contributed equally

Abstract

Temperature is still increasing, and anthropogenic activities such as urbanization, CO_2 emissions, and overall air pollution sharpen it. There is an urgent need to measure human exposure to air temperature (Ta), specifically in countries such as France, where summers are becoming extremely hot and dangerous. This study proposes an innovative methodology to reconstruct Ta from 2014 to 2024 at 1 km resolution for continental France, and 200 m resolution for its principal urban towns. We leverage an innovative approach based on feature engineering (FE) to predict daily three Ta measures: minimum (T_{min}), average (T_{mean}), and maximum (T_{max}) temperatures. To start, we employed FE techniques to develop geographically weighted predictors and impute missing gaps in satellite-derived land surface temperature (LST) data. Once continuous surface temperature (Ts) fields were obtained, we applied Extreme Gradient Boosting (XGBoost) models to calibrate and predict Ta using other spatial and spatio-temporal predictors (elevation, population density, other temperature datasets from ECMWF Reanalysis, and EUMETSAT SEVIRI). Finally, we downscaled 1 km Ta estimations to 200 m resolution by calibrating residuals of the 1 km model (R_{Ta}) with variables at higher spatial detail, such as Landsat LST and NDVI data. Our resulting models demonstrate excellent cross-validated performance, with RMSE ranging from 0.72°C to 1.19°C for 1 km Ta, and 0.44°C to 0.88°C for 200m RTa residuals. This methodology has been shown to provide significant improvements in exposure modelling approaches such as air temperature.

Keywords

Air temperature; land surface temperature; MODIS; Feature Engineering; K-Nearest Neighbors; Machine Learning algorithms,

1. Introduction

Rising air temperatures (Ta), driven by a changing global climate, have emerged as a major concern for public health and environmental stability. Understanding how this warming trend affects human health has become an urgent priority. Numerous studies have highlighted the physiological and clinical impacts of temperature variability, linking it to a range of health issues—including complications in pregnancy (Basu et al., 2010; Kloog et al., 2015), cardiovascular strain such as elevated heart rate and blood viscosity, disruptions in blood circulation, and increased cholesterol levels (Keatinge, 1997). Additionally, extreme heat can lead to dehydration and kidney failure (Schwartz et al., 2004). In France, recent epidemiological research estimates that over 33,000 people died due to heatwaves between 2014 and 2022 (Santé publique France, 2023). While economic studies have suggested that the impact of temperature on mortality has decreased thanks to adaptation strategies (Salesse, 2024; INRAE), the main efforts have relied on data. Public institutions such as Santé publique France and Météo-France continue to strengthen the response. Through syndromic surveillance and emergency department monitoring, these agencies update the heat warning system and preventive measures annually (Santé publique France, 2023).

To accurately assess the impact of Ta, high-resolution estimates that account for both spatial and temporal variability are essential. Understanding the spatio-temporal variation of Ta enables the identification of critical exposure windows, which is particularly important for vulnerable populations such as the elderly, individuals with chronic illnesses, and those living alone (Zhou et al., 2020). This approach reduces errors in exposure assessment and mitigates biases that could lead to inaccurate conclusions regarding the health effects of temperature (Zeger et al., 2000). In contrast, relying solely on remotely sensed measurements at fixed locations may yield similar results for both urban and rural areas, which can be misleading. Cities typically experience higher temperatures due to anthropogenic factors and slower heat dissipation. Although temperature has demonstrated health effects in rural populations (Lee et al.), understanding these effects is more difficult due to the lack of high-resolution Ta estimates in such areas (Hough et al., 2020).

To achieve this, several strategies have been developed. Traditional numerical prediction models, based on physical laws, are limited by current computational capacity, resulting in medium spatial resolutions (typically around 4 km), which may not be sufficient for capturing finer-scale variations. However, more recent approaches combining machine learning (ML) and satellite data have proven to be more robust. In the case of Ta, studies such as Kloog et al. (2014) have shown that MODIS Land Surface Temperature (LST) exhibits a strong correlation with Ta and effectively captures its spatiotemporal variability, due to its high temporal and spatial resolution (daily, 1 km). On the other hand, similar ML methods have been used to model Ta across diverse geographic contexts, exhibiting strong predictive performance. These models, commonly evaluated with root mean square error (RMSE), have reported values of 2.2 °C in the USA (Kloog et al., 2014), 1.06 °C in Warsaw, Poland (Hassani et al., 2024), 1.7 °C in France (Kloog et al., 2017), and 1.2 °C in Israel (Rosenfeld et al., 2017). These results highlight the robustness of ML-based approaches for predicting Ta.

For the future, the panorama appears promising with the application of newest ML approaches in the context of exposure modelling models, such as for Ta. For instance, ensemble learning combines different ML models to enhance predictive accuracy leveraging the strengths of multiple models across different areas and periods. As an example, RMSEs of 0.4-0.6°C were achieved in France by ensembling a random forest, and an XGBoost model, within a generalized additive model (Hough et al., 2020) for air temperature residuals. Additionally, the use of Feature Engineering (FE)—which involves transforming and developing data to create new predictors—to boost the generalization capacity of models. For instance, in a model using satellite data to predict $PM_{2.5}$ (Just et al., 2025), a new feature was created using Inverse Distance Weighting (IDW), which interpolated same-day particulate measurements from sites within 500 km. This feature contributed to the development of a powerful model for predicting $PM_{2.5}$ levels in the United States.

This study introduces several innovations to enhance Ta estimation, including novel feature engineering (FE) techniques applied to LST and Ta for a more robust reconstruction of diurnal air temperatures. A key contribution is the use of FE approaches based on the k-nearest neighbors (KNN) algorithm, which leverages distance-based relationships with neighboring values to improve regression performance.

2. Data and methods

2.1 Study area

This study focuses on continental France, encompassing the entirety of mainland French territory in Europe, excluding the island of Corsica. Spanning approximately 543,000 km², the region presents a wide range of topographical and climatic conditions, with elevations stretching from below sea level (−10 m) to the peak of Mont Blanc at 4,809 m. The diverse landscape contributes to a complex climatic pattern, which can be broadly categorized into several zones (Fig.1). The northern and western coastal areas experience a humid, temperate oceanic climate, while the central-northern regions exhibit a drier and cooler variation of this pattern. Mountainous regions in the east, south-central, and southwest are characterized by montane and semi-continental climates, often marked by cold winters. In contrast, the southeastern Mediterranean coast features hot, dry summers and mild, wet winters. Similar but slightly cooler conditions extend inland and to parts of the western coastline. The southwestern basin shares many of these characteristics but tends to have drier winters (Hough et al., 2020).

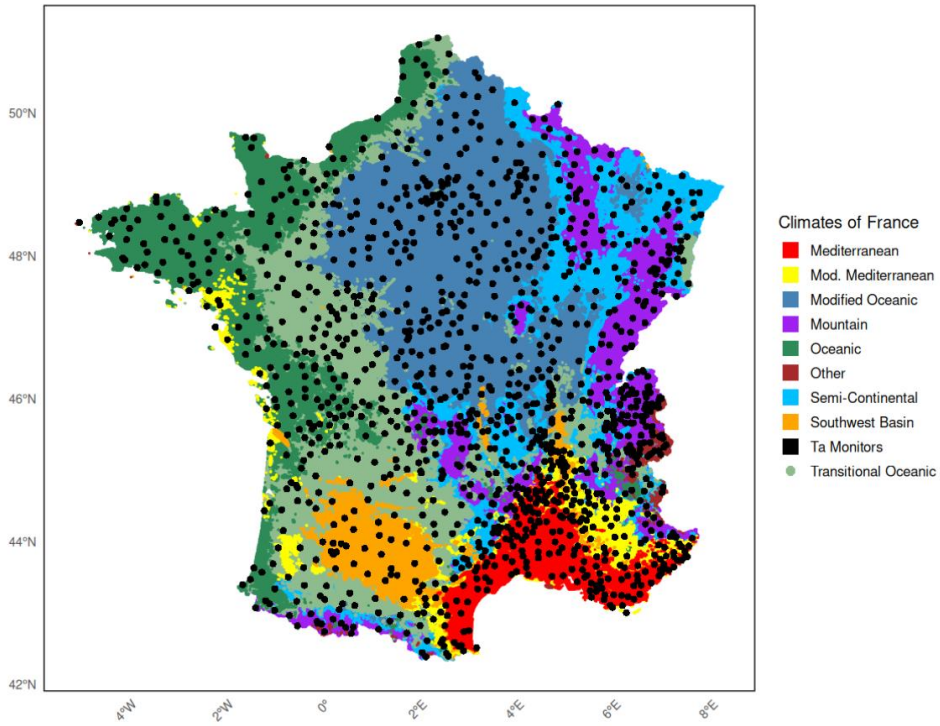


Fig. 1. French climatic regions Joly et al. (2010). As of 2024 year, 1729 Ta monitors are distributed across France (continental France except Corsica), this data is provided by Météo France.

As of January 1, 2024, the population of metropolitan France was estimated at approximately 66.55 million (INSEE, 2024). Around 82.1% of the population resided in urban areas,

corresponding to roughly 54.66 million people (Worldometers, 2024). The Paris urban area remained the most densely populated, with an estimated 11.28 million residents in 2024 and a projected increase to 11.35 million in 2025 (Macrotrends, 2024). The six next largest cities—Lyon, Marseille, Lille, Toulouse, Bordeaux, and Nice—each had populations between 0.9 and 1.8 million, collectively representing about 14–15% of the national total (Statistics Times, 2024). Mid-sized cities with populations ranging from 0.5 to 1 million accounted for approximately 10%, while smaller urban areas with fewer than 500,000 residents made up an additional 37% of the country's population.

2.2 Meteorological data

2.2.1 Monitor data

As shown in Figure 1, a total of 1729 Ta monitors were operational across France in 2024, Their control and maintenance is managed by Météo-France. Hourly data was collected and aggregated, from which daily minimum (T_{min}), mean (T_{mean}), and maximum (T_{max}) temperatures were calculated for the period 2014–2024. Monitors are distributed across the study area, with higher densities in major urban centers (Paris, Lyon, Marseille, Toulouse), mountainous regions (Alps, Pyrenees, Massif Central), and coastal zones (Atlantic, Mediterranean, and English Channel). Annual monitor deployment is provided in Appendix 1.

A preliminary analysis of the 2024 Ta data (Fig. 2) reveals that in France the hottest months of the year are June, July, and August, with average T_{max} ranging from 22°C to 27°C. Some measurements during these months even exceeded 40°C.

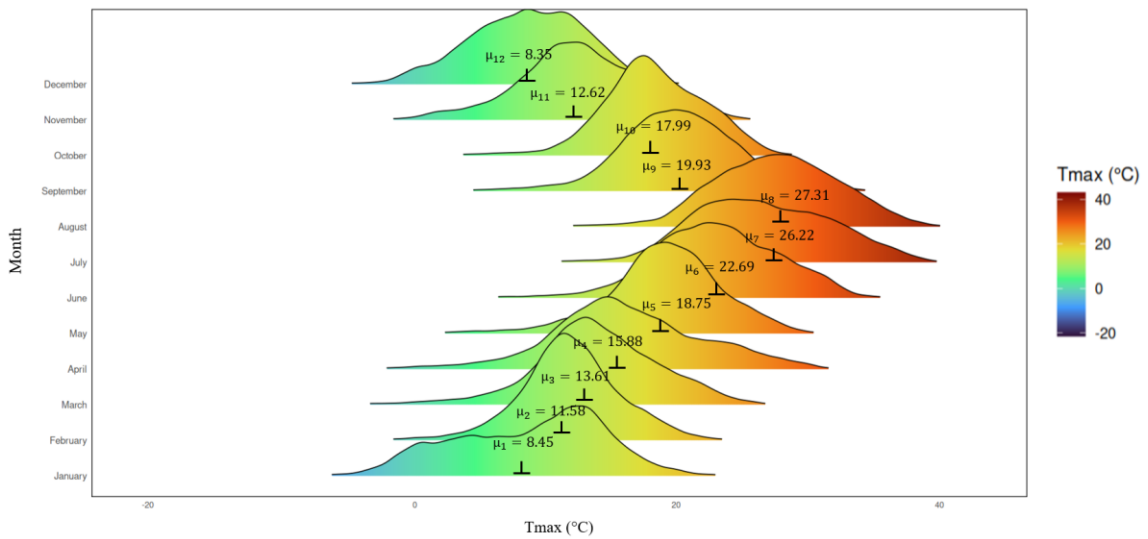


Fig. 2. Monthly distribution of observed T_{max} in France during 2024.

Regional analysis shows that coldest T_{max} values happen in the mountain region, where temperatures can drop as low as -11.2°C. In contrast, the Mediterranean region experiences the highest average T_{max} temperature at 41.2°C (Fig. 3). Temperatures surpass 40°C in the Transitional Oceanic, Southwest Basin, and Mediterranean regions. Moreover, maximum temperature spikes exceeding 35°C, a threshold considered dangerous for human health (Basu, 2009), were observed across all regions.

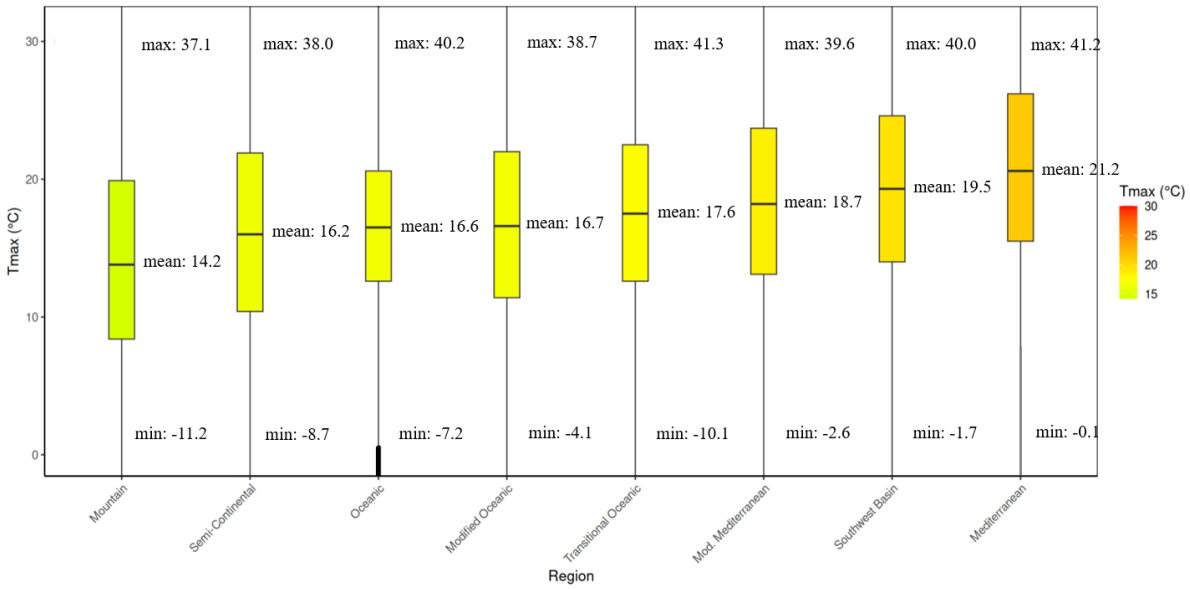


Fig. 3. Regional observed T_{max} air temperature in France in 2024.

2.2.2 Land surface temperature datasets

Land Surface Temperature (LST) data were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard NASA's Terra and Aqua satellites, which provide observations at a spatial resolution of 1 km. These satellites follow distinct polar orbits, allowing each to pass over the same location once per day; combining daytime and nighttime observations from both satellites provides up to four daily measurements per location. Although slight differences in overpass times may introduce some variability, the two satellites offer complementary LST information. For continental France, the corresponding MODIS tiles include h07v04, h07v05, h08v04, h08v05, h09v04, h09v05, h10v04, and h10v05. Due to cloud cover, MODIS data can be affected by missing observations.

Reanalysis LST data was obtained from the ERA5 dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service (C3S). ERA5 provides hourly global estimates of LST at a spatial resolution of approximately 0.1° (~11 km). In this study, LST data was extracted from the pressure levels product, accessed via the Copernicus Climate Data Store (CDS) in December 2024. The data was aggregated to daily means and resampled to match the 1 km MODIS grid. Less than 5% of the data was missing and was filled using simple mean imputation.

In addition, an LST covariate derived from the Spinning Enhanced Visible and InfraRed Imager (SEVIRI) was included. SEVIRI offers high temporal resolution (15-minute intervals) but lower spatial resolution (~3 km). The data was downloaded from the LSA SAF collections (accessed on December 10, 2024) and aggregated to daily values following the methodology described by Zhou et al. (2020).

To further support high-resolution modeling of air temperature (Ta), we also integrated LST data from Landsat 8–9 Collection 2. These observations were composited into monthly averages over a 20-year period (2004–2024), capturing fine-scale spatial variability in Ta at a resolution of 200 m (see Table 1).

2.2.3 ERA5 2-meter air temperature and Skin temperature

In this study, we used 2-meter air temperature and skin temperature from the ERA5 single levels product, which is available through the Copernicus Climate Change Service (C3S) Climate Data Store (https://cds.climate.copernicus.eu). These atmospheric datasets are produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) using data assimilation techniques similar to the ones of Integrated Forecasting System (IFS). This method combines satellite observations and historical in-situ measurements to generate reliable climate estimates. The data provides hourly estimates with a spatial resolution of approximately 0.1° (~11 km) and we retrieved the concerned for the 2014-2024 years. Less than 5% of the data was missing and was filled using simple mean imputation for gap filling.

2.2.4 NDVI

The Normalized Difference Vegetation Index (NDVI) from the MODIS Aqua satellite was employed to characterize vegetation dynamics across the study area. NDVI is derived from atmospherically-corrected surface reflectances in the blue (469 nm), red (645 nm), and near-infrared (858 nm) bands, providing a standardized measure of vegetation greenness and density (NASA). Sixteen-day composite NDVI products at a spatial resolution of 250 m were

retrieved and subsequently averaged to 1 km, matching the two-week temporal resolution of the study grids. This approach captures both seasonal and interannual variations in vegetation cover. Vegetation indices such as NDVI are relevant for air temperature modeling because they influence local temperature patterns through mechanisms including shading, modification of surface energy balance, evapotranspiration, and alteration of boundary layer dynamics (Zhou et al., 2021).

Table 1. A summary of spatio-temporal covariates used in this study.

Name	Source	Temporal resolution	Spatial resolution	Units	Nature	Stages use
Elevation	EU-DEM (Copernicus)	-	25m	m	Time invariant	1-5
Population	INSEE	-	200m	(people/km²)	Time invariant	1-5
Land cover	Corine Land Cover (CLC)	-	100m	- (%)	Time invariant	5
Climatic regions	Joly et al. (2010)	-	-	-	Time invariant	1-5
MODIS LST Aqua	Nasa	12 hours	1km	Kelvin	Time variant	1-4
MODIS LST Terra	Nasa	12 hours	1km	Kelvin	Time variant	1-4
Reanalysis LST	Reanalysis - pressure levels	Hourly	~11km	Kelvin	Time variant	1-4
2 meters air temperature	Reanalysis - single-level	Hourly	~11km	Kelvin	Time variant	1-4
Skin temperature	Reanalysis - single-level	Hourly	~11km	Kelvin	Time variant	1-4
LST SEVIRI	MVIRI - SEVIRI	Every 15mins	3km	°C	Time variant	1-4
Landsat 8,9 collection 2 LST	TIRS Nasa	16 days	100m	Kelvin	Time variant	5
NDVI Landsat 5,7,8 collection 2 Surface reflectance	TIRS Nasa	16 days	30m	- (ratio)	Time variant	5

2.2.4 Elevation

Elevation data was retrieved from version 1.1 of the European Digital Elevation Model (EU-DEM), provided by the Copernicus Land Monitoring Service. This data has a spatial resolution of 25 meters and a vertical root mean square error (RMSE) of ±7 meters (Tøttrup, 2014).

2.2.5 Population

For population data, we used the 2010 census data from INSEE, the French National Institute of Statistics and Economic Studies, and incorporated it into our 200m grid (INSEE, 2017).

2.2.6 Land cover and Climatic regions

For this study, we also used 100 m Corine Land Cover (CLC) dataset for the year 2012. We grouped the land cover classes into four main categories: artificial, vegetation, bare, and water as in (Hough et al.,2020). Additionally, we include as a covariate the French climatic regions defined by Joly et al. (2010), whose divisions are based on temperature and precipitation patterns (Fig. 1).

2.3 Statistical methods

To estimate Ta, we developed a five-step machine learning (ML) framework that builds upon and extends methodologies from previous studies (e.g., Zhou et al., 2020; Hough et al., 2020; Kloog et al., 2017; Kloog et al., 2014). A key innovation of our approach is the integration of feature engineering (FE) within the imputation and calibration phases (see Figure 4). Rather than flattening satellite imagery into data tables, we generate geographically weighted predictors based on MODIS LST and ground-based Ta measurements. This allows us to preserve local spatial relationships between individual locations and their surrounding areas—information that is often lost during standard data preprocessing and randomization. Including these new predictors in our models improved the accuracy of our 1 km Ta models for France.

Stages 1 through 4 of the workflow were dedicated to generating Ta estimates at 1 km resolution. In Stage 5, the final phase of our approach, we enhanced the spatial resolution of Ta from 1 km to 200 m, focusing on predicting over principal French urban areas. Following the same approach described by Hough et al. (2020), we first calculated residuals by subtracting the 1 km model predictions from observed Ta values at monitoring sites. These residuals, which reflect local variations not captured by the coarser model, were then mapped onto a 200 m grid. We trained an extreme gradient boosting (XGBoost) model on these high-

resolution residuals to capture fine-scale spatio-temporal variability. Finally, the 200 m residual predictions were added to the original 1 km Ta estimates, resulting in the final downscaled 200 m Ta surfaces for urban zones. For all model training stages, we split the available data into training (80%) and testing (20%) sets, incorporating 5-fold cross-validation to reduce bias and minimize the risk of overfitting.

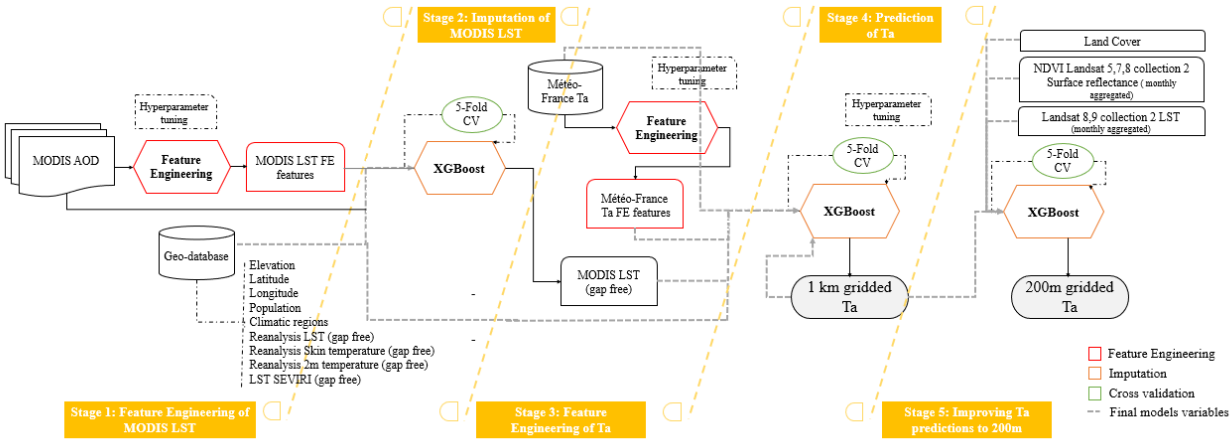


Fig. 4. Schematic diagram of the five-stage model for estimating Ta in France.

Modelling Ta at 1km

Stage 1 model: Feature Engineering of MODIS LST

In Stage 1, we applied a gap-filling process to MODIS LST Aqua variables (Night and Day) and MODIS LST mean, which is the average of all four MODIS overpasses. To accomplish this, we created new features using a strategy based on the K-Nearest Neighbors (KNN) algorithm, a non-parametric supervised machine learning method. KNN works by keeping all available data and predicting missing values by finding the most similar existing data points, based on a chosen distance metric (Swathi et al., 2022).

Having a matrix containing P points representing France. For each point *p* in P, we identify a set of neighboring points (i, j, g, …l in P | i, j, g, …l ≠ p) that share for example similar latitude and longitude values. Similarity in this algorithm is determined solely by geographical attributes, like longitude (x_p) and latitude (y_p). Although other variables, such as elevation (E_p) or population density (Pop_p), can be included, we limited our analysis to the first two spatial coordinates (see Equation 1). Distances between points are organized into a distance matrix D. For each point *p*, we identify a set of *k* nearest neighbors (*k*N*p*). Then, for

each point *p* and date *t*, we compute the average of LST values of its neighbors, resulting in a new feature that estimates LST at those locations and times.

Because an FE can have missing values—due to missing neighboring LST observations around a point *p*—we fill gaps applying simple imputation. Specifically, we replaced missing values (NAs) in the feature with the average LST of the point *p* within the same week of the year (each year divided into 52–54 weeks, and our models trained annually). Alternatively, missing values could be imputed using an ML model, where the FE is predicted based on other available spatio-temporal predictors (such as latitude, longitude, elevation, population, reanalysis LST, reanalysis 2-meter air temperature, etc.). In practice, the number of missing values in the FE is always lower than in the original LST product, and it decreases as more neighboring points are used in the calculation. The optimal number of neighbors should be selected through hyperparameter tuning.

To select the optimal number of nearest neighbors *k*, we create FE predictors iteratively by varying *k* (see Appendix 2 for further details on the FE generation). We identified the optimal FE variable to include in our model by minimizing the root mean square error (RMSE). This was done by sequentially adding each FE variable into a regression model of the MODIS LST variable, one at a time. We then determined the optimal feature, denoted as FE(b), where k = b, as the one beyond which no significant improvement in RMSE was observed when increasing k from b to b + a (with a = 1, 2, 3, 4, …). The feature FE(b + a) was also tested in the regression model, but it did not yield a meaningful reduction in RMSE compared to FE(b), confirming FE(b) as the optimal choice. We named this methodology for tuning *k* as the "cut-off point rule".

$$x_{\mathrm{p}}: \text{ Longitude of cell p} \in \mathrm{P}, \qquad y_{\mathrm{p}}: \text{ Latitude of cell p} \in \mathrm{P}$$

$$D_{\mathrm{pj}}: \text{ Distance from cell p} \in \mathrm{P} \text{ to cell j} \in \mathrm{P}.$$

$$D_{\mathrm{pj}} = \sqrt{\left(x_{\mathrm{p}} - x_{\mathrm{j}}\right)^2 + \left(y_{\mathrm{p}} - y_{\mathrm{j}}\right)^2} \quad \forall \mathrm{p,j} \in \mathrm{P} \qquad \text{Eq. 1}$$

$$D = \begin{vmatrix} 0 & D_{1,2} & \dots & D_{1,|P|} \\ D_{2,1} & 0 & \dots & \dots \\ \dots & \dots & 0 & D_{|P|-1,|P|} \\ D_{|P|,1} & \dots & D_{|P|,|P|-1} & 0 \end{vmatrix}$$

For each MODIS LST variable, we generated 20 FEs based on different values of the *k* parameter (see Equation 2). The distance matrix D was calculated using latitude and longitude

coordinates (see Figure 5). To keep the model efficient, we followed a parsimonious approach using XGBoost for each iteration. Date-related variables were binarized, including calendar months (e.g., Jan, Feb, Mar), weekdays (e.g., Mon, Tue, Wed), and weeks of years (e.g., 1,2,3…52).

$$\mathrm{FE}^{k}{}_{LST}(p,t) = \frac{1}{|\ kN_p{}^{valid}\ |} \sum_{\mathrm{j} \in kN_p{}^{valid}} LST_{j,t}\ \forall \mathrm{p} \in \mathrm{P}, \mathrm{t} \in \mathrm{T} \qquad \text{Eq. 2}$$

$$|\ kN_p{}^{valid}\ | \leq \mathrm{k}$$

$kN_p{}^{valid}$: Set of valid nearest neighbors for point p where the LST value is not missing.

$|\ kN_p{}^{valid}\ |$: Number of valid neighbors for point p where the LST is not missing.

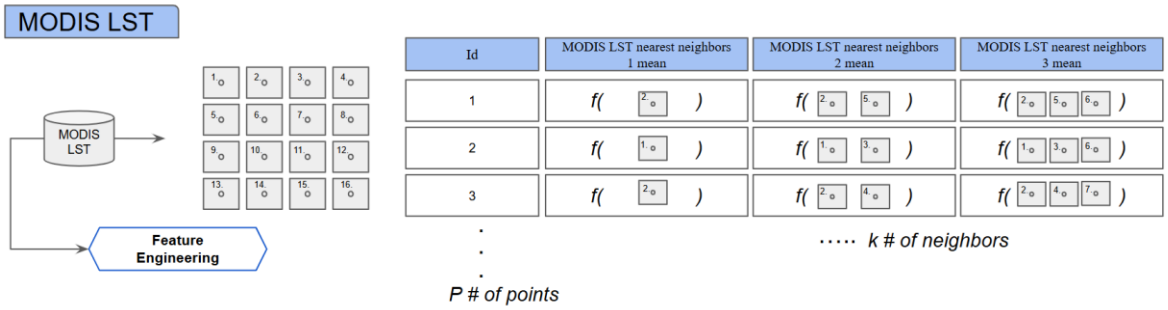


Fig. 5. A simplified example for a specific day is provided, illustrating how distance is quantified using latitude and longitude geographical coordinates when calculating FEs for MODIS LST variables.

Stage 2 model: Imputation of MODIS LST

In Stage 2, we filled gaps in MODIS LST variables by training XGBoost models using the following predictors: latitude, longitude, elevation, population, binarized date variables, LST SEVIRI, reanalysis LST, 2-meter temperature, skin temperature, and the selected FE from Stage 1. Equation 3 summarizes our final Stage 2 model, which includes the optimized selected FE feature related to the latitude and longitude geographical coordinates used for the distance matrix D.

$$\mathrm{LST} \sim f(\mathrm{lat}, \mathrm{lon}, \mathrm{elevation}, \mathrm{population}, \mathrm{Reanalysis\ LST}, \mathrm{Reanalysis\ skin\ temperature},$$
$$\mathrm{Reanalysis\ 2m\ temperature}, \mathrm{LST\ SEVIRI}, \mathrm{Monday}, \mathrm{Tuesday}, \mathrm{Wednesday} \ldots, \mathrm{January}, \mathrm{Febrary},$$
$$\mathrm{March}, \ldots\ \mathrm{Week1}, \mathrm{Week2}, \mathrm{Week3}..\mathrm{Week52},$$

FE MODIS LST) Eq. 3

Stage 3 model: Feature Engineering of T_a

To produce daily estimates of Ta, we modified our initial feature engineering (FE) approach by restricting the FE generation exclusively to the locations of Ta monitors. A key distinction in this stage, compared to our initial FE process in Stage 1, is the dimensionality reduction of the distance matrix D. Specifically, the matrix was reduced from a full (P,P) form—where P denotes the total number of points across France—to a more targeted (P,s) structure, with s representing the number of points with active Ta monitors. Appendix 3 provides a comprehensive explanation of the changes applied to D during this modified FE process. Again, to select the most relevant FE to include in our models, we applied our selection method known as the "cut-off point rule". Moreover, we incorporated inverse distance weighting (IDW) into the FE procedure to ensure that observations from nearby monitors had greater influence when computing Ta averages for a given location (Fig.6). Equation 4 exhibits how the FE is calculated. As illustrated in Figure 7, the IDW method assigns higher weight—such as (W1, S2) —to measurements of Ta monitors like S2 that are geographically closer to a particular point *p* of interest (for instance, *p*=1). This weighting logic dynamically adjusts for each grid point based on its spatial relationship to surrounding monitors.

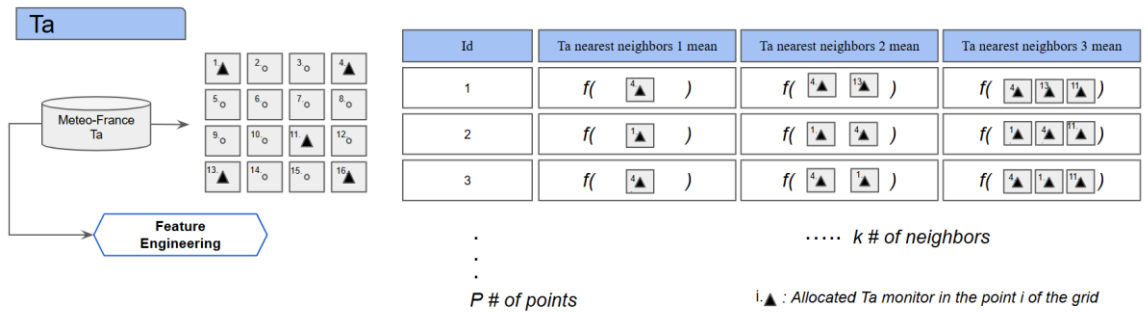


Fig. 6. Explanation of the Ta FE calculation. This minimalistic example illustrates the process for a specific day, using latitude and longitude geographical coordinates to quantify distance in the distance matrix D.

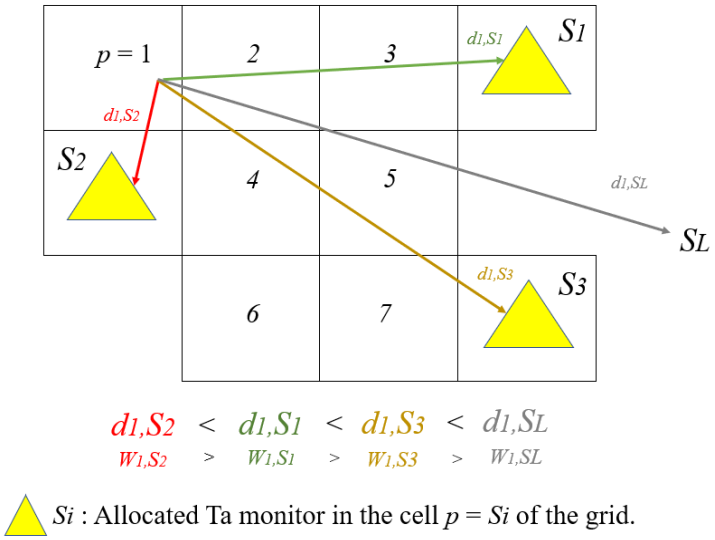


Fig. 7. Logic behind the implementation of inverse distance weighting (IDW) for the FE calculation of Ta.

We trained an XGBoost model for each *i-th* value used as k for the feature generation. Our final model for this stage's hyperparameter tuning is summarized by equation 5.

$$\mathrm{FE}^{k}{}_{Ta}(p,t) = \frac{1}{|\, ks_p{}^{valid} \,|} \sum_{\mathrm{j} \in ks_p{}^{valid}} W_{p,j} * Ta_{j,t} \;\forall \mathrm{p} \in \mathrm{P}, \mathrm{t} \in \mathrm{T} \quad \text{Eq. 4}$$

$$|\, ks_p{}^{valid} \,| \leq \mathrm{k} \leq \mathrm{s}$$

$ks_p{}^{valid}$: Set of valid nearest neighbors monitors of point p where the Ta value is not missing.

$|\, ks_p{}^{valid} \,|$: Number of valid nearest neighbors monitors for point p where the Ta is not missing.

$\mathrm{Ta}^{b} \sim f$(lat, lon, elevation, population, Reanalysis LST, Reanalysis skin temperature, Reanalysis 2m temperature, LST SEVIRI, Monday, Tuesday, Wednesday … , January, Febrary, March, … , Week1, Week2, Week3, … Week52,

$$\text{gap} - \text{free MODIS LST}, \ \mathrm{FE}^{b}\ \mathrm{Ta}) \qquad \text{Eq. 5}$$

Stage 4 model: prediction of T_a

In Stage 4, we predicted daily Ta variables (T_{min}, T_{mean}, and T_{max}) using latitude, longitude, elevation, population density, binarized date variables (day of week, month, and week of the

year), LST from SEVIRI, reanalysis LST, 2-meter air temperature, skin temperature, gap-free MODIS LST, and the corresponding FE Ta feature.

For the T_{min} model, the previously predicted T_{mean} was added as an additional predictor. Similarly, for the T_{max} model, both predicted T_{mean} and T_{min} were included as features. Incorporating these hierarchical dependencies ensures that, for any given day *t* and location *p*, the logical constraint $T_{min} < T_{mean} < T_{max}$ is satisfied. Equations 6 (A, B, C) summarize these mathematical formulations.

$T_{mean} \sim f$(lat, lon, elevation, population, Reanalysis LST, Reanalysis skin temperature, Reanalysis 2m temperatur

LST SEVIRI, Monday, Tuestday, Wednesday … , January, Febrary, March, … , Week1, Week2, Week3 … Week52,

$$\text{gap} - \text{free MODIS LST}, \text{selected FE } T_{mean}) \qquad \text{Eq. 6. A}$$

$T_{min} \sim f$(lat, lon, elevation, population, Reanalysis LST, Reanalysis skin temperature, Reanalysis 2m temperature

LST SEVIRI, Monday, Tuestday, Wednesday … , January, Febrary, March, … , Week1, Week2, Week3 … Week52,

$$\text{gap} - \text{free MODIS LST}, \text{selected FE } T_{min}, \text{predicted } T_{mean}) \qquad \text{Eq. 6. B}$$

$T_{max} \sim f$(lat, lon, elevation, population, Reanalysis LST, Reanalysis skin temperature, Reanalysis 2m temperature

LST SEVIRI, Monday, Tuestday, Wednesday … , January, Febrary, March, … , Week1, Week2, Week3 … Week52,

$$\text{gap} - \text{free MODIS LST}, \text{selected FE } T_{max}, \text{predicted } T_{mean}, \text{predicted } T_{min}) \qquad \text{Eq. 6. C}$$

Spatial blocking cross-validation

To reduce potential spatial bias in the 1 km Ta estimates—particularly the risk of overfitting—we implemented a spatial blocking strategy. In this alternative validation, we held out a set of Ta monitors before splitting the data into training and testing sets, ensuring models had no access to the held-out monitors measures during learning (Fig. 8). Given that Ta monitors are densely and fairly evenly distributed across France, we began by selecting a representative sample of 20 monitors using clustering (Fig. 8B). Around each of these 20 locations, we identified 6 to 8 nearby monitors and held out all of them (Fig. 8C). In total, 199 out of 1729 monitors were excluded from training (2024 year), covering a wide range of geographic areas across France. We evaluated our final models by predicting Ta at these held-out locations—where ground-truth Ta values were available—and compared the predictions accordingly.

To prevent data leakage during this validation, we recalculated the Ta FE features while excluding the held-out monitors values from all averages. Including their values in feature computation would introduce bias and artificially inflate model performance. This validation approach provides a more realistic and robust assessment of the model's ability to generalize in previously unobserved areas.

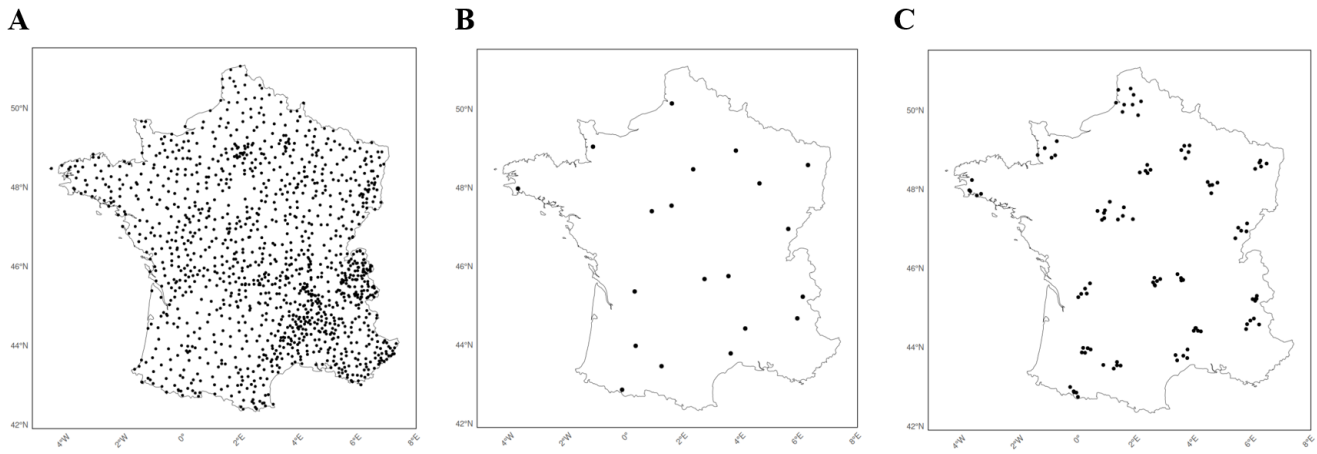


Fig. 8. Steps for selecting held out monitors in the spatial blocking cross-validation of 1km Ta models.

A. The total number of Ta monitors in France (2024).

B. The initial selected sample of 20 Ta monitors to be held out, identified using clustering.

C. The complete set of 199 held out monitors used to validate our models' performance.

Stage 5 model: improving Ta predictions to 200 m

In the final phase of our approach (Stage 5), we improved the spatial resolution of Ta estimates to 200 meters to more accurately reflect intra-urban temperature variability within major French cities. This refinement involved first computing residuals at monitoring station locations, by subtracting the predicted 1 km Ta of our previous stages' models and the corresponding monitors' measurements. These residuals were subsequently used to train XGBoost models, which enabled us to extrapolate high-resolution temperature estimates across the full 200 m grid. The model inputs included 200 m resolution geographic coordinates (latitude and longitude), elevation, population density, land cover classifications, French climatic region, binarized date variables, monthly averaged Land Surface Temperature (LST) from Landsat Collection 2 (products 8 and 9), and monthly averaged Normalized Difference Vegetation Index (NDVI) derived from Landsat Collection 2 surface reflectance,

covering the period from 2014 to 2024. Additionally, we integrated the previously predicted 1 km resolution Ta into the corresponding residual models—daily predicted T_{min} for R_{Tmin}, predicted T_{mean} for R_{Tmean}, and predicted T_{max} for R_{Tmax}—as outlined in Equations 7 and 8. This integration of spatial, temporal, and remote sensing variables enabled a more accurate representation of urban microclimates across the studied French metropolitan areas

$$\text{Residuals } T_{mean} \text{ at 200 m} \sim \text{XGBoost}(\text{lat, lon, population, elevation, land cover, French climatic regions, monthly averaged Landsat LST 200 m, monthly averaged Landsat NDVI 200 m, Monday, Tuestday, Wednesday ..., January, Febrary, March, ..., Week1, Week2, Week3 ... Week52, } predicted\ T_{mean} \text{ at 1km}) \quad \text{Eq. 7}$$

$$T_{mean} \text{ at 200 m} = predicted\ T_{mean} \text{ at 1 km} + \text{Residuals } T_{mean} \text{ at 200 m} \quad \text{Eq. 8}$$

3. Results

3.1 Correlation heatmap

Figure 9 displays the correlations calculated between all time-variant 1km predictors used in this study.

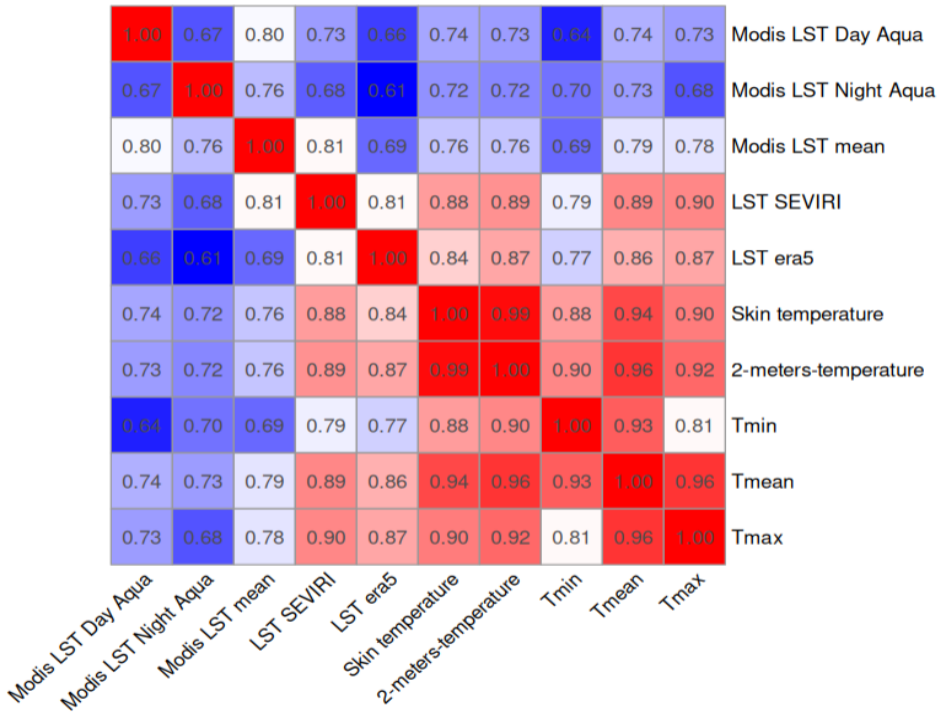


Fig. 9. Correlation heatmap of all time-variant 1km predictors, and Ta for 2024 year.

3.2 Tuning *k* parameter in stage 1 model

We started working on the MODIS LST Night Aqua variable, with a feature generation varying from k=1 to k=20. We conducted a simple grid search for XGBoost and found that setting the number of trees to 85 and col-sample-rate-per-tree to 0.5 yielded strong modeling performance. Initially, the XGBoost model trained without any FE resulted in an RMSE of 2.05°C. However, after incorporating the first FE (k = 1), the RMSE dropped to 0.82°C. With further iterations, we found that the optimal cut-off point was at k = 5, which reduced the RMSE to 0.33°C (Figure 10.A). A similar pattern was observed with the other two MODIS LST variables—Day Aqua and the mean of 4 overpasses—where the best performance also occurred at k = 5, yielding RMSEs of 0.46°C and 0.95°C, respectively (Figures 10.B and 10.C). These results correspond to the year 2024. We applied the same methodology with other years of our study (2014-2023).

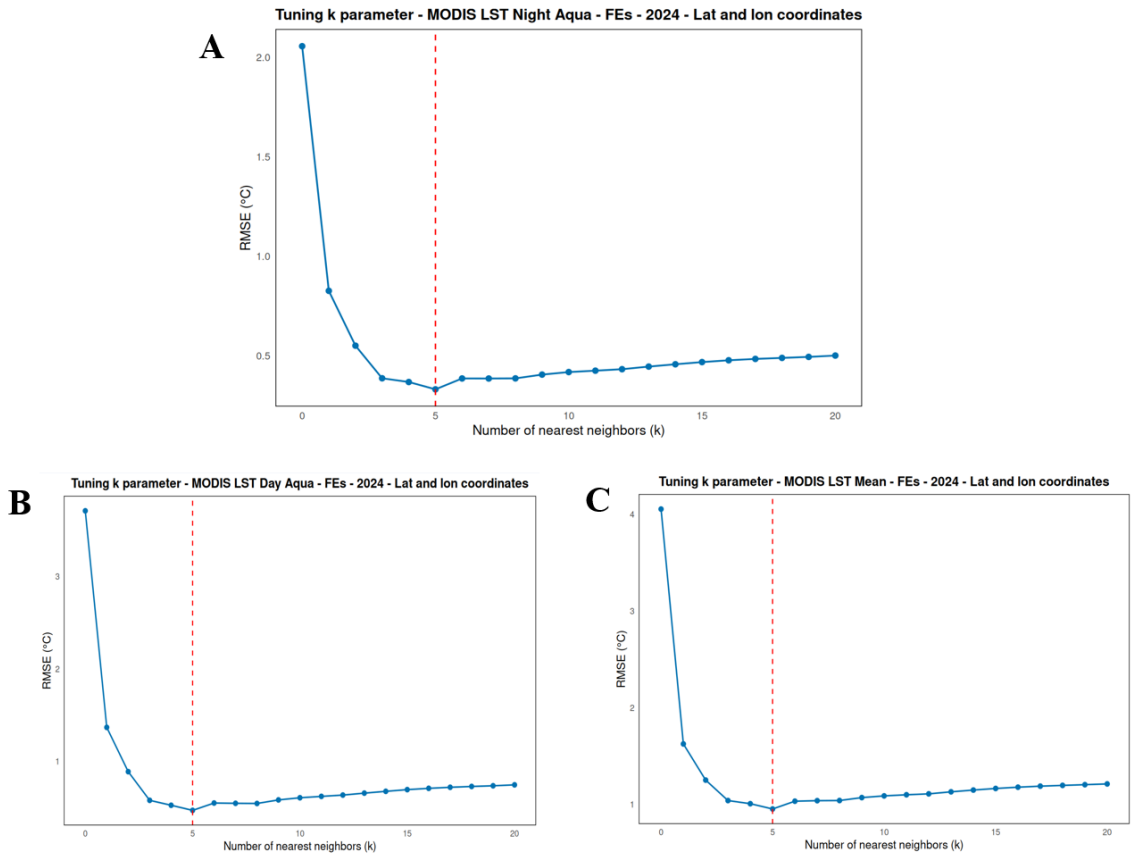


Fig. 10. FE selection for the model that imputes MODIS LST variables, year 2024.

A. Hyperparameter tuning for k to select FE of MODIS LST Night Aqua variable.

B. Hyperparameter tuning for k to select FE of MODIS LST Day Aqua variable.

C. Hyperparameter tuning for k to select FE of MODIS LST mean variable.

3.3 Performance of the stage 2 model

Table 2 presents the 5-fold cross-validation performance of the stage 2 models across all years (2014-2024), incorporating the optimal FE features found in the previous stage. The average RMSEs over the years were 0.33°C for MODIS LST Night Aqua, 0.55°C for MODIS LST Day Aqua, and 1.01°C for MODIS LST Mean.

Table 2. Stage 2 model performance of MODIS LST variables. 5-fold cross-validation have been used across all years (2014-2024). RMSE metrics were obtained from XGBoost models with 85 trees and a column sampling rate of 0.50.

	MODIS LST Night Aqua	MODIS LST Day Aqua	MODIS LST mean
Year	RMSE (°C)	RMSE (°C)	RMSE (°C)
2024	0.33	0.46	0.95
2023	0.32	0.50	0.98
2022	0.34	0.58	0.99
2021	0.34	0.50	1.08
2020	0.33	0.57	1.03
2019	0.33	0.58	1.01
2018	0.317	0.59	1.01
2017	0.33	0.55	1.02
2016	0.33	0.56	0.95
2015	0.34	0.57	1.00
2014	0.32	0.58	1.01

3.2 Tuning *k* parameter in stage 3 model

To determine the optimal value of *k* and select the final FE features to include in the Ta models, we generated 20 features using Inverse Distance Weighting (IDW). Feature selection was based on the predefined cut-off rule. As in stage 1, we performed a simple grid search for XGBoost and found that using 1000 trees and setting col-sample-rate-per-tree to 0.45 provided good modeling performance. The tuning process led to an optimal *k* of 2 for the FE features associated with the Ta variables (T_{min}, T_{mean}, and T_{max}), resulting in RMSEs of 1.09°C, 0.76°C, and 0.84°C, respectively (in 2024). These results were consistent across all

study years (2014–2024) when using latitude and longitude coordinates to construct the distance matrix D, suggesting that annual re-optimization of this parameter is not necessary. Figure 11 shows the tuning results for 2024, while results for 2023 are included in Annex A.

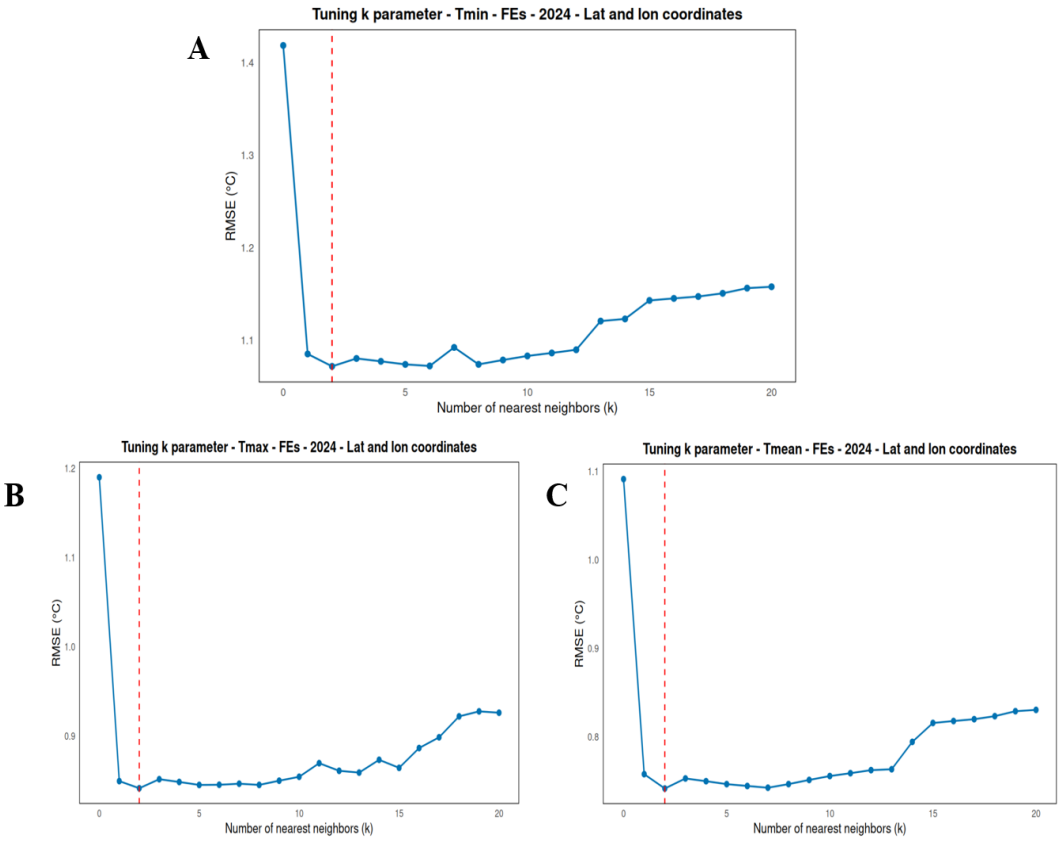


Fig. 11. FE selection for the model that imputes Ta variables, year 2024.

A. Hyperparameter tuning for k to select FE of Tmin.

B. Hyperparameter tuning for k to select FE of Tmax.

C. Hyperparameter tuning for k to select FE of Tmean.

3.3.1 Performance of the stage 4 model

Table 3 presents the overall results for the 1 km Ta models across the years. In general, the models performed well, with average RMSE values of 0.79°C for T_{mean}, 1.08°C for T_{min}, and 0.91°C for T_{max} (2014-2024). We wanted our performance analysis to be deepened and for that we checked out the temporal performance of our models by months. That allowed us to conclude that our Ta predictions are more accurate during the periods May-Sept, and worst during months comprising December-February (see Table 4). At the same time, we evaluated

the models spatially across the different French climatic regions. We concluded the models perform better in regions such us the Modified Oceanic, Oceanic, and the Southwest Basin. On the other side, worst performance is usually gotten in regions like Mod. Mediterranean and Mountain (see Table 5).

Table 3. Stage 4 model (predicting daily 1 km Ta): 5-fold cross-validated performance across all years (2024–2014), with RMSE values from XGBoost models with 1000 trees and a column sampling rate of 0.45.

	T_{min}	T_{mean}	T_{max}
Year	RMSE (°C)	RMSE (°C)	RMSE (°C)
2024	1.07	0.74	0.84
2023	1.09	0.76	0.86
2022	1.12	0.79	0.87
2021	1.08	0.79	0.88
2020	1.02	0.80	0.89
2019	1.01	0.76	0.92
2018	0.98	0.72	0.87
2017	1.09	0.83	0.92
2016	1.09	0.83	0.95
2015	1.19	0.87	1.03
2014	1.18	0.89	1.07

Table 4. Stage 4 model performance for all years by month, using XGBoost models with 1000 trees and a column sampling rate of 0.45.

	T_{min}	T_{mean}	T_{max}
Month	RMSE (°C)	RMSE (°C)	RMSE (°C)
Jan	1.16	0.94	0.98
Feb	1.12	0.85	0.92
Mar	1.18	0.77	0.90
Apr	1.03	0.77	0.85
May	1.05	0.69	0.85
Jun	1.04	0.67	0.86
Jul	1.01	0.68	0.81
Aug	1.03	0.72	0.85
Sep	1.03	0.71	0.88

Oct	1.11	0.77	0.97
Nov	1.07	0.91	1.01
Dec	1.09	0.97	1.02

Table 5. Stage 4 model performance for all years by climatic region, using XGBoost models with 1000 trees and a column sampling rate of 0.45.

	T_{min}	T_{mean}	T_{max}
Climatic region	RMSE (°C)	RMSE (°C)	RMSE (°C)
Mediterranean	1.12	0.87	0.96
Mod. Mediterranean	1.22	0.86	0.94
Modified Oceanic	0.86	0.74	0.81
Mountain	1.28	0.85	1.05
Oceanic	0.94	0.67	0.76
Other	1.14	0.75	0.86
Semi-Continental	1.16	0.82	0.95
Southwest Basin	0.97	0.78	0.87
Transitional Oceanic	1.02	0.76	0.86

3.3.2 Performance of the stage 4 model with Spatial blocking cross-validation

To assess spatial bias in our final models, we applied spatial blocking, 199 Ta monitors were hold out before training stages, applying the strategy in 2024 and 2023 years. This validation approach demonstrated strong performance on the held-out monitors. However, the RMSE for each Ta variable increased slightly after applying it. For instance, the RMSE for T_{mean} rose from 0.74°C to 1.02°C, T_{min} RMSE increased from 1.07°C to 1.33°C, and T_{max} RMSE went from 0.84°C to 1.10°C in 2024. Table 6 presents our results for the held-out monitors, which are shown in Figure 8.

Table 6. 5-fold cross-validated RMSE of Ta models when spatially blocking 199 monitors (2024-2023).

	2024			2023		
	T_{min}	T_{mean}	T_{max}	T_{min}	T_{mean}	T_{max}
Held out Ta monitors (By regions)	Results held out monitors – RMSE (°C)					
Mediterranean	1.72	1.35	1.25	1.75	1.38	1.28
Mod. Mediterranean	1.24	0.94	1.04	1.26	0.96	1.06
Modified Oceanic	1.19	0.52	0.78	1.21	0.53	0.79

Mountain	1.49	1.34	1.48	1.51	1.37	1.51
Oceanic	1.14	0.77	0.82	1.16	0.79	0.84
Other	1.48	2.18	1.60	1.50	2.22	1.63
Semi-Continental	1.48	1.20	0.93	1.51	1.22	0.94
Southwest Basin	1.10	0.61	0.87	1.11	0.62	0.88
Transitional Oceanic	0.85	0.46	0.76	0.86	0.47	0.77
Average RMSE (Held out monitors)	1.33	1.02	1.10	1.35	1.04	1.12

3.3.3 Predictions of the stage 4 model

The analysis of predictions from our 1 km stage 4 models reveals a clear upward trend in Ta over the years (2014-2024), with increasingly intense summer heatwaves. Notable shifts occurred during the periods 2014–2015, 2016–2018, and 2021–2022, which showed significant year-to-year temperature changes (ΔT_{mean}), as illustrated in Figures 12 and 13. The highest average maximum temperatures (T_{max}) were observed in 2018 and 2022. Furthermore, an in-depth comparison of predicted temperatures on a fixed date (February 18) across the years 2014, 2019, and 2024 indicates a marked warming trend in the central and northern regions of the country. Even traditionally colder, snow-prone areas in the southeast now exhibit significantly higher temperatures than those recorded a decade ago (Figure 14).

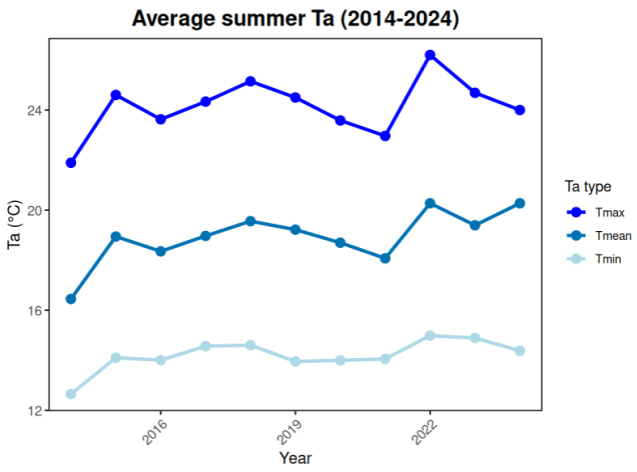


Fig. 12. Mean summer Ta (T_{min}, T_{mean}, T_{max}) between 2014 and 2024 in continental France.

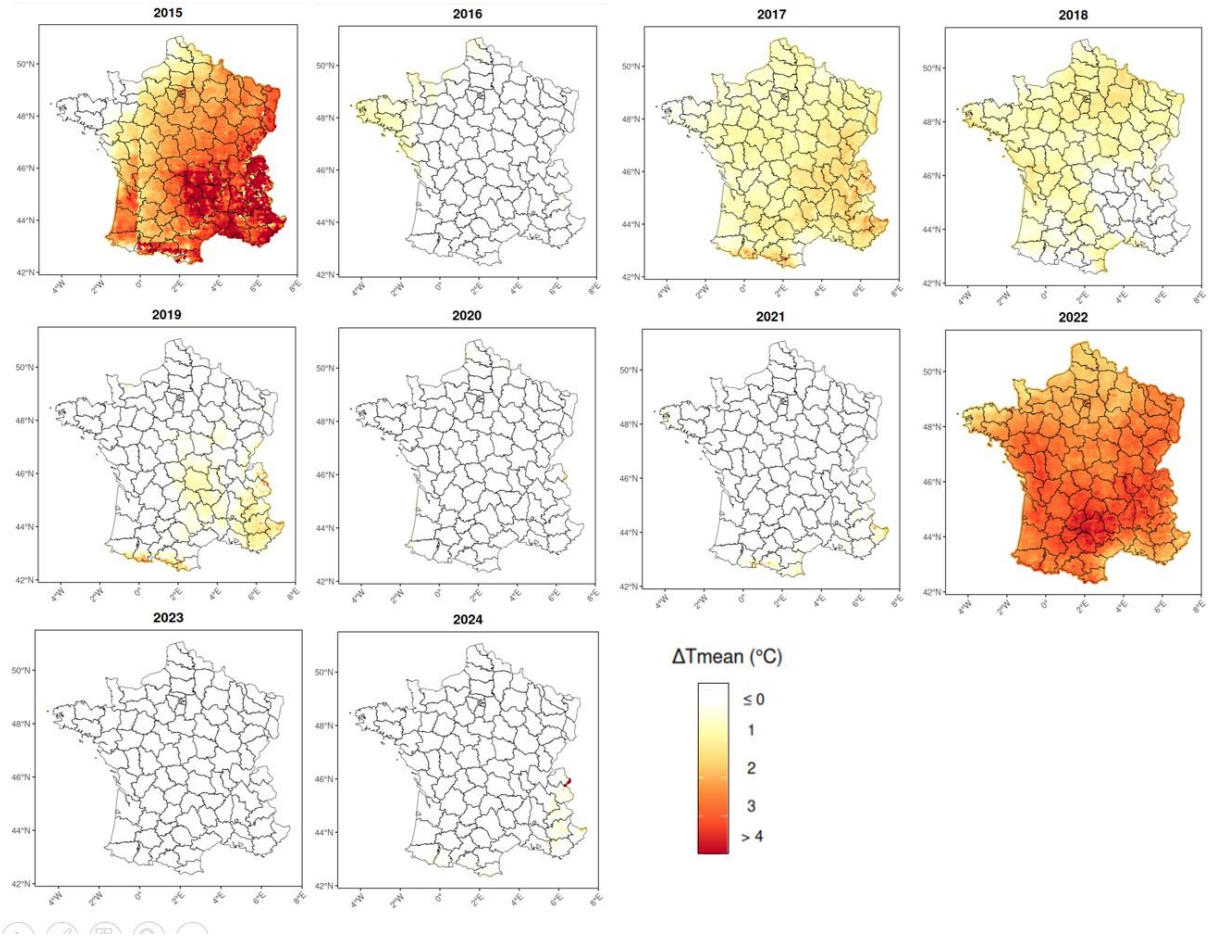


Fig. 13. Average summer ΔTmean between 2015 and 2024 in continental France.

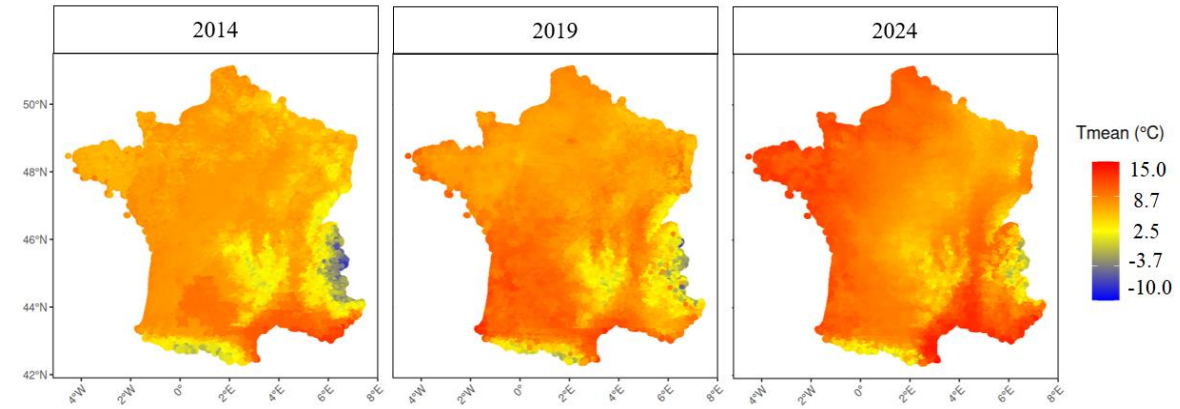


Fig. 14. Predicted 1 km T_{mean} from Stage 4 models on 18th of February in 2014, 2019, and 2024.

3.3.4 Variable importance

In general, the most important predictors for our final 1 km Ta models are the FE Ta variables conceived in earlier stages, the gap-free MODIS LST variables, lat-lon coordinates (y, x), elevation, and other temperature variables such as LST SEVIRI, reanalysis LST, skin

temperature, and 2-meter air temperature (particularly for the T_{mean} model). Each Ta model has its own top 10 explanatory predictors, as shown in Fig. 15.

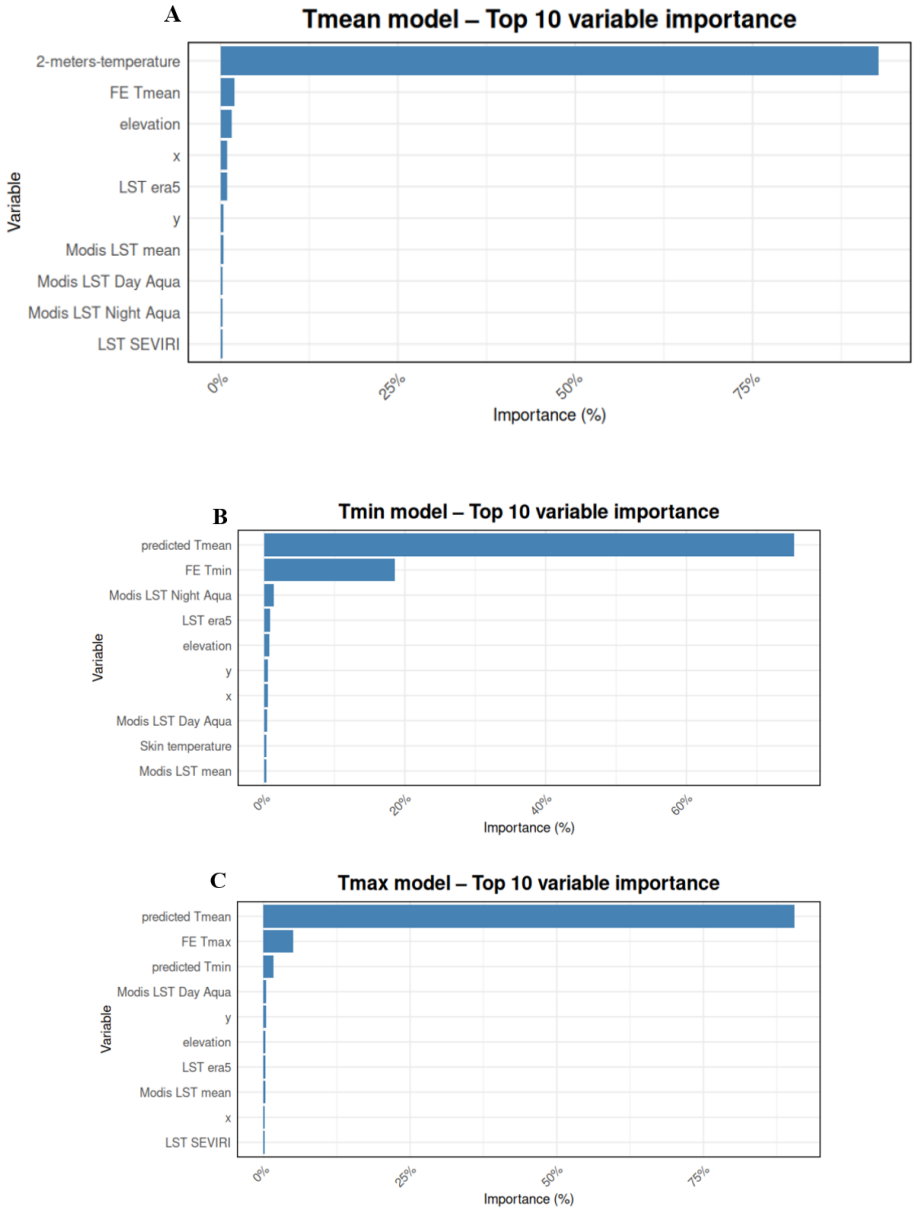


Fig. 15. Variable importance for final 1km Ta models.

A. Feature importance T_{mean} model (top 10 variables).

B. Feature importance T_{min} model (top 10 variables).

C. Feature importance T_{max} model (top 10 variables).

3.4 Performance of the stage 5 model

Table 7 presents the overall results for the 200 m models estimating Ta residuals across all years of the study. In summary, these models performed well, with average RMSE values of 0.55°C for R_{Tmean}, 0.66°C for R_{Tmax}, and 0.76°C for R_{Tmin} on the validation sets.

Table 7. RMSEs of Ta residuals at a 200 m resolution, calculated using 5-fold cross-validation across all years (2024–2014). XGBoost models were trained with the following parameters 150 trees and a column sampling rate of 0.50.

	R_{Tmin}	R_{Tmean}	R_{Tmax}
Year	RMSE (°C)	RMSE (°C)	RMSE (°C)
2024	0.65	0.44	0.57
2023	0.62	0.45	0.55
2022	0.64	0.47	0.58
2021	0.65	0.45	0.59
2020	0.66	0.46	0.54
2019	0.88	0.63	0.76
2018	0.87	0.44	0.74
2017	0.86	0.69	0.76
2016	0.85	0.66	0.74
2015	0.86	0.68	0.76
2014	0.88	0.69	0.75

Additionally, we examined the performance of these 200 m models temporally and spatially by evaluating RMSEs by months and by climatic regions for all analyzed years. Table 8 and 9 presents our findings. Similar to the 1 km models, the 200 m models perform better in regions such as Modified Oceanic, Oceanic, and Transitional Oceanic, but worse in others like Modified Mediterranean and Mountain.

Table 8. Stage 5 model performance for all years by month, using XGBoost models with 150 trees and a column sampling rate of 0.50.

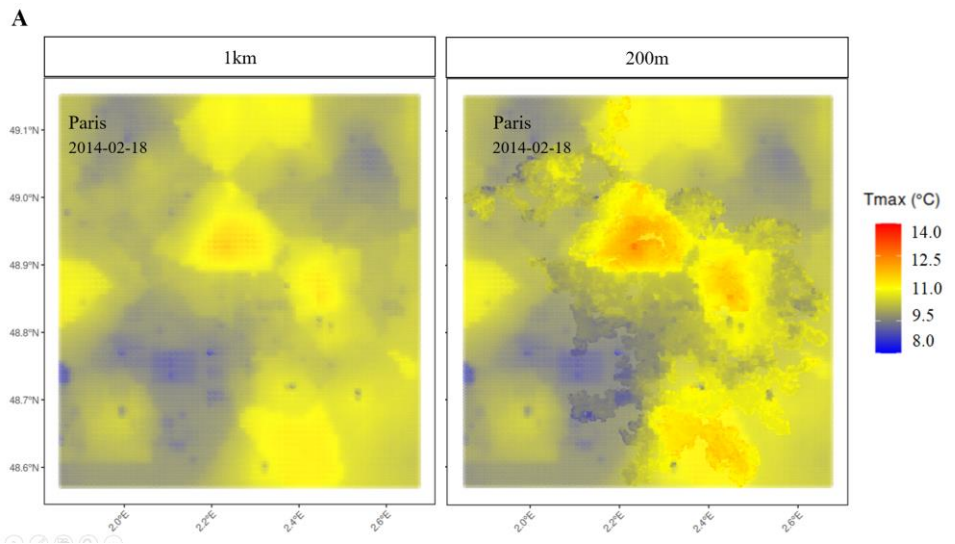
	R_{Tmin}	R_{Tmean}	R_{Tmax}
Month	RMSE (°C)	RMSE (°C)	RMSE (°C)
Jan	0.82	0.65	0.71
Feb	0.79	0.59	0.67
Mar	0.83	0.53	0.65

Apr	0.73	0.53	0.62
May	0.74	0.47	0.62
Jun	0.73	0.46	0.62
Jul	0.71	0.47	0.58
Aug	0.73	0.49	0.61
Sep	0.73	0.49	0.64
Oct	0.78	0.53	0.70
Nov	0.75	0.63	0.73
Dec	0.77	0.67	0.74

Table 9. Stage 5 model performance for all years by climatic region, using XGBoost models with 150 trees and a column sampling rate of 0.50.

	R_{Tmin}	R_{Tmean}	R_{Tmax}
Climatic region	RMSE (°C)	RMSE (°C)	RMSE (°C)
Mediterranean	0.71	0.46	0.60
Mod. Mediterranean	0.71	0.50	0.59
Modified Oceanic	0.61	0.35	0.51
Mountain	0.69	0.52	0.65
Oceanic	0.55	0.38	0.48
Other	0.62	0.47	0.54
Semi-Continental	0.67	0.47	0.59
Southwest Basin	0.64	0.39	0.54
Transitional Oceanic	0.63	0.42	0.54

3.4.1 Predictions of the stage 5 model



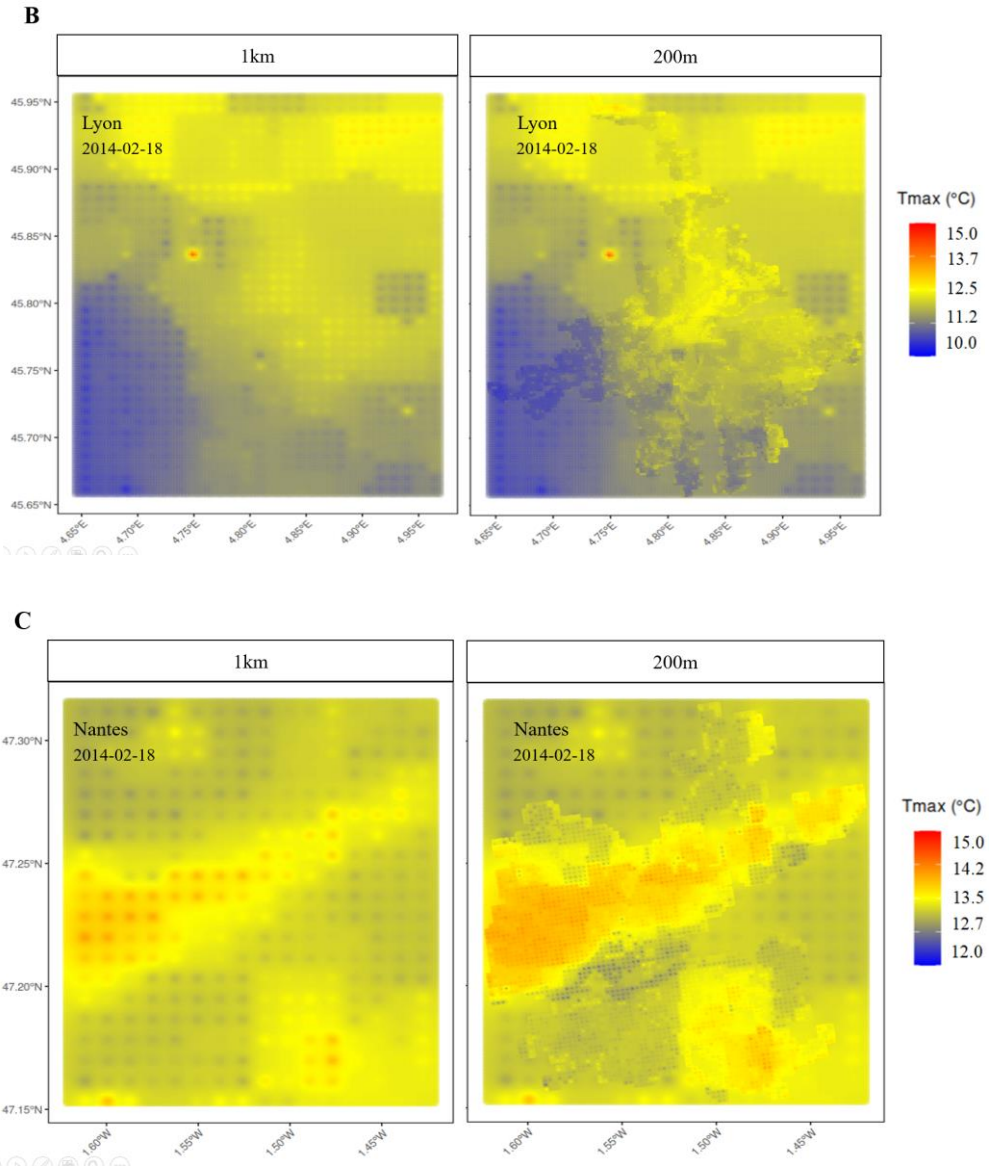


Fig. 16. Predicted 200m T_{max} from Stage 5 models on 18th of February in 2014.

A. Improvement of spatial resolution in T_{max} from 1km (stage 4) to 200m (stage 5) over Paris.
B. Improvement of spatial resolution in T_{max} from 1km (stage 4) to 200m (stage 5) over Lyon.
C. Improvement of spatial resolution in T_{max} from 1km (stage 4) to 200m (stage 5) over Nantes.

3.4.2 Variable importance

The most important predictors in our final 200 m R_{Ta} models are the1km Ta variables predicted in stage 4, each having approximately 30% of importance to their corresponding residual model. Monthly aggregated Landsat LST and NDVI are important also, along with certain land cover types and spatial variables such as population density, elevation, and geographic coordinates (latitude and longitude). Figure 17 summarises these findings.

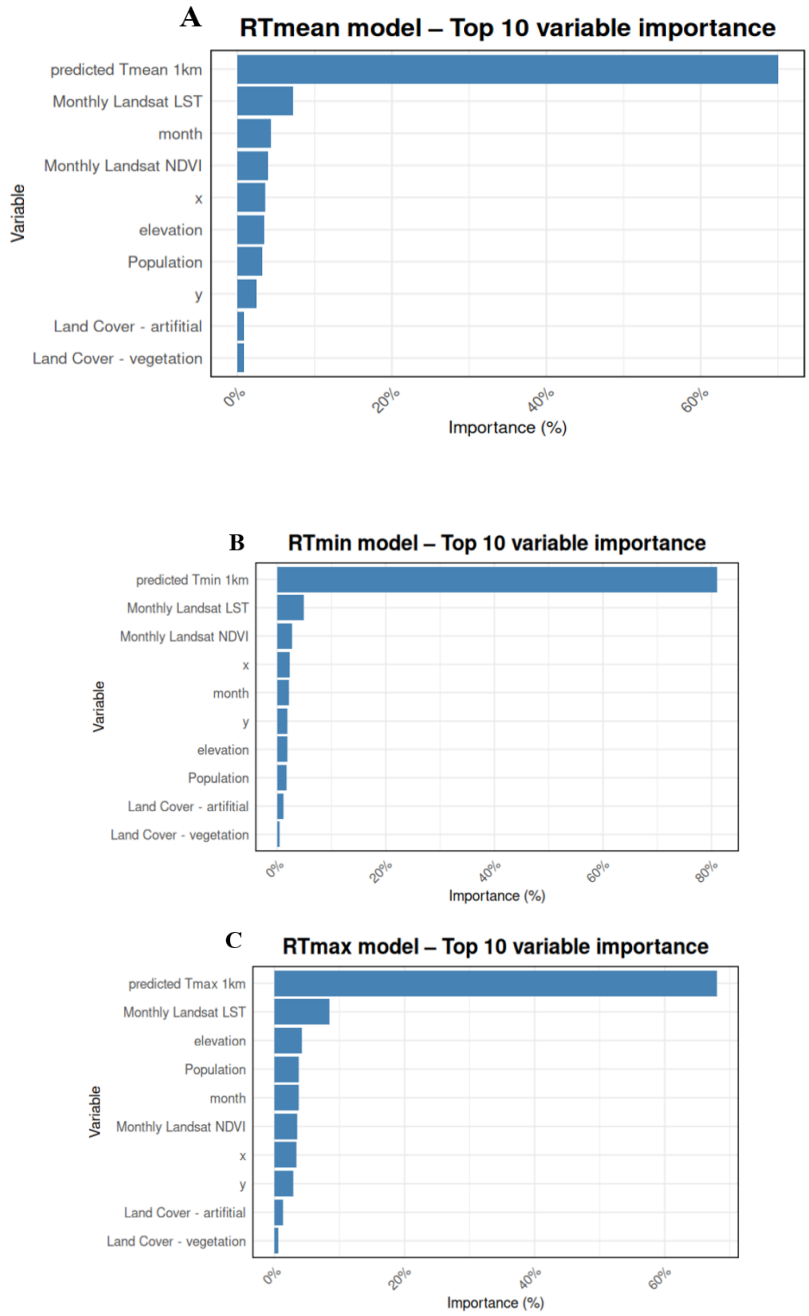


Fig. 17. Variable importance for final 200 R_{Ta} models.

A. Feature importance R_{Tmean} model (top 10 variables).

B. Feature importance R_{Tmin} model (top 10 variables).

C. Feature importance R_{Tmax} model (top 10 variables).

4. Discussion

Comparison with Other Studies

Our methodology demonstrates the potential of hybrid approaches integrating classical ML algorithms with more advanced techniques such as FE, which we consider reflects a new era of data-driven exposure models powered by artificial intelligence (AI). These kinds of combinations notably improve predictive accuracy and model robustness. For instance, Allan et al. (2024) reported RMSE improvements of 29.42% and 5.12% for daily T_{min} and T_{max}, respectively, at a spatial resolution of 1–4 km over the contiguous US (2015–2021), compared to existing Ta models covering the same area (PRISM, gridMET, and Daymet). For France, we compared our results with those of Hough et al. (2020), who predicted Ta for the period 2000–2016. We observed notable improvements, with global average RMSE reductions of 42.66%, 38.11%, and 49.27% for 1 km daily T_{min}, T_{mean}, and T_{max}, respectively, over the period 2014–2024.

In comparison with other recent studies predicting Ta across different territories, our model's performance is consistent with what is expected from models using FE, either directly or indirectly, such as it happens when applying Deep Learning architectures. For instance, while many advanced models achieve average RMSE values close to 1°C, our models outperformed this benchmark for T_{mean} and T_{max}, with average RMSEs of 0.79°C and 0.91°C, respectively. Comparable performance has been reported in other contexts. A study in Hong Kong (Adeniran et al., 2024), using a federated learning artificial neural network (FLANN) to predict diurnal Ta, achieved an RMSE of 0.97°C. Another study in Sweden applied an ensemble metamodel combining four base learners (Random Forest, XGBoost, Generalized Additive Model, and Generalized Additive Mixed Model) to predict daily T_{mean} at 1 km resolution, reporting an RMSE of 1.38°C (Jin et al., 2022). Similarly, Flückiger et al. (2022) reported RMSEs ranging from 0.5°C to 1.86°C for daily T_{min}, T_{mean}, and T_{max} at 100 m spatial resolution in Switzerland.

Insights into Feature Engineering Methods

The exploration of recent literature on environmental exposure modeling reveals that newer studies—even those focused on pollutants like particulate matter 2.5 µm ($PM_{2.5}$) rather than Ta—have started using FE techniques similar to ours. For example, Allan et al. (2025) developed a daily spatiotemporal ML model for average $PM_{2.5}$ in the contiguous US, constructing an IDW FE based on measurements from stations within 500 km to the sites of interest. Their method reduced the mean absolute error (MAE) by 17% compared to existing

models and revealed a stronger link between $PM_{2.5}$ and social vulnerability than previous datasets like FAQSD.

Compared to their method, we believe the selection of nearby monitoring stations for interpolation can be improved. Our study showed that the model's RMSE is sensitive to the number of neighboring monitors used in the IDW interpolation; often, only one or two nearby stations are sufficient. Moreover, choosing a fixed number of nearest monitors for each site—rather than all stations within a fixed radius—helps the model better adapt to spatial variability without being limited by a circular search area. To address this, we optimized the number of nearest stations (k) for the IDW features using a parsimonious approach that selects k based on RMSE minimization (cut-off point rule).

Weaknesses of Modeling Approaches Relying on FE

We highlight a key limitation of approaches that rely on FE. Deriving features from actual monitoring measurements can introduce bias into models' predictions, particularly at sites located far from monitoring stations. This is not the case in our study, as Ta monitors in France are homogeneously distributed across the territory, with high density nationwide. However, in other settings, this issue warrants further investigation to determine whether the inclusion of such features consistently improves model performance and prediction logic. Our results suggest that combining satellite-based models with FE derived from the original target variable (Ta) is most effective in regions with dense monitoring networks. In contrast, for areas with sparse monitoring coverage, we recommend combining predictions from models that use FE features with those that do not—for example, through an ensemble approach such as a linear mixed model that weights predictions based on local performance. This strategy can better leverage the strengths of both model types and improve accuracy across diverse landscapes.

Challenges with the Downscaling Approach

We also encountered challenges when applying the downscaling method used by Hough et al. (2020) using residuals from 1 km stage models as the new target at finer resolution. As the accuracy of the 1 km model improves, residuals become smaller and less variable, leading to unrealistic downscaled predictions with low variance. Instead, we propose directly modeling Ta at higher resolution (e.g., 200 m) by incorporating 1 km model predictions as one of several high-resolution predictors, alongside variables such as Landsat NDVI-LST. We

expect that as monitoring networks expand and ML techniques advance, direct modeling at finer scales will provide more accurate estimates, making residual-based downscaling less favorable. In our case, we aimed to directly compare our results with those of Hough et al. (2020), and therefore applied the same downscaling approach. Using this method, our RMSE for the residuals was higher by 39.39% for R_{Tmean} and 34.21% for R_{Tmax}, but improved by 8.33% for R_{Tmin}.

5. Conclusion

This study introduces a novel five-stage ML methodology for estimating Ta using satellite data from multiple sources. Leveraging advancements in computational power, we enhanced traditional ML algorithms by integrating FE techniques, specifically through distance matrix functions and the KNN model. This approach enables high-resolution Ta estimates across both spatial and temporal dimensions.

We provide a dataset that captures Ta variability through minimum, mean, and maximum daily predictions across continental France. Our methodology includes detailed steps for deriving FEs from raw data, offering a framework that is adaptable to other study regions or environmental variables.

The models showed strong performance based on RMSE metrics over yearly periods, achieving spatial resolution of up to 200 m. By using globally accessible geospatial datasets, our approach ensures broad applicability. While our focus was on France, the methodology can be readily applied in other regions with similar data availability.

We believe our findings will help reduce exposure misclassification in epidemiological research, thereby minimizing bias and improving the reliability of scientific inferences. We encourage the scientific community to adopt and build upon this framework for more robust and scalable exposure modeling.

Appendix 1

Table A1. (Appendix 1) Results for the number of Ta monitors allocated in France from 2014 to 2024. Data were retrieved from MeteoFrance.

Year	Number of T_a monitors
2014	1270

2015	1267
2016	1251
2017	1224
2018	1339
2019	1488
2020	1584
2021	1666
2022	1636
2023	1622
2024	1729

Appendix 2

Distance is calculated with respect to stationary variables (l_1, l_2, l_3, etc.); for example, $l_i = x_p$ (longitude), $l_i = y_p$ (latitude), or $l_i = E_p$ (elevation). Once distances between points are computed, sets of nearest neighbors are determined by selecting the column arguments with the smallest scores.

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2} \text{ or } D_{pj} = \sqrt{(E_p - E_j)^2}, \text{or even}$$

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2 + (E_p - E_j)^2} \quad \forall p, j \in P \qquad \text{Eq. 1}$$

kN_p : Set of first k nearest neighbors points of the point $p \in P$.

$$kN_p = [N_1 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}), N_2 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}) \mid D_{p,N1} \notin \{D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}\}, N_3 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}) \mid D_{p,N1}, D_{p,N2} \notin \{D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}\} \ldots, N_k = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}) \mid D_{p,N1}, D_{p,N2}, D_{p,N3} \ldots, D_{p,NK-1} \notin \{D_{p,1}, D_{p,2}, D_{p,3\ldots} D_{p,|P|}\}]$$

Eq. A2.1

To calculate one of the MODIS LST FEs, we used the following expression:

$$\text{LST nearest neighbors } k \text{ mean}_{p,t} = \sum_{j \in kN_p}^{k} LST_{j,t}/k \quad \forall p \in P, t \in T \qquad \text{Eq. A2.2}$$

Appendix 3

$P_{monitors}$: Set of points with monitors allocated in Israel.

S_i : Denotes a monitor $i \in P_{monitors}$; $P_{monitors} = \{S_1, S_2, S_3, .. S_s\}$; $P_{monitors} \in P$

$$D = \begin{vmatrix} D_{1,S_1} & D_{1,S_2} & \dots & D_{1,S_S} \\ D_{2,S_1} & D_{2,S_2} & \dots & D_{2,S_S} \\ \dots & \dots & \dots & D_{|P|-1,S_S} \\ D_{|P|,S_1} & D_{|P|,S_2} & D_{|P|,S_S-1} & D_{|P|,S_S} \end{vmatrix}$$

kS_p : Set of first k monitors to the point p ∈ P.

$$k < s$$

If a point p is itself a monitor, then we change the value $D_{p=S_i,S_i} = \infty$, thus the point itself is not selected as one of its nearest neighbors.

Annex A

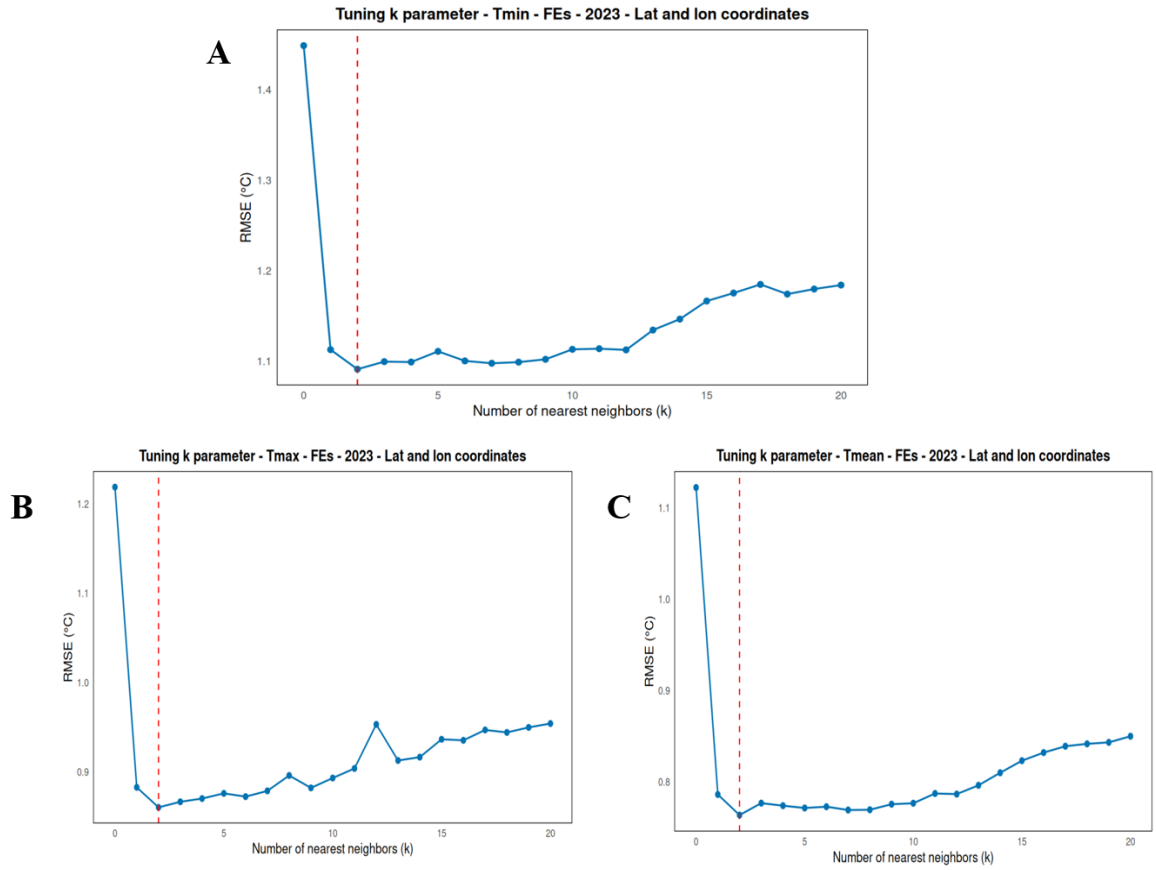


Fig A. (Annex A) FE selection for the model that imputes Ta variables, year 2023.

A. Hyperparameter tuning for k to select FE of T_{min}.

B. Hyperparameter tuning for k to select FE of T_{max}.

C. Hyperparameter tuning for k to select FE of T_{mean}.

ACKNOWLEDGEMENTS

This work was funded by the U.S.-Israel Binational Science Foundation (grant number 2021197) and the French National Research Agency (ANR-15-IDEX-02). Juan David Briceno is supported by a French public grant under the "Investissements d'avenir" program and a scholarship from the Kreitman School for Advanced Graduate Studies at Ben-Gurion University of the Negev. We would like to thank the EUMETSAT Satellite Application Facility on Land Surface Analysis (LSA SAF) for providing SEVIRI LST data, NASA LP DAAC for MODIS LST data, and IMS for remotely sensed weather data.

AUTHOR CONTRIBUTIONS

Juan David Briceno: Data curation; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing.

Itai Kloog: Formal analysis; Methodology; writing – review and editing.

Johanna Lepeule: Formal analysis; methodology; writing – review and editing.

Guillaume Barbalat: writing – Review and editing.

Elena Colicino: Writing – Review and editing.

Massimo Stafoggia: Methodology; conceptualization; writing – review and editing.

Michael Dorman: Data curation; visualisation.

Data Availability Statement

The data supporting the findings of this study are not publicly available due to confidentiality restrictions. Access to the data can be requested from the IAB (Institut pour l'Avancée des Biosciences), UGA, where the data is managed.

Funding information

U.S.-Israel Binational Science Foundation (BSF), Grant/Award Numbers: 2021197; and French National Research Agency in the framework of the "Investissements d'avenir" program, Grant/Award Numbers: ANR-15-IDEX-02 IRGA 2021 PAPex project.

ORCID

Juan David Briceno https://orcid.org/0009-0006-3771-528X

References

Hough, I., Just, A. C., Zhou, B., Dorman, M., Lepeule, J., & Kloog, I. (2020). A multi-resolution air temperature model for France from MODIS and Landsat thermal data. *Environmental Research, 183*, 1–11. https://doi.org/10.1016/j.envres.2020.109244.

Zhou, B., Erell, E., Hough, I., Rosenblatt, J., Just, A. C., Novack, V., & Kloog, I. (2020). Estimating near-surface air temperature across Israel using a machine learning-based hybrid approach. *International Journal of Climatology, 40*, 6106–6121. https://doi.org/10.1002/joc.6570.

Basu, R., Malig, B., & Ostro, B. (2010). High ambient temperature and the risk of preterm delivery. *American Journal of Epidemiology, 172*(10), 1108–1117. https://doi.org/10.1093/aje/kwq170.

Kloog, I., Chudnovsky, A., Koutrakis, P., & Schwartz, J. (2012). Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. *Science of the Total Environment, 432*, 85–92. https://doi.org/10.1016/j.scitotenv.2012.05.095.

Kloog, I., Melly, S. J., Coull, B. A., Nordio, F., & Schwartz, J. D. (2015). Using satellite-based spatiotemporal resolved air temperature exposure to study the association between ambient air temperature and birth outcomes in Massachusetts. *Environmental Health Perspectives, 123*(10), 1053–1058. https://doi.org/10.1289/ehp.1308075.

Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2014). Predicting spatiotemporal mean air temperature using MODIS satellite surface temperature measurements across the Northeastern USA. *Remote Sensing of Environment, 150*, 132–139. https://doi.org/10.1016/j.rse.2014.04.024.

Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., & Schwartz, J. (2017). Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data. *International Journal of Climatology, 37*(1), 296–304. https://doi.org/10.1002/joc.4705.

Hassani, A Hassani, A., Santos, G. S., Schneider, P., & others. (2024). Interpolation, satellite-based machine learning, or meteorological simulation? A comparison analysis for spatio-temporal mapping of mesoscale urban air temperature. Environmental Modeling & Assessment, 29(2), 291–306. https://doi.org/10.1007/s10666-023-09943-9

Ochsenwald, W. L., Sicherman, H., et al. (2019). Israel. *Encyclopaedia Britannica*. Encyclopaedia Britannica, Inc.

Rosenfeld, A., Dorman, M., Schwartz, J., Novack, V., Just, A. C., & Kloog, I. (2017). Estimating daily minimum, maximum, and mean near-surface air temperature using hybrid satellite models across Israel. *Environmental Research, 159*, 297–312. https://doi.org/10.1016/j.envres.2017.08.017.

Schwartz, J., Samet, J. M., & Patz, J. A. (2004). Hospital admissions for heart disease. *Epidemiology, 15*(6), 755–761. https://doi.org/10.1097/01.ede.0000134875.15919.0f.

Keatinge, W. R. (1997). Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. *The Lancet, 349*(9062), 1341–1346. https://doi.org/10.1016/S0140-6736(96)12338-2.

Green, M., Pri-or, N., Capeluto, G., Epstein, Y., & Paz, S. (2013). Climate change and health in Israel: Adaptation policies for extreme weather events. *Israel Journal of Health Policy Research, 2,* 23. https://doi.org/10.1186/2045-4015-2-23.

Dudhia, J. (2008). Fundamental issues in numerical weather prediction. In *Large-Scale Disasters Prediction, Control, and Mitigation* (pp. 447-452). Cambridge University Press. https://doi.org/10.1017/CBO9780511535963.021.

Israel Meteorological Service (IMS). (n.d.). *Israel Meteorological Service website*. Israeli Ministry of Transport and Road Safety. https://ims.gov.il/en (accessed 20 February 2022).

Central Bureau of Statistics. (2018). *Statistical Abstract of Israel 2018*.

Tøttrup, C. (2014). *EU-DEM Statistical Validation Report*.

Swathi, N., Bhat, M., Reddy, N., & Rao, A. (2022). Study of distance metrics on k-nearest neighbor algorithm for star categorization. *Journal of Physics: Conference Series, 2161*, 012004. https://doi.org/10.1088/1742-6596/2161/1/012004.

Zhang, T., Zhou, Y., Zhao, K., Zhu, Z., Chen, G., Hu, J., & Wang, L. (2022). A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003–2020). *Earth System Science Data, 14*, 5637–5649. https://doi.org/10.5194/essd-14-5637-2022.

Zhou, K., Liu, H., Deng, X., Wang, H., Zhang, S., & Bonafoni, S. (2020). Comparison of machine-learning algorithms for near-surface air-temperature estimation from FY-4A AGRI data. *Advances in Meteorology*. https://doi.org/10.1155/2020/8887364.

Kohli, S., Godwin, G. T., & Urolagin, S. (2021). Sales prediction using linear and KNN regression. In S. Patnaik, X. S. Yang, & I. Sethi (Eds.), *Advances in Machine Learning and Computational Intelligence: Algorithms for Intelligent Systems*. Springer, Singapore. https://doi.org/10.1007/978-981-15-5243-4_29.

Kaggle. (2018). *Home Credit Default Risk*. https://www.kaggle.com/c/home-credit-default-risk (accessed 18 May 2022).

Kaggle. (2020). *Fashion MNIST-CNN Beginner (98% Accuracy)*. https://www.kaggle.com/general/207202 (accessed 19 May 2022).

Mukid, M. A., Widiharih, T., Rusgiyono, A., & Prahutama, A. (2018). Credit scoring analysis using weighted k-nearest neighbor. *Journal of Physics*. https://iopscience.iop.org/article/10.1088/1742-6596/1025/1/012114/pdf.

Kaggle. (2018). *Can you predict how capable each applicant is of repaying a loan?* https://www.kaggle.com/c/home-credit-default-risk (accessed 18 May 2022).

De Blesser, P., Hooghe, B., Vlieghe, D., & Van Roy, F. (2007). A distance difference matrix approach to identifying transcription factors that regulate differential gene expression. *Genome Biology*, *8*(5), R83. https://doi.org/10.1186/gb-2007-8-5-r83.

Watson, G., Telesca, D., Reid, C., Pfister, G., & Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution, 254*, 112792. https://doi.org/10.1016/j.envpol.2019.06.088.

Dirksen, M., Knap, W., Steeneveld, G., Holtslag, A., & Klein Tank, A. (2020). Downscaling daily air-temperature measurements in the Netherlands. *Theoretical and Applied Climatology, 142*, 751–767. https://doi.org/10.1007/s00704-020-03313-1.

Wang, P., Mao, K., Meng, F., Qin, Z., Shu, F., & Bateni, S. (2022). A daily highest air temperature estimation method and spatial–temporal changes analysis of high temperature in China from 1979 to 2018. *Geoscientific Model Development*, *15*, 6059–6083. https://doi.org/10.5194/gmd-15-6059-2022.

Li, X., Zhou, Y., Asrar, G., & Zhu, Z. (2018). Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.201

Just, A. C., Arfer, K. B., Rush, J., & Kloog, I. (2025). XIS-temperature: A daily spatiotemporal machine-learning model for air temperature in the contiguous United States. Environmental Research, 270, 120731. *https://doi.org/10.1016/j.envres.2024.120731*

Jin, Z., Ma, Y., Chu, L., Liu, Y., Dubrow, R., & Chen, K. (2022). Predicting spatiotemporally-resolved mean air temperature over Sweden from satellite data using an ensemble model. Environmental Research, 204, 111960. https://doi.org/10.1016/j.envres.2021.111960

4.3 Article 3

Full reference:

Briceno, J. D., Alderson, J., Barbalat, G., Colicino, E., Stafoggia, M., Lepeule, J., & Kloog, I. (2025). A new hybrid approach applying feature engineering to model high-resolution particulate matter across Israel.

Status: The article was completed in September 2025 and is currently finalized, awaiting approval from the thesis directors for submission.

PhD candidate's contribution:
The PhD candidate designed the study, downloaded and processed the data, performed the statistical analyses, interpreted the results, and wrote the manuscript. The candidate was the main author of the paper.

(Article 3 in its complete version follows on the next pages. Tables, figures, and equations are numbered as in its last updated version and have not been modified.)

A new hybrid approach applying Feature Engineering to model high resolution particle matter across Israel

Juan David Briceno, Jakob Alderson, Guillame Barbalat, Elena Colicino, Massimo Stafoggia, Johanna Lepeule, Itai Kloog

A new hybrid approach applying Feature Engineering to model high resolution particle matter across Israel

Juan David Briceno[a,b*], Jakob Alderson[b], Guillame Barbalat[a], Elena Colicino[d], Massimo Stafoggia[c], Johanna Lepeule[a1], Itai Kloog[b,d1*]

[a] University Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Development and Respiratory Health, Institute for Advanced Biosciences (IAB), La Tronche, France.
[b] Department of Environmental, Geoinformatics and Urban Planning Sciences, Ben-Gurion University of the Negev, P.O.B. 653, Beer Sheva, Israel
[c] Lazio Regional Health Service, Department of Epidemiology, Unit of Environmental and Occupational Epidemiology and Cancer Registry, Roma, Italy
[d] Department of Environmental Medicine and Climate Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA

* Corresponding author. Institute for Advanced Biosciences (IAB), Site Santé – Allée des Alpes, 38700 La Tronche, France. E-mail: juan.briceno-guerrero@univ-grenoble-alpes.fr, Tel: +33768296374

[1] these authors contributed equally

Abstract

Exposure to particulate matter (PM_{10} and $PM_{2.5}$) remains a major public health concern in Israel due to its geographic location, complex geo-climatic conditions, industrial activities, and diverse anthropogenic sources, underscoring the need for accurate and continuous air-quality modeling. This study introduces an innovative framework for estimating daily PM_{10} and $PM_{2.5}$ concentrations at a 1-km resolution across Israel by implementing a novel feature engineering (FE) approach. Using a k-nearest neighbors (KNN) model, we developed new spatio-temporal predictors and enhanced the imputation of MODIS aerosol optical depth (AOD), generating continuous AOD fields that exploit the correlation between AOD and particulate matter. We then calibrated predictive models that incorporated spatial and spatio-temporal variables, ground-based PM observations, satellite data, and FE-based predictors to estimate daily PM concentrations. The models achieved strong cross-validated performance, with RMSE values ranging from 10.29 to 26.97 µg/m³ for PM_{10} and from 2.73 to 13.96 µg/m³ for $PM_{2.5}$. This methodology advances exposure assessment by providing high-resolution PM estimates to support epidemiological research and environmental policy planning in regions with complex geographic characteristics.

Keyword

Particle matter; Aerosol optical depth; MODIS AOD; Feature Engineering; K-Nearest Neighbors; Machine Learning; Extreme Gradient Boosting

1. Introduction

Air pollution, particularly particulate matter (PM_{10} and $PM_{2.5}$), remains one of the most critical environmental health challenges worldwide. Exposure to elevated PM levels is associated with respiratory and cardiovascular diseases, adverse pregnancy outcomes, and increased mortality (Shtein et al., 2018). These effects are driven by the particles' ability to penetrate deep into the respiratory system and enter the bloodstream, amplifying systemic inflammation and oxidative stress. In Israel, the issue is further exacerbated by the country's diverse geo-climatic conditions, transportation and industrial patterns, and frequent natural dust events, all of which contribute to complex and highly variable particulate matter profiles across space and time.

To accurately assess PM-related health risks, it is essential to develop models that provide high-resolution spatio-temporal PM predictions. Such models enable the identification of critical exposure windows, support nationwide epidemiological studies, and inform policy measures aimed at reducing air pollution impacts on vulnerable populations. However, ground-based monitoring networks in Israel are sparse and unevenly distributed, resulting in discontinuous exposure assessment that introduces bias and limits the reliability of epidemiological outcomes. Reliance on a limited number of fixed monitors fails to capture localized variations in PM concentrations, particularly in regions with diverse land use, topography, and meteorological conditions. For instance, Israel's PM levels are strongly influenced by external sources, such as dust transported from the Sahara Desert and the Arabian Peninsula toward the Negev, where the low density of monitoring stations may compromise the quality of health-related studies (Kloog et al., 2015).

Satellite-derived data products have become indispensable for bridging these gaps. AOD from the Moderate Resolution Imaging Spectroradiometer (MODIS), especially the high-resolution Multi-Angle Implementation of Atmospheric Correction (MAIAC) retrievals, has demonstrated considerable potential in estimating PM concentrations (Kloog et al., 2015; Shtein et al., 2018; Hough et al., 2021; Wang et al., 2025). Nonetheless, AOD products are prone to missing observations due to cloud cover, high surface reflectance over bright surfaces, and sensor degradation (Kloog et al., 2015). Overcoming these limitations requires novel modeling approaches for gap-filling, such as leveraging spatio-temporal predictors

derived from historical data patterns to improve estimation accuracy generating realistic imputations.

This study introduces several innovations based on feature engineering (FE) for PM prediction in Israel. Advanced techniques based on the KNN model were employed to derive novel spatio-temporal predictors enhancing both the imputation of MODIS AOD and the prediction of PM, leading to more accurate, high-resolution continuous PM fields. These newly derived predictors were combined with ground-based observations and satellite products to generate daily estimates at a 1 × 1 km resolution across Israel. This approach represents a significant advancement in exposure modeling, enabling more reliable epidemiological analyses and providing policymakers with scientifically grounded information to support evidence-based decisions.

2. Data and methods

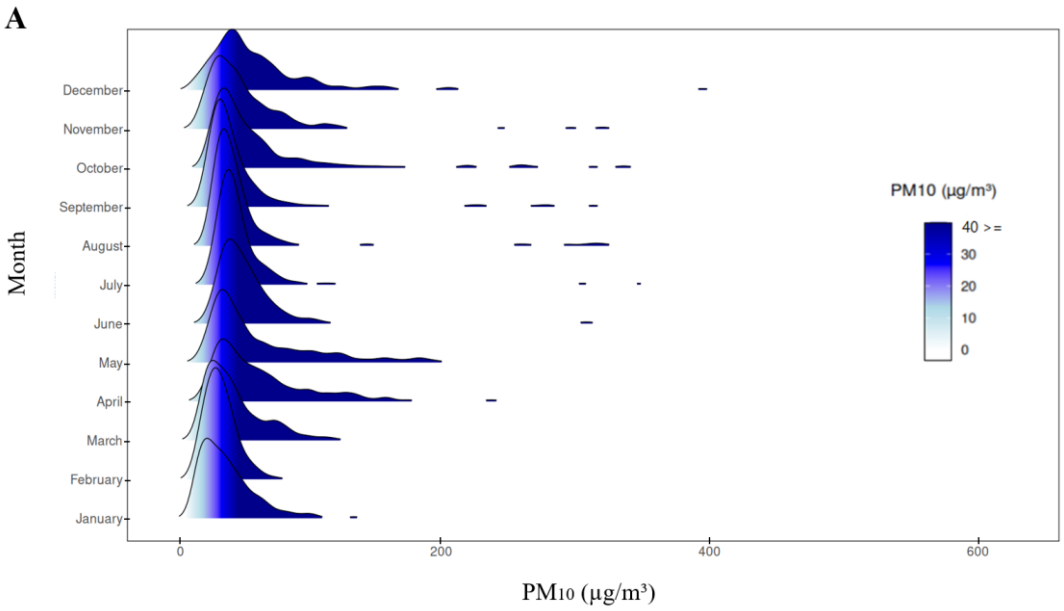
2.1 Study area

This study focuses on the State of Israel, a relatively small but environmentally diverse country that spans approximately 21,670 km². Israel stretches about 470 km from north to south and up to 135 km at its widest point, with dramatic elevation changes from Mount Hermon (2,807 m above sea level) to the Dead Sea (around 430 m below sea level). The country encompasses four main geographical regions (Zhou et al., 2020): the Mediterranean coastal plain, the Central Hills, the Jordan Rift Valley, and the Negev Desert. Each of these regions exhibits distinctive meteorological and land-use characteristics that directly affect PM dynamics. For example, the densely populated coastal plain is dominated by urban and industrial activity, while the Negev Desert features sparse settlements and extensive dry landscapes. This spatial heterogeneity, combined with the country's sharp climatic gradients, creates a unique setting for studying variations in air quality at fine scales.

Particulate matter concentrations across Israel reflect a complex interplay between natural and anthropogenic sources. Major industrial and energy hubs such as the Ne'ot Hovav (Ramat Hovav) industrial complex in the south, the Ashdod Oil Refinery on the central coast, and the Barkan and Shahak industrial parks contribute combustion and process-related PM emissions alongside precursor gases for secondary aerosols (Cohen, 2011). The densest highway and road networks occur in the Tel Aviv metropolitan area and northern coastal corridor, where heavy commuter traffic, freight movement, and urban expansion combine to generate

substantial traffic-related emissions and resuspended road dust (OECD, 2023). Ben Gurion International Airport near Tel Aviv is also a significant localized source of particulate emissions from aircraft operations and ground-support activities (Israel Airports Authority, n.d.). In addition to these anthropogenic inputs, natural dust intrusions regularly affect Israel, with the Negev Desert and the Jordan Rift Valley serving as principal entry points for mineral dust transported from the Sahara and Arabian Peninsula during spring and autumn (Kishcha et al., 2016). These events cause sudden spikes in coarse particulate matter (PM_{10}), elevating concentrations above typical levels, whereas fine particulate matter ($PM_{2.5}$) is primarily driven by anthropogenic emissions under normal conditions, particularly during stagnant summer periods with limited atmospheric mixing (Shtein et al., 2018).

Figure 1 illustrates the distribution of PM across Israel during 2024. PM_{10} values above 70–90 µg/m³ are uncommon (Achilleos et al., 2020) and generally occur due to environmental phenomena such as dust storms.



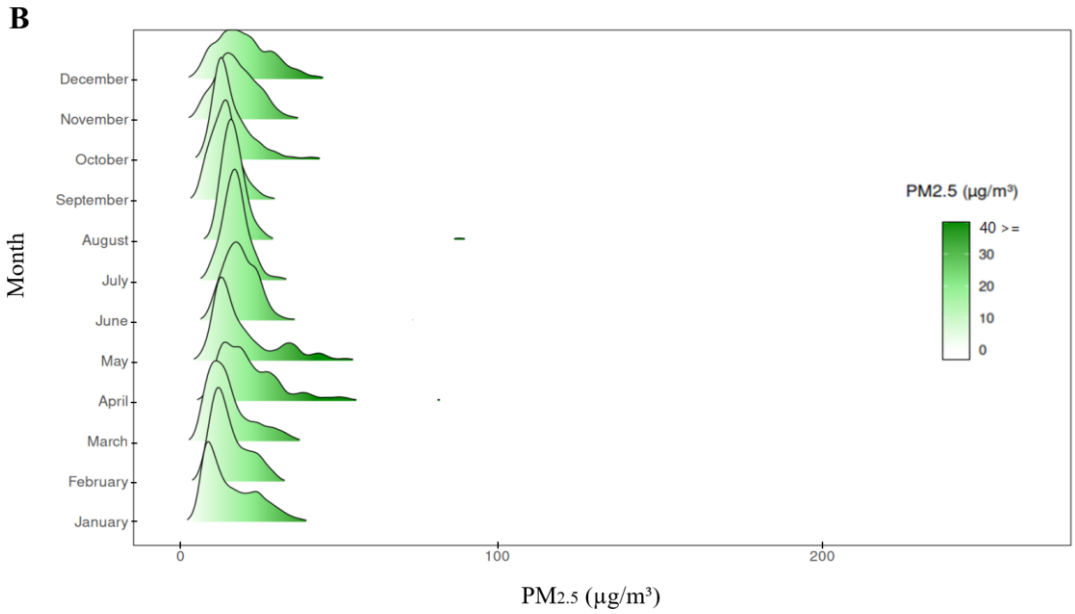


Fig 1. Distribution of monitored particulate matter (TCEEH) in Israel during 2024.

A. Monthly distribution of PM_{10} across the year.

B. Monthly distribution of $PM_{2.5}$ across the year.

Regional analysis revealed that the highest average concentrations of PM_{10} and $PM_{2.5}$ occur in the Mediterranean Coastal Plain, where levels reach up to 72 µg/m³ for PM_{10} and 19 µg/m³ for $PM_{2.5}$. In contrast, the Central Hills exhibit the lowest average concentrations, at 39 µg/m³ for PM_{10} and 15 µg/m³ for $PM_{2.5}$ (Fig. 2). Concentration spikes of up to 370 µg/m³ for PM_{10} and 77 µg/m³ for $PM_{2.5}$ were observed across all regions, exceeding the recommended safe environmental thresholds of 50 µg/m³ for PM_{10} and 25 µg/m³ for $PM_{2.5}$. These exceedances indicate the presence of potentially hazardous exposure levels for human health in this country (WHO, 2021).

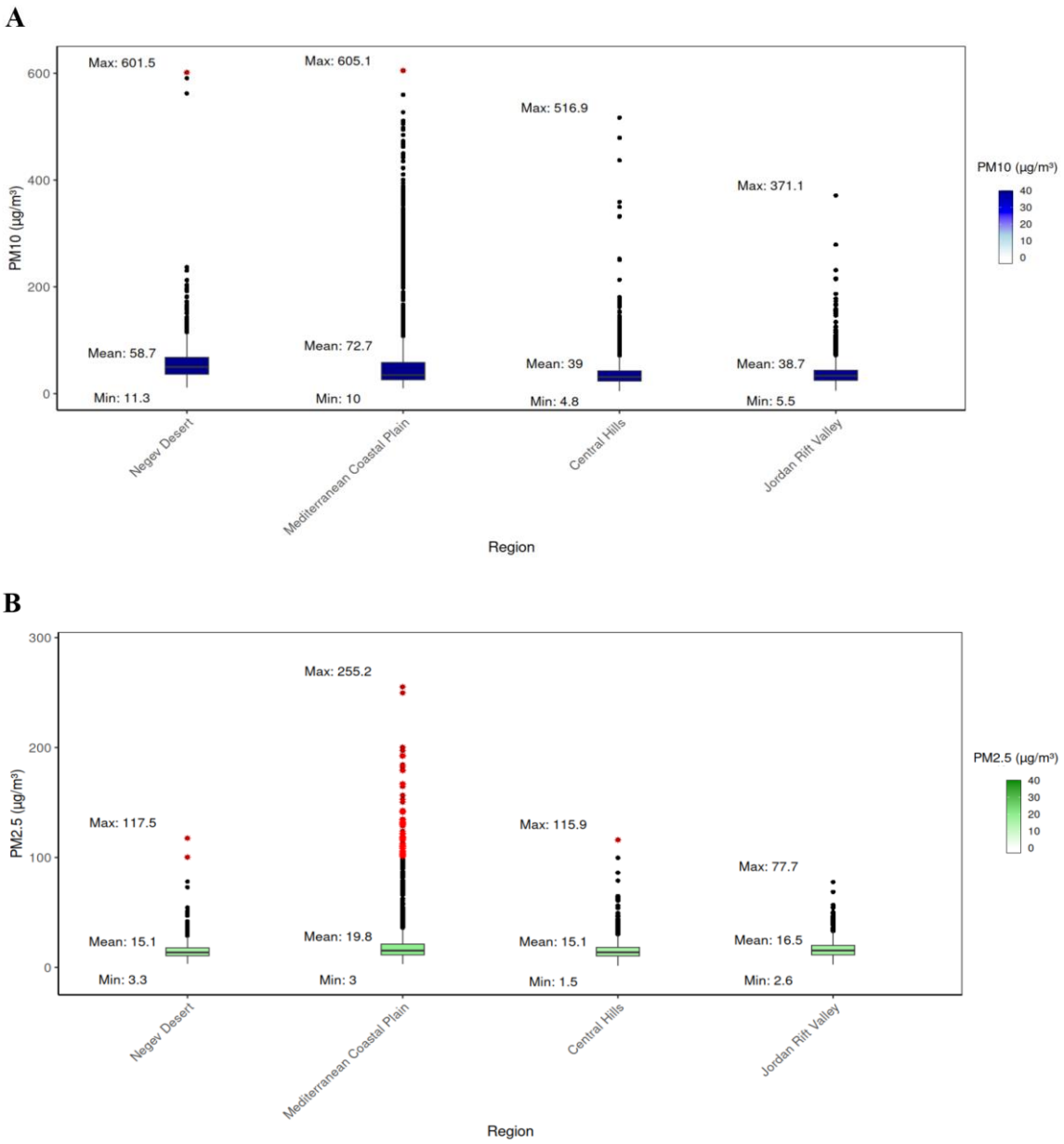


Fig. 2. Particulate matter concentrations in Israel (TCEEH) per region during 2024.

A. Regional distribution of PM_{10} throughout the year.

B. Regional distribution of $PM_{2.5}$ throughout the year.

2.2 Particle matter data

PM levels in Israel are assessed using data from the nationwide air quality monitoring network. Not all monitors measure both PM_{10} and $PM_{2.5}$, and the number of operational monitors varies over time due to maintenance, failures, and the installation of new devices. The network is concentrated primarily in densely populated coastal regions and major urban centers, including Tel Aviv, Jerusalem, Haifa, Ashdod, and Beer-Sheva (see Fig. 3). Data is

managed and provided by the Technion Center of Excellence in Exposure Science and Environmental Health (TCEEH), with measurements available at half-hour intervals. For this study, we retrieved data for the years 2003–2024 and calculated daily averages from the half-hourly observations.

Air quality monitoring in Israel is conducted by various local and national agencies under the supervision of the Ministry of Environmental Protection, following standardized procedures. Particulate matter is measured using Tapered-Element Oscillating Microbalance (TEOM) instruments, maintained according to US-EPA protocols, with an estimated accuracy of approximately ±5% (Shtein et al., 2018). To ensure data reliability, only monitors with sufficient continuous records were used. For 2024, there were 24 monitors measuring $PM_{2.5}$, 7 monitors measuring PM_{10}, and 10 monitors measuring both $PM_{2.5}$ and PM_{10} in Israel.

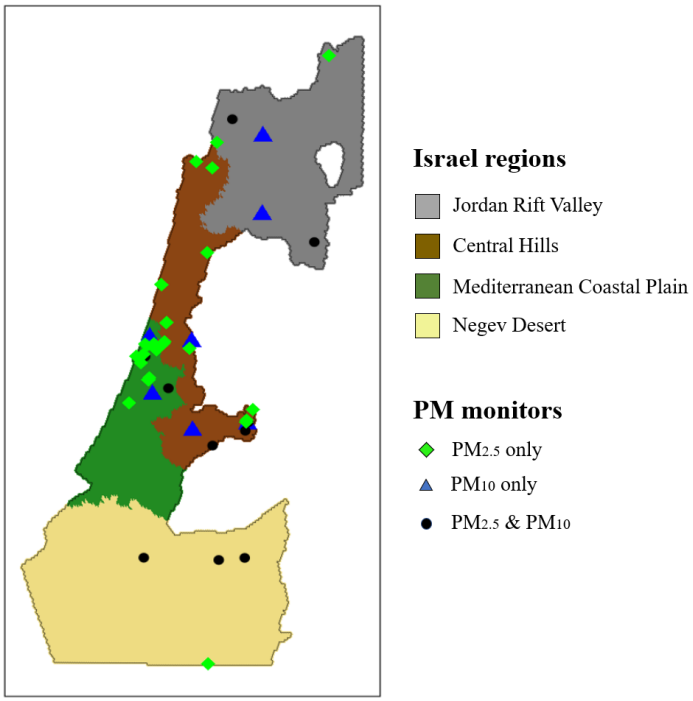


Fig. 3. Distribution of 41 PM monitors across Israel in 2024. Four regions from Zhou et al. (2020) are also shown.

2.3 Satellite derived Aerosol Optical Depth

MODIS satellites (Aqua and Terra) provide ~1 km² resolution AOD measurements, with each satellite following distinct polar orbits resulting in roughly one observation per day per satellite. Combining Aqua and Terra yields approximately two daily observations. Although

differences in overpass times may introduce slight temporal mismatches with ground measurements, the combined datasets provide complementary information. Data gaps may occur due to clouds, surface brightness, or retrieval limitations. As in previous studies, we applied the latest version of the MAIAC algorithm (Lyapustin et al., 2011a,b; Arvani et al., 2016; de Hoogh et al., 2018) to enhance accuracy over bright and desert surfaces. MAIAC incorporates bidirectional reflectance modeling and uses up to 16 days of measurements for temporal smoothing while providing daily AOD estimates. MAIAC AOD data underwent quality assurance to exclude pixels flagged as unreliable. Previous studies demonstrated a strong correlation between MAIAC AOD and ground-based AERONET AOD ($r = 0.85$; Kloog et al., 2015), confirming the reliability of the satellite product for capturing columnar aerosol load. This high agreement indicates that satellite AOD provides a robust predictor in statistical models for surface PM concentrations. Consistent with previous work, we used the AOD 470 nm band from MODIS AOD for our analysis.

2.4 Spatial and temporal predictors of Particle matter

Numerous spatial and spatio-temporal predictors were incorporated into our hybrid models for estimating PM concentrations. These predictors were retrieved for the full study period (2003–2024). The following sections describe each dataset in detail, and Table 1 provides a summary specifying source, resolution, units and others.

2.4.1 Spatio-temporal predictors

2.4.1.1 Reanalysis data: CAMS AOD

We used AOD from the Copernicus Atmosphere Monitoring Service (CAMS) Reanalysis, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). This reanalysis combines satellite observations, ground-based measurements, and atmospheric modeling within an Integrated Forecasting System (IFS) to provide consistent global estimates of atmospheric composition. To capture the spectral variability of aerosols, we retrieved AOD at 469, 550, 670, 865, and 1240 nm. We also retrieved dust AOD to complement our analysis. The shorter wavelengths (469, and 550nm) are particularly sensitive to fine-mode and anthropogenic aerosols, while the intermediate (670nm) and near-infrared (865nm) bands help characterize the transition from fine to coarse particles. The 1240 nm band and the dedicated dust AOD product provide enhanced sensitivity to coarse mineral aerosols, such as desert dust, and were specifically used to identify days affected by dust storms, improving discrimination over bright and arid surface. Together, these multi-

wavelength AOD datasets enable a more robust representation of aerosol optical properties across particle sizes, which is critical for estimating PM. CAMS Reanalysis provides data at a spatial resolution of approximately 0.75° (~80 km) with a 3-hourly temporal resolution, covering the period 2003–2024. All datasets were downloaded from the Copernicus Atmosphere Data Store (https://ads.atmosphere.copernicus.eu; accessed March 1, 2025). Standard quality assurance procedures were applied, and grid cells flagged as unreliable were excluded prior to further analysis.

2.4.1.2 MODIS Vegetation indices

The Normalized Difference Vegetation Index (NDVI) from the MODIS Aqua satellite was employed to characterize vegetation dynamics across the study area. NDVI is derived from atmospherically-corrected surface reflectance in the blue (469 nm), red (645 nm), and near-infrared (858 nm) bands, providing a standardized measure of vegetation greenness and density (NASA, n.d.). Sixteen-day composite NDVI products at a spatial resolution of 250 m were retrieved and subsequently averaged to 1 km to match the spatial resolution of the study grids. For daily PM-AOD modeling, the same NDVI value was applied to all days within each 16-day period. The Enhanced Vegetation Index (EVI) was also included as it improves sensitivity over dense vegetation and reduces effects from canopy background and atmospheric contamination. Vegetation indices such as NDVI and EVI are relevant for particulate matter modeling because they influence local air quality through mechanisms including dust suppression, deposition of airborne particles, and modification of boundary layer dynamics (Zhou, M., et al., 2021; Zhou, X., et al., 2024).

2.4.1.3 Wind Speed and Direction

Hourly wind speed and direction data were retrieved from the MERRA-2 reanalysis, produced by NASA's Global Modeling and Assimilation Office. The dataset provides global estimates of atmospheric variables at a spatial resolution of approximately 0.5° (~50 km). Wind speed and direction are derived from assimilating satellite observations, surface measurements, and model forecasts to produce consistent hourly fields. For this study, hourly values were averaged to obtain daily estimates at 50 km resolution. These variables are important for PM

modeling because they influence the transport and dispersion of aerosols, affecting local PM concentrations.

2.4.1.4 Dust storm classification

To capture the occurrence of dust storm events in the region, all observations throughout the study period were classified as either affected by dust (dust flag = 1) or not affected by dust (dust flag = 0). A decision tree classifier and then a logistic regression model were applied to extend this classification to all grid points and dates. The daily dust storm flag was assigned following an extended scheme based on Achilleos et al. (2020).

2.4.2 Spatial predictors

2.4.2.1 Elevation

We retrieved elevation data from the ASTER Global Digital Elevation Model (GDEM) (ASTER, 2014), which provides near-global coverage from 83°N to 83°S. The dataset has a spatial resolution of 30 m, capturing fine-scale variations in terrain. For the purposes of this study, elevation values were averaged to a 1 km resolution to match the modeling grids. Due to substantial elevation differences across the study area, elevation was included as a spatial predictor in our PM modeling framework.

2.4.2.2 Population density

We also included population density data from the Israeli Central Bureau of Statistics (ICBS, 2014). To align with the modeling grids, population values were aggregated by calculating the population-weighted average for each 1 km grid cell.

2.4.2.3 Distance from main roads, highways, and cities

To account for anthropogenic transportation effects on PM, the distance from each grid cell centroid to major roads, highways, and cities was included as a predictor. Distances were calculated in meters.

2.4.2.4 Length all roads and highways

A total road density raster for each grid cell was generated following the approach of Shtein et al. (2018), using the line density tool to calculate the density of polyline features (all roads, and separately for highways) within a 1-km radius around each cell centroid.

2.4.2.5 Land use, Israel's regions and natural areas

Land use, administrative regions, and natural areas were included as spatial predictors in the models. For each point in the grid, land use describes functional characteristics of the surface, such as residential, industrial, or agricultural areas, which can influence pollutant dispersion. Israel's regions classify the country into administrative zones, providing contextual information for modeling temperature and particulate matter at each grid location. Natural areas identify protected or natural zones that may affect the local microclimate and the dispersion of pollutants at each grid point.

Table 1. A summary of the time-invariant and time-variant covariates employed in this study

Name	Source	Temporal Resolution	Spatial Resolution	Units	Nature	Stages Used
Elevation	GDEM	-	30 m	m	Time-invariant	1–6
Population density	ICBS	-	Point data	people/km²	Time-invariant	1–6
Distance from main roads	GISrael – MAPA	-	Point data	m	Time-invariant	3–6
Distance from highways	GISrael – MAPA	-	Point data	m	Time-invariant	3–6
Distance from cities	GISrael – MAPA	-	Point data	m	Time-invariant	3–6
Length all roads	GISrael – MAPA	-	Point data	km/km²	Time-invariant	3–6
Length highways	GISrael – MAPA	-	Point data	km/km²	Time-invariant	3–6
Land-use	ICBS	-	Point data	%	Time-invariant	3–6
Israel regions	GIS Maps of Israel – Resources	-	Polygons	-	Time-invariant	3–6
Natural areas	GIS Maps of Israel – Resources	-	Polygons	-	Time-invariant	3–6
MODIS NDVI	NASA	16 days	250 m	-	Time-variant	1–6
MODIS EVI	NASA	16 days	250 m	-	Time-variant	1–6
MODIS AOD (470 nm)	NASA	Daily	~1 km	-	Time-variant	1–6
CAMS AOD (469 nm)	ECMWF	Hourly	80 km	-	Time-variant	1–6

CAMS AOD (550 nm)	ECMWF	Hourly	80 km	-	Time-variant	1–6
CAMS AOD (670 nm)	ECMWF	Hourly	80 km	-	Time-variant	1–6
CAMS AOD (865 nm)	ECMWF	Hourly	80 km	-	Time-variant	1–6
CAMS AOD (1240 nm)	ECMWF	Hourly	80 km	-	Time-variant	1–6
CAMS Dust AOD	ECMWF	Hourly	80 km	-	Time-variant	3–6
Wind speed	NASA	Hourly	50 km	m/s	Time-variant	3–6
Wind direction	NASA	Hourly	50 km	°	Time-variant	3–6
Dust storm classification	Achilleos et al., 2020	Daily	1 km	-	Time-variant	3–6

2.5 Statistical methods

To model particulate matter (PM), we developed a six-stage machine learning (ML) approach, where stages three to four and five to six are treated independently. This approach integrates methodologies from previous studies (Kloog et al., 2015; Shtein et al., 2018; Hough et al., 2021) and introduces a novel FE technique (see Figure 4). The method generates geographically weighted predictors for MODIS AOD and monitored PM, allowing us to capture spatio-temporal patterns while preserving the positional integrity of specific geographic points and their nearest neighbors, which is often lost when transforming raster images into tabular data for training and randomization. Incorporating these new predictors significantly improves the accuracy of daily 1km PM models across Israel.

Stages one and two focused on imputing MODIS AOD data affected by missing values. Following stage two, two independent modeling paths were conducted using the gap-filled AOD outputs. Stages three and four were used to predict PM_{10}, while stages five and six were used to predict $PM_{2.5}$. We modeled PM_{10} and $PM_{2.5}$ separately to prevent error propagation between stages, while ensuring that in the final stage the expected relationship $PM_{10} > PM_{2.5}$ was preserved. For all stages of the modeling, extreme-gradient-boosting (XGBoost) models were used. In addition, we applied basic feature engineering to selected spatial and spatio-temporal predictors to generate and transform covariates, thereby slightly improving the representation of our models.

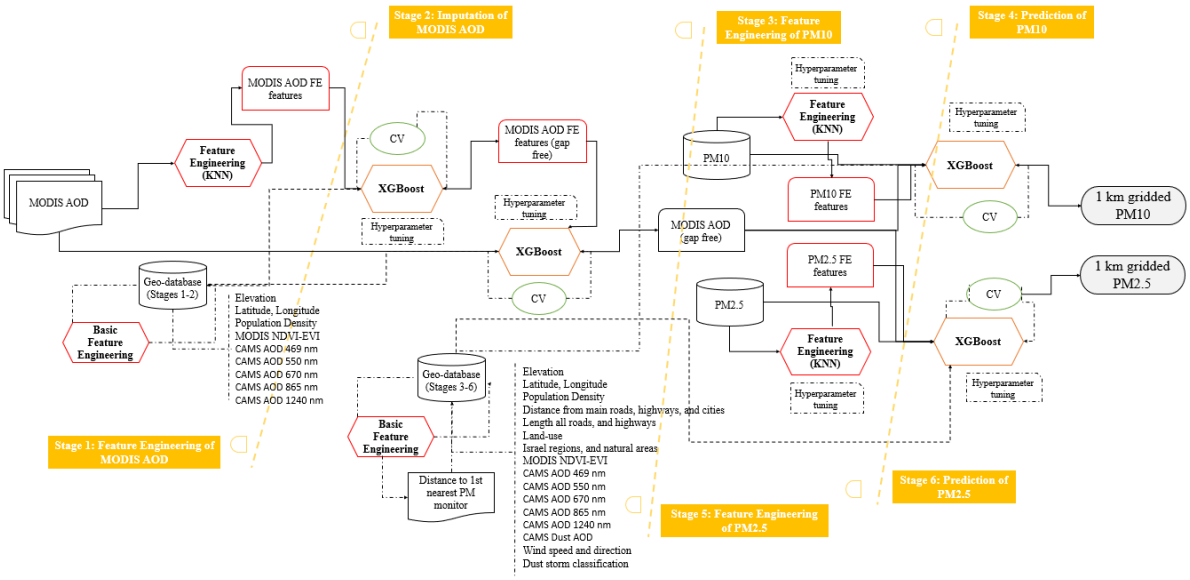


Fig 4. Schematic diagram of the six-stage model for estimating PM_{10} and $PM_{2.5}$ in Israel.

Modeling Satellite-Based Aerosol Optical Depth

Stage 1 model: Feature Engineering of MODIS AOD

In Stage 1, we performed gap-filling of the MODIS 470 nm AOD band. To begin, we produced a functional set of predictors from the original spatial and spatio-temporal predictors selected for this stage (see Table 1). These new features, derived through basic transformations and calculations (feature engineering), include encoded representations of calendar date variables with cyclical properties (sine and cosine), local and regional MODIS AOD patterns, CAMS AOD spectral differences, and measures of variability and relationships across different AOD wavelengths. Table A1 in the Appendix briefly defines and explains these variables. When missing values occurred, we imputed them using simple averages.

We then applied more advanced feature engineering using the KNN model, a non-parametric supervised machine learning method. KNN operates by retaining all available data and predicting missing values based on the most similar existing data points, identified through a selected distance metric (Swathi et al., 2022).

We defined a grid representing Israel with P points. For each point $p \in P$, we identified a corresponding set of neighboring points $(i, j, g, \ldots l \in P \mid i, j, g, \ldots l \neq p)$ sharing similar topographic properties such as latitude and longitude. In general, similarity in this algorithm is determined by geographical spatial attributes like longitude (x_p) and latitude (y_p). Although

other variables, such as elevation (E_p) or population density (Pop_p), could be used, in this study we restrict our analysis to the first two spatial coordinates (see Equation 1). Distances between grid points are arranged into a distance matrix D. For each point p, we identify a set of k nearest neighbors (kNp). Then, for each point p and date t, we compute the average AOD values of its neighbors, resulting in a new feature that provides an estimate of AOD across different locations and times.

Because these feature-engineered predictors (FEs) can be susceptible to missing values, due to absent neighboring AOD observations at a point p, we address gaps using a model-based approach. To avoid confusion, we clarify that we train XGBoost models (1000 trees, column sampling rate of 0.45) using all other available spatial and spatio-temporal predictors to estimate values at missing locations. No cross-validation or hyperparameter tuning is performed at this imputation step. XGBoost was chosen to maintain consistency across the analysis. Specifically, missing values (NAs) in FEs were replaced with model-based predictions of the form FE ~ f(latitude, longitude, elevation, …, CAMS AOD 469 nm, CAMS AOD 1,240 nm, …). MODIS AOD was excluded as a predictor since it is the target variable being imputed at this stage. In practice, the number of missing values in the FEs is always lower than in the original AOD product and decreases as more neighboring points are incorporated into the calculation. The optimal number of neighbors (k) was subsequently determined through hyperparameter tuning.

To select the optimal number of nearest neighbors (k), we generated FE predictors by varying k (see Appendix 4 for details on FE construction) and filled their gaps as described previously. Each gap-free FE predictor was then included iteratively in the MODIS AOD model. The optimal FE variable was identified by minimizing the root mean square error (RMSE) on the test sets. This was achieved by sequentially adding each FE variable into a regression model of the MODIS AOD variable, one at a time, and identifying the point at which no further significant improvement was obtained when increasing k. We refer to this methodology as the "cut-off point rule" for selecting k.

$$x_p : \text{Longitude of cell } p \in P, \qquad y_p : \text{Latitude of cell } p \in P$$

$$D_{pj} : \text{Distance from cell } p \in P \text{ to cell } j \in P.$$

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2} \quad \forall p, j \in P \qquad \text{Eq. 1}$$

$$D = \begin{vmatrix} 0 & D_{1,2} & \dots & D_{1,|P|} \\ D_{2,1} & 0 & \dots & \dots \\ \dots & \dots & 0 & D_{|P|-1,|P|} \\ D_{|P|,1} & \dots & D_{|P|,|P|-1} & 0 \end{vmatrix}$$

To select the final FE candidate for the MODIS AOD model, we generated 20 FEs based on different values of the k parameter (see Equation 2). The distance matrix D was from latitude and longitude coordinates (see Figure 5). To maintain model efficiency, we adopted a parsimonious approach using XGBoost. At each iteration, all available spatial and spatio-temporal predictors, including the FE variables, were incorporated. Averages for FE calculation were computed using only non-missing neighboring AOD values, and the value of the point itself was excluded from its own FE calculation.

$$\mathrm{FE}^{k}{}_{AOD}(p,t) = \frac{1}{|\ kN_p{}^{valid}\ |} \sum_{\mathrm{j} \in kN_p{}^{valid}} AOD_{j,t}\ \forall \mathrm{p} \in \mathrm{P}, \mathrm{t} \in \mathrm{T} \qquad \text{Eq. 2}$$

$$|\ kN_p{}^{valid}\ | \leq \mathrm{k}$$

$kN_p{}^{valid}$: Set of valid k nearest neighbors for a point p where the AOD value is not missing.

$|\ kN_p{}^{valid}\ |$: # of valid k neighbors for a point p, neighbors whose AOD values are not missing.

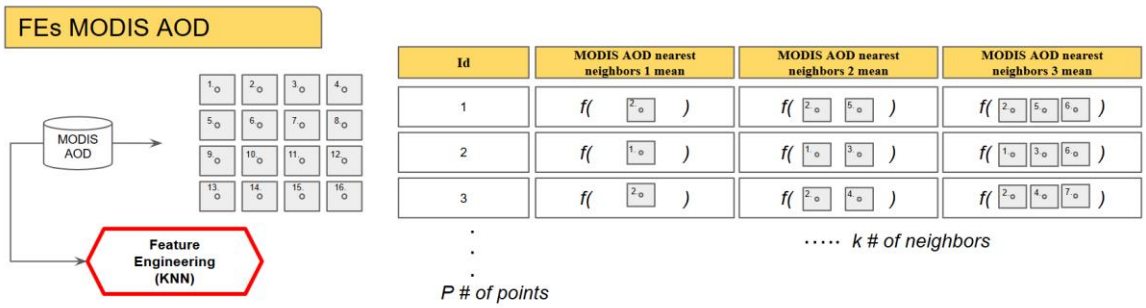


Fig. 5. Simplified example for a specific day, illustrating how distances are quantified using latitude and longitude coordinates when calculating FE predictors for MODIS AOD.

Stage 2 model: Imputation of MODIS AOD

In Stage 2, gaps in MODIS AOD were filled by training XGBoost models with the following predictors: latitude, longitude, population density, elevation, sine and cosine transformation for calendar date variables, NDVI, EVI, CAMS AOD 469, CAMS AOD 550, CAMS AOD 670, CAMS AOD 865, CAMS AOD 1240, Weekly mean MODIS AOD (global), Weekly

mean MODIS AOD (per grid cell), Weekly minimum MODIS AOD (per grid cell), Weekly maximum MODIS AOD (per grid cell), CAMS AOD ratio (550/670 nm), CAMS AOD ratio (550/1240 nm), CAMS AOD difference (550–670 nm), CAMS AOD difference (1240–469 nm), Normalized Difference Index (NDI 550–1240), Minimum CAMS AOD (multi-band), Mean CAMS AOD (multi-band), Maximum CAMS AOD (multi-band), Standard deviation CAMS AOD (multi-band), FE MODIS AOD selected from Stage 1. Equation 3 summarizes our final Stage 2 model, which incorporates the optimally selected FE feature, derived from latitude and longitude coordinates used to compute the distance matrix *D*.

$$\begin{aligned}\mathrm{AOD} \sim f(&\mathrm{lat}, \mathrm{long}, \mathrm{elevation}, \text{population density},\\ &\sin(\mathrm{DoY}), \cos(\mathrm{DoY}) \ldots, \sin(\mathrm{weekday}), \cos(\mathrm{weekday}), \ldots\\ &, \mathrm{NDVI}, \mathrm{EVI}, \mathrm{CAMS\ AOD\ 469}, \mathrm{CAMS\ AOD\ 550}, \ldots, \mathrm{FE\ MODIS\ AOD}) \end{aligned} \quad \text{Eq. 3}$$

Stages 3 and 5 model: Feature Engineering of PM

The feature generation approach was extended and applied in similar manner to both PM_{10} and $PM_{2.5}$, and we therefore describe the procedure for both pollutants together. As in Stage 1, basic feature engineering was applied to create and transform predictors aimed at improving model performance. Some features from Stage 1 were reused, while additional features introduced in this stage included categorical target encoding of natural areas, categorical target encoding of regions, and distance to the nearest PM monitoring station. A detailed description of these new predictors is provided in Table A2 of Appendix 2.

To produce daily estimates of PM, we modified the initial KNN-based feature engineering approach by restricting FE generation exclusively to the locations of PM monitors. A key difference of this stage compared to Stage 1, is that distance matrix *D* is now structured along its columns dimension. Specifically, the matrix was reduced from a full (*P*,*P*) structure, where *P* denotes the total number of points across Israel, to a (*P*,*s*) structure, with *s* representing the number of active PM monitors. To prevent, a monitor location from selecting itself as a nearest neighbor, the corresponding entries in *D* were replaced with infinity before computing FEs. Appendix 3 provides a detailed explanation of the modifications applied to *D* during this process.

To select the optimal FE for inclusion in our models, we again applied the "cut-off point rule". In addition, inverse distance weighting (IDW) was incorporated into the FE procedure to ensure that observations from nearby monitors had greater influence when computing

neighbor averages for a given location (Fig. 6). Equation 4 shows how an FE is calculated at this stage. As illustrated in Figure 7, the IDW method assigns higher weight, such as (W_1, S_2), to measurements from monitors like S_2 that are geographically closer to a particular point of interest (e.g., p=1). This weighting logic dynamically adjusts for each grid point based on its spatial relationship to surrounding monitors.

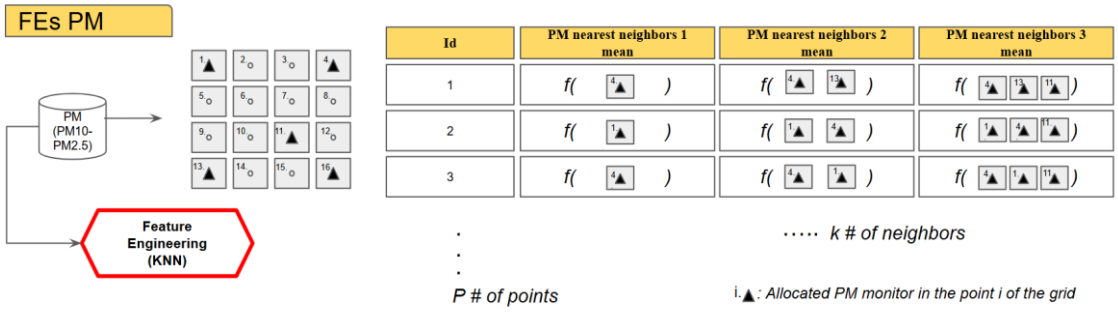


Fig. 6. Illustration of FE calculation for PM. This minimal example shows the process for a specific day, using latitude and longitude coordinates to compute distances in the distance matrix D.

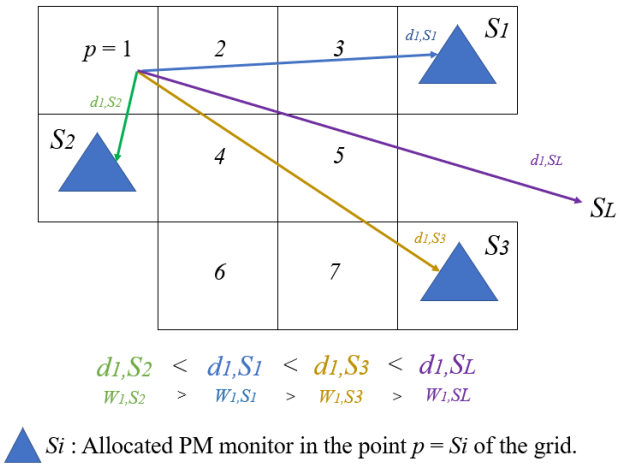


Fig. 7. Illustration of the inverse distance weighting (IDW) logic used in the FE calculation for PM.

We trained an XGBoost model for each i-th value of k used in feature generation. Our final model for this stage's hyperparameter tuning is summarized in Equation 5. All predictors associated with this stage are included, such as distance from main roads, distance from highways, distance from cities, and others.

$$\mathrm{FE}^{k}{}_{PM}(p,t) = \frac{1}{|\, ks_p{}^{valid} \,|} \sum_{j \in ks_p{}^{valid}} W_{p,j} * PM_{j,t}; \forall p \in P, t \in T \quad \text{Eq. 4}$$

$$| ks_p{}^{valid} | \leq k \leq s$$

$ks_p{}^{valid}$: Set of k valid nearest neighbors monitors of a point p which PM value is not missing.

$| ks_p{}^{valid} |$: # of k valid nearest neighbors monitors for a point p where PM is not missing.

$$\begin{aligned} \mathrm{PM}^k \sim f(&\text{latitude, longitude, elevation, population density, distance from main roads, ...} \\ &\text{, sin(DoY) , cos(DoY) ..., sin(weekday), cos(weekday), ...} \\ &\text{, NDVI, EVI, CAMS AOD 469, CAMS AOD 550, ..., gap} - \text{free MODIS AOD, ...} \\ &\text{, categorical target encode of regions, distance to first PM nearest monitor, FE}^k \text{ PM)} \end{aligned} \quad \text{Eq. 5}$$

Stage 4 and 6 model: prediction of PM

In Stages 4 and 6, we predicted daily PM concentrations using the following predictors: latitude, longitude, population density, elevation, distance from main roads, distance from highways, distance from cities, length all roads, length highways, land-use, sine and cosine transformations for calendar date variables, NDVI, EVI, CAMS AOD 469, CAMS AOD 550, CAMS AOD 670, CAMS AOD 865, CAMS AOD 1240, CAMS AOD ratio (550/670 nm), CAMS AOD ratio (550/1240 nm), CAMS AOD difference (550–670 nm), CAMS AOD difference (1240–469 nm), Normalized Difference Index (NDI 550–1240), Minimum CAMS AOD (multi-band), Mean CAMS AOD (multi-band), Maximum CAMS AOD (multi-band), CAMS DUST AOD, wind speed, wind direction, Dust storm classification, categorical target encode of natural areas, categorical target encode of regions, distance to first PM nearest monitor, gap-free MODIS AOD and the selected FE PM feature from the previous stage.

Equations 6A and 6B summarize the mathematical formulations for PM_{10} and $PM_{2.5}$, respectively. XGBoost models (1000 trees, column sampling rate of 0.45) were trained for this purpose.

$$\begin{aligned} PM_{10} \sim f(&\text{latitude, longitude, elevation, population density, distance from main roads, ...} \\ &\text{, sin(DoY) , cos(DoY) ..., sin(weekday), cos(weekday), ...} \\ &\text{, NDVI, EVI, CAMS AOD 469, CAMS AOD 550, ..., gap} - \text{free MODIS AOD, ...} \end{aligned}$$

, categorical target encode of regions, distance to first PM_{10} nearest monitor, selected FE PM_{10}) Eq. 6. A

$PM_{2.5} \sim f$(latitude, longitude, elevation, population density, distance from main roads, …

, sin(DoY) , cos(DoY) …, sin(weekday), cos(weekday), …

, NDVI, EVI, CAMS AOD 469, CAMS AOD 550, … , gap − free MODIS AOD, …

, categorical target encode of regions,
distance to first $PM_{2.5}$ nearest monitor, selected FE $PM_{2.5}$) Eq. 6. B

Validation

Our results were cross-validated using a standard 5-fold cross-validation (CV) approach for Stages 1 and 2 of the model. To address potential spatial bias in the 1 km PM estimates, particularly in areas with low monitoring density, we implemented a leave-one-out spatial blocking cross-validation approach. Specifically, we selected a group of monitors distributed across the territory, excluded one monitor at a time from the training data as a held-out fold, and trained the models on the remaining data. RMSE results were then averaged for these monitors, discriminating by the regions defined in Zhou et al. (2020), which allowed us to identify areas where the models performed weakly (Figure 8).

To correctly implement the feature engineering approach, KNN-derived features were generated exhaustively using only the training data at each iteration of CV, ensuring that the excluded monitor's PM values were not used when computing the FEs. This evaluation provided a robust assessment of model performance in regions with limited monitoring coverage, ensuring that no data leakage occurred during feature engineering.

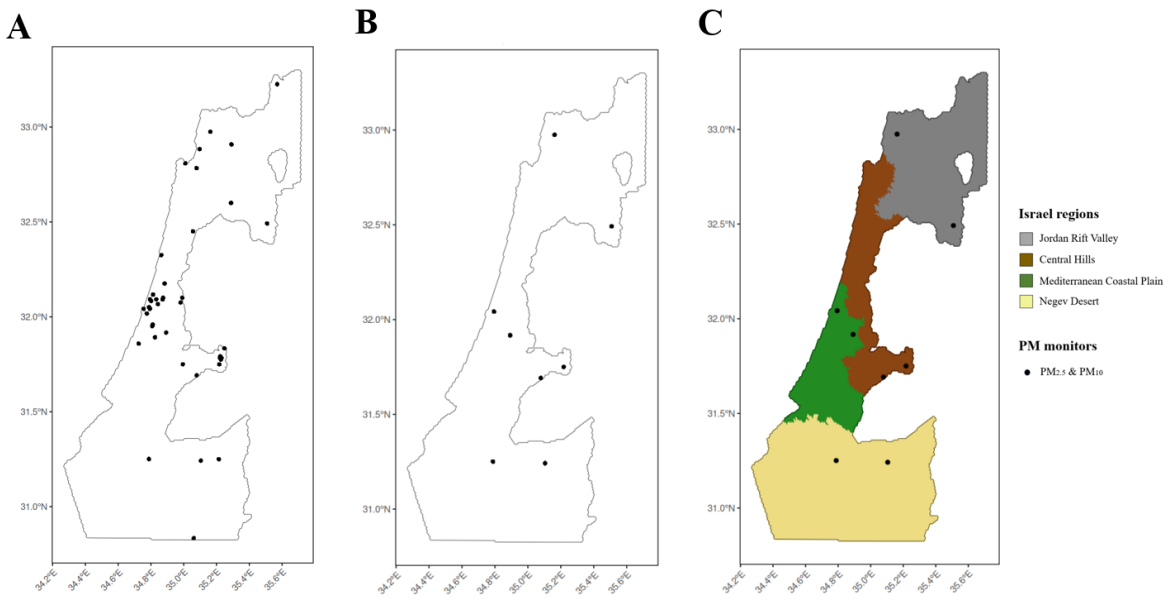


Fig. 8. Procedure for selecting held-out monitors in the leave-one-monitor-out spatial blocking cross-validation of the 1 km PM models. Eight monitors were selected**.**

A. The total number of PM monitors in Israel (2024).

B. A selected sample of 8 PM monitors to be held out, identified using clustering.

C. The 8 selected PM monitors within their corresponding regions.

3. Results

3.1 Correlation heatmap

Figure 9 presents the correlations between all AOD types and PM, prior to any imputation or prediction stage.

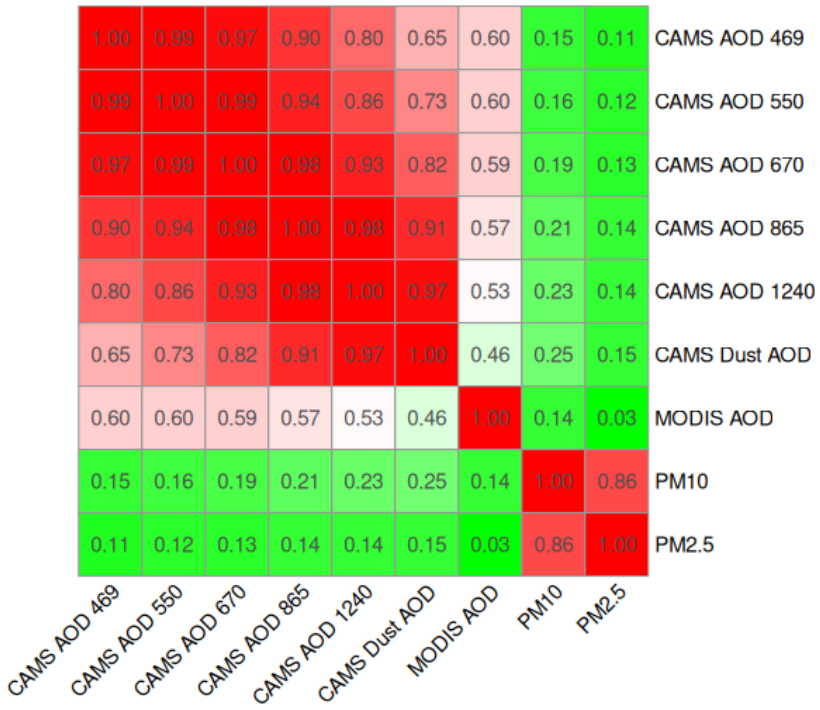


Fig 9. Correlation heatmap of all AOD types, and PM for 2024 year.

3.2 Tuning *k* parameter in stage 1 model

We began the hyperparameter tuning process for the FEs of MODIS AOD, varying the k parameter from 1 to 20 (Figure 10), and applied a random K-fold CV using 80% of the data for training and 20% for testing. Initially, the RMSE on the test sets was approximately 0.042 for a model trained without any FE features. Adding the first calculated FE (k = 1) reduced the RMSE from 0.042 to nearly 0.019. After further iterations, the optimal cut-off point was identified at k = 3, which further decreased the RMSE to 0.017. These results were obtained for the year 2024. A similar approach was applied for the other years in the study, yielding comparable results with slight variations in k, indicating that this optimization process only needs to be performed once. Tuning results for FEs of MODIS AOD in 2015 are provided in Annex A (Figure A.1).

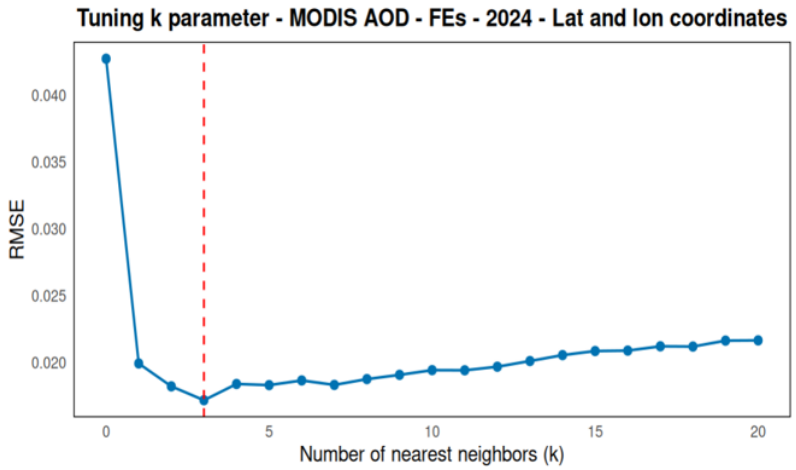


Fig 10. FE selection for the model that imputes MODIS AOD, year 2024.

3.3 Performance of the stage 2 model

Table 2 presents the 5-fold CV performance of the Stage 2 models across all years, incorporating the selected FE feature from the previous stage. The average MODIS AOD RMSE on the test sets across the years was 0.0177. As an example, we show the results for the gap-filling of MODIS AOD on March 23 for 2003, 2013, and 2024, when the percentage of missing values was high, with more than 60% gaps in each case (see Figure 11).

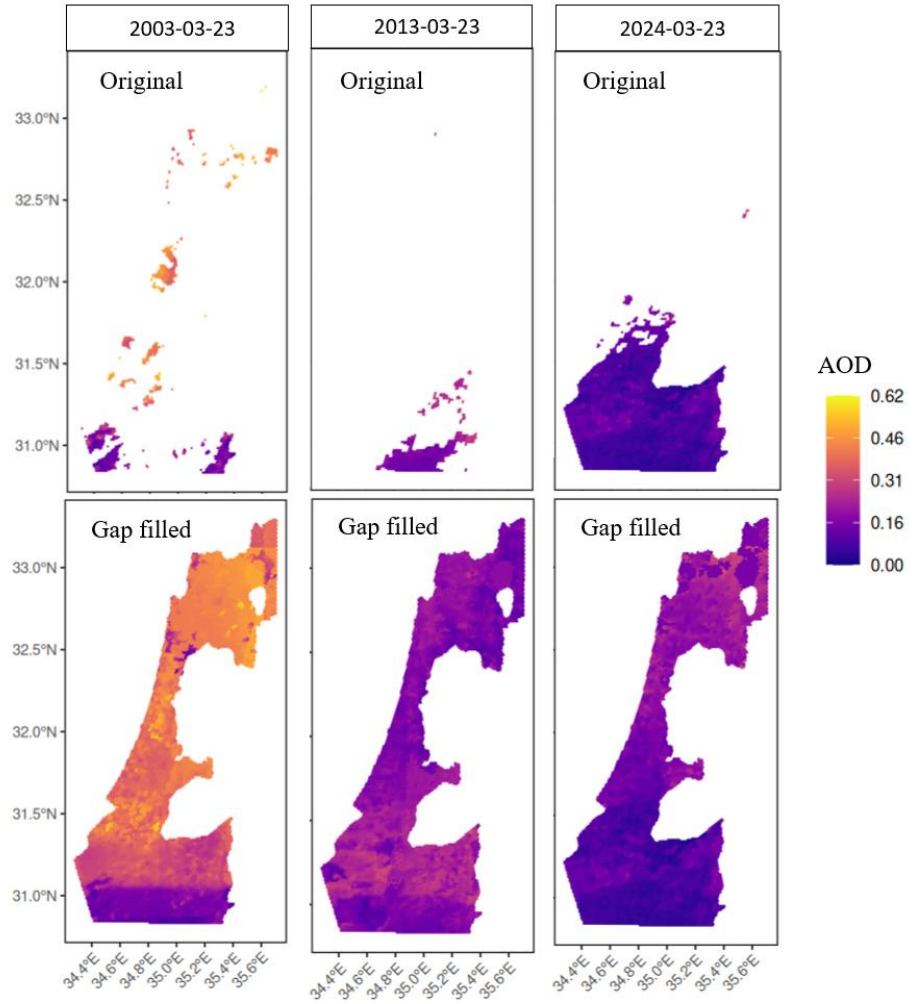


Fig. 11. Gap-filling of MODIS AOD on March 23 in 2003, 2013, and 2024.

Table 2. Stage 2 model performance of MODIS AOD imputation, 5-fold CV has been used across all years (2003–2024). RMSE metrics were obtained from XGBoost models with 1000 trees and a column sampling rate of 0.45. Training and test RMSE metrics are reported.

Year	RMSE train	RMSE test
2003	0.019	0.020
2004	0.016	0.016
2005	0.012	0.013
2006	0.023	0.024
2007	0.016	0.016
2008	0.017	0.016
2009	0.013	0.014
2010	0.024	0.025

2011	0.019	0.019
2012	0.020	0.020
2013	0.016	0.017
2014	0.015	0.015
2015	0.023	0.023
2016	0.020	0.021
2017	0.013	0.012
2018	0.015	0.015
2019	0.019	0.019
2020	0.014	0.014
2021	0.016	0.016
2022	0.022	0.022
2023	0.016	0.016
2024	0.017	0.017

The feature importance analysis for this stage indicates that the engineered predictors developed through feature engineering contribute substantially to the modeling of MODIS AOD, with six of the top ten variables being FE-based. The most important predictors were the KNN-based MODIS AOD FE, weekly mean MODIS AOD (per grid cell), weekly minimum MODIS AOD (per grid cell), CAMS AOD 550, mean CAMS AOD (multi-band), and weekly maximum MODIS AOD (per grid cell), collectively accounting for nearly 94% of the model's importance. These results are illustrated in Figure 12.

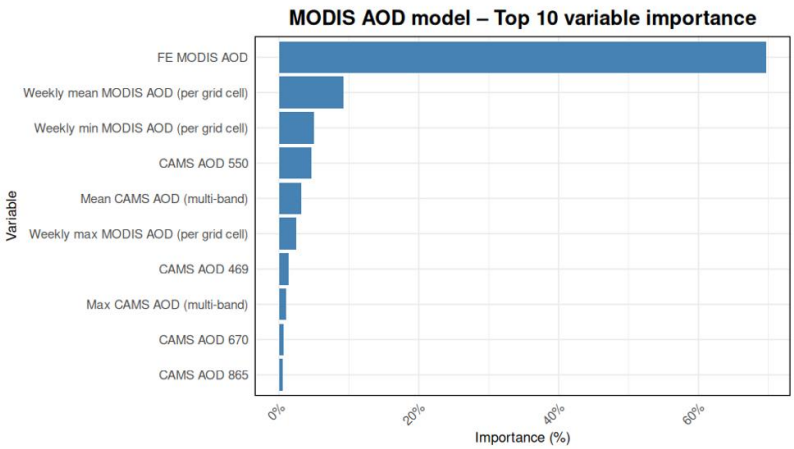


Fig. 12. Variable importance for final MODIS AOD model.

3.2 Tuning *k* parameter in stages 3 and 5 model

To determine the optimal value of *k* and select the best FE features for the PM models, we generated 15 features using Inverse Distance Weighting (IDW). Feature selection was performed based on our predefined cut-off point rule.

The tuning process identified $k = 6$ for the FEs corresponding to PM_{10} and $k = 4$ for the FEs corresponding to $PM_{2.5}$. These results were consistent across the study years, so we decided to maintain the same *k* values for all periods (2003–2023). This indicates that annual optimization of this parameter is unnecessary and only needs to be performed once.

Figures 13 and 14 illustrate the tuning results for PM_{10} and $PM_{2.5}$ in 2024, where latitude and longitude were used as the geographical attributes for computing *D*. Tuning results for PM models in 2015 are provided in Annex A (Figures A.2 and A.3).

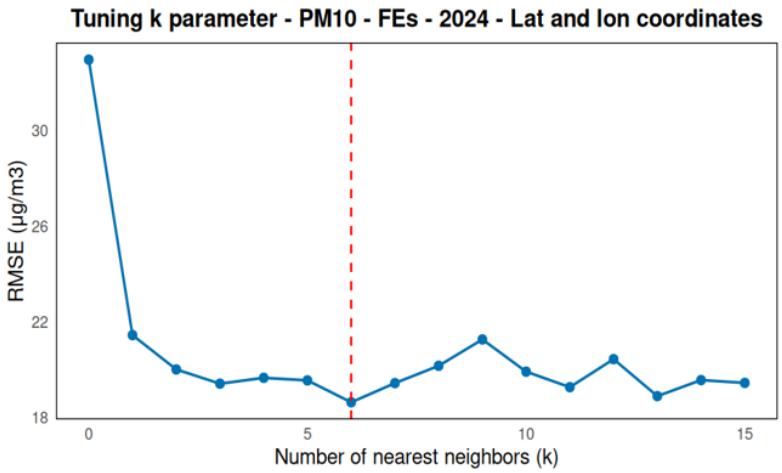


Fig 13. Selection of the best FE (using latitude and longitude dimensions in *D*) for the model that predicts PM_{10}, 2024.

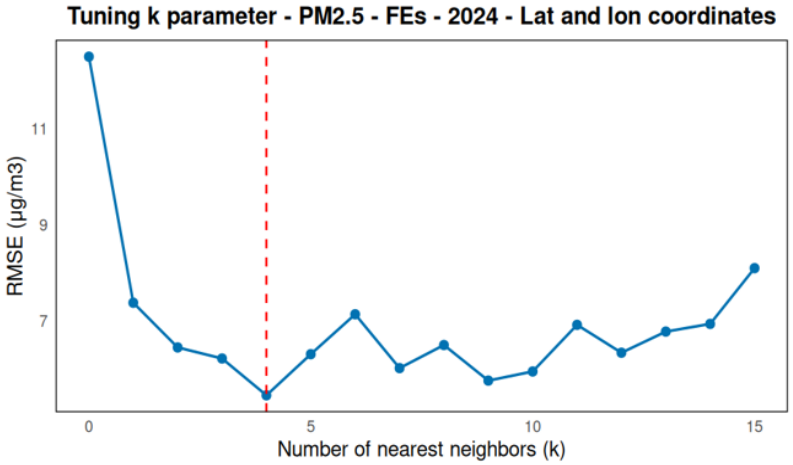


Fig 14. Selection of the best FE (using latitude and longitude dimensions in *D*) for the model that predicts $PM_{2.5,}$ year 2024.

3.3 Performance of the stage 4 and 6 model

Performance in these stages was evaluated using a leave-one-monitor-out CV with eight monitors, calculating summary metrics per region. Overall results by region are provided in Tables 3 and 5 for PM_{10} and $PM_{2.5}$, respectively. Tables 4 and 6 show results across years for both PM models. In general, the models performed well, with average RMSE values of 17.88 µg/m³ for PM_{10} and 6.60 µg/m³ for $PM_{2.5}$ (2003-2024). When spatially blocking eight monitors (two per region), the results indicate better model performance in the Jordan Rift Valley and poorer performance in the Central Hills for PM_{10}. For $PM_{2.5}$, all regions generally exhibited good generalization, although the Mediterranean Coastal Plain showed the poorest evaluation metric scores. These differences can be attributed to the higher density of $PM_{2.5}$ monitors compared to PM_{10}, which enhances model precision (see Appendix 3, Table A.3). Comparison between training and test results indicated no overfitting, as the metrics were similar across both sets (Ng, 2017).

Results PM_{10}

Table 3. Cross-validated average results for PM_{10} from the eight spatially blocked monitoring stations, reported overall for all years.

Region	RMSE Train (µg/m⁻³)	RMSE Test (µg/m⁻³)
Central Hills	22.07	27.71
Jordan Rift Valley	6.72	8.44

Mediterranean Coastal Plain	15.85	19.90
Negev Desert	12.31	15.46
Overall	**14.24**	**17.88**

Table 4. Cross-validated results for PM_{10} from the 8 spatially blocked monitoring stations across all years.

Year	RMSE train ($\mu g/m^{-3}$)	RMSE test ($\mu g/m^{-3}$)
2003	17.784	18.760
2004	20.843	24.300
2005	8.822	12.820
2006	10.042	11.370
2007	21.362	22.180
2008	9.819	15.630
2009	15.079	22.710
2010	23.190	26.870
2011	13.700	17.390
2012	17.597	21.630
2013	12.924	14.130
2014	11.608	12.960
2015	25.674	26.970
2016	13.041	15.890
2017	9.726	13.870
2018	14.271	16.290
2019	9.058	13.060
2020	8.345	10.290
2021	11.222	11.410
2022	14.006	24.010
2023	11.824	22.110
2024	13.276	18.63

Feature importance in PM_{10} models

Feature importance at this stage supports FE and its implementation. Figure 13 shows the top ten most important predictors for the PM_{10} model, with the nearest-neighbor FE being the most influential. Other important predictors include variables related to cyclical temporal patterns, several CAMS longwave bands, NDVI, and gap-free MODIS AOD imputed in Stage 2.

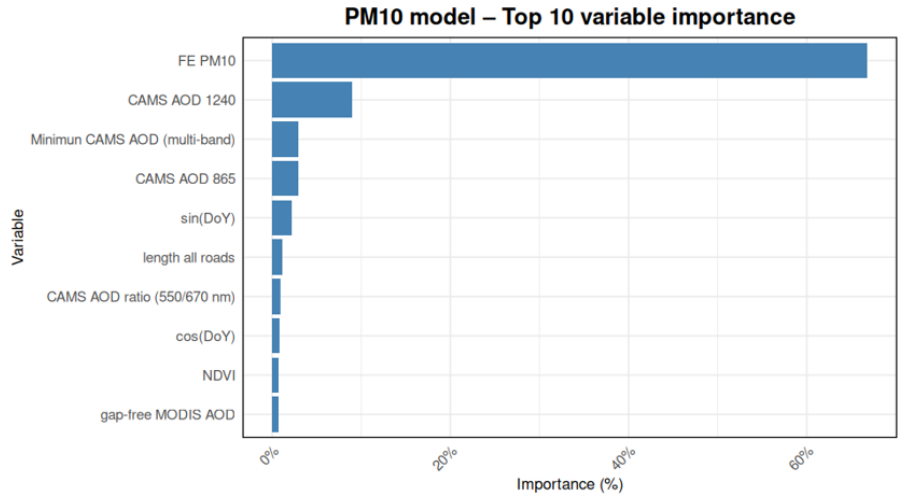


Fig 15. Feature importance of PM_{10} models.

Results PM2.5

Table 5. Cross-validated average results for $PM_{2.5}$ from the eight spatially blocked monitoring stations, reported overall for all years.

Region	RMSE Train (µg/m⁻³)	RMSE Test (µg/m⁻³)
Central Hills	2.33	3.62
Jordan Rift Valley	3.73	5.79
Mediterranean Coastal Plain	7.97	12.36
Negev Desert	2.98	4.62
Overall	**4.25**	**6.60**

Table 6. Cross-validated results for $PM_{2.5}$ from the 8 spatially blocked monitoring stations across all years.

Year	RMSE train (µg/m⁻³)	RMSE test (µg/m⁻³)
2003	6.65	9.18
2004	11.35	13.96

2005	3.16	5.03
2006	7.66	13.09
2007	4.33	6.55
2008	3.84	6.37
2009	6.85	9.39
2010	8.02	10.02
2011	5.20	7.29
2012	4.35	6.82
2013	2.20	4.28
2014	2.41	5.35
2015	8.37	10.84
2016	3.88	5.62
2017	2.70	4.46
2018	1.85	3.91
2019	0.91	2.98
2020	0.66	2.73
2021	2.03	4.06
2022	1.53	3.52
2023	1.88	3.89
2024	3.73	5.84

Feature importance in $PM_{2.5}$ models

Figure 16 shows the ten most important predictors for the $PM_{2.5}$ model, with the nearest-neighbor FE being the most influential. Other key predictors include the categorical target encoding of natural areas, distance from cities, distance to the first nearest-neighbor $PM_{2.5}$ monitor, and others.

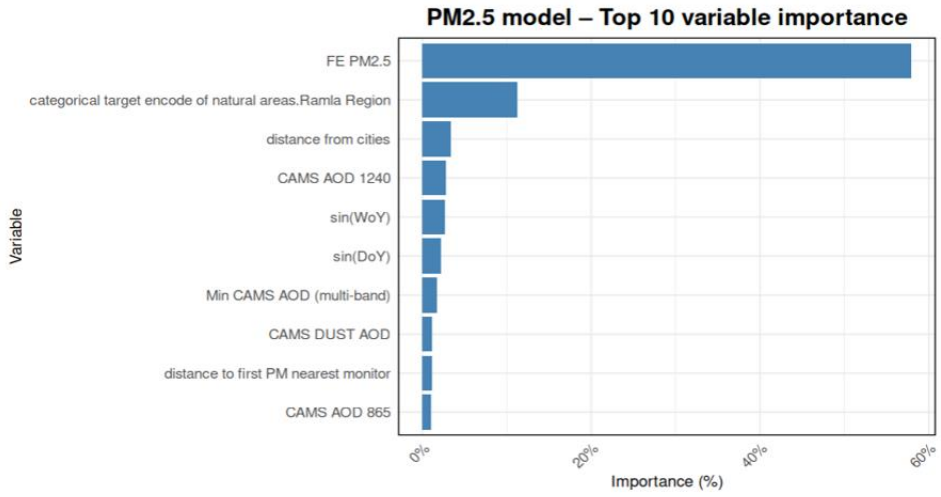


Fig 16. Feature importance of $PM_{2.5}$ models.

Dust Storm Classification for Approximating Anthropogenic Emissions

Using the methodology of Achilleos et al. (2020), we classified locations experiencing dust storm events. To distinguish PM values from anthropogenic sources, we excluded PM measurements recorded on dust storm days (DDS) from our model predictions and then calculated annual averages to identify areas in Israel with higher exposure to human emissions. Figure 17 illustrates zones with particularly high exposure (2024), including major roads in the country, Tel Aviv, Ben Gurion Airport, Ashdod Port, Jerusalem, the Ramat Hovav industrial zone, and Beersheba.

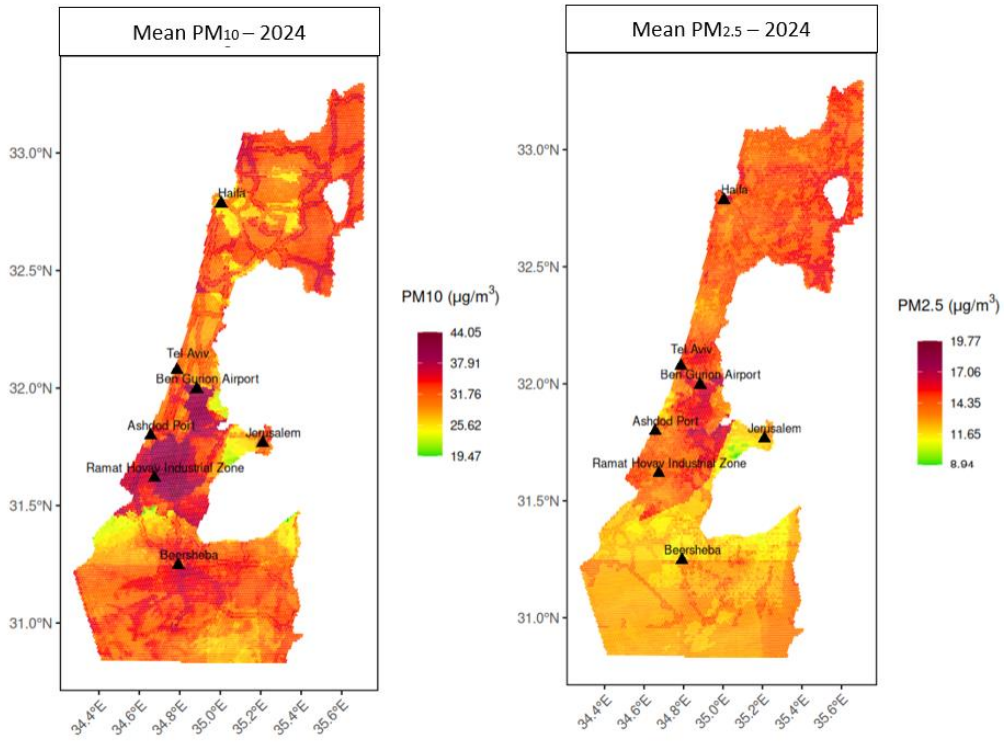


Fig. 17. Predicted annual mean 1 km PM from anthropogenic sources, 2024.

4. Discussion

Our modeling demonstrated robust performance for particulate matter models at 1 km resolution. From 2003 to 2024, the models achieved RMSE values ranging from 10.29 to 26.97 µg/m³ for PM_{10} and from 2.73 to 13.96 µg/m³ for $PM_{2.5}$. A key innovation of this study was the integration of both basic and advanced feature engineering techniques for PM modeling across Israel, approaches that have not been previously tested in this context. Results were compared with previous studies that primarily focused on 1 km scales (Stein et al., 2018; Kloog et al., 2015). Additionally, we developed a specialized method for imputing AOD, reporting metrics for each year, which supports the integration of feature engineering in both prediction and gap-filling stages.

Our methodology performs slightly worse than the most recent study conducted in Israel by Shtein et al. (2018), despite covering a longer time span (2003–2024 vs. 2005–2015). However, when comparing the overlapping years (2005–2015), our results were pretty similar (under our validation approach), with overall RMSE values of 18.60 (this study) versus 17.91 µg/m³ for PM_{10}, and 7.73 (this study) versus 6.10 µg/m³ for $PM_{2.5}$. This suggests that our validation scheme is more realistic, as it employed spatially blocked cross-validation with

folds of equal size while accounting for regional differences. Averaging results by monitor density per area would likely have produced lower RMSE values, potentially below those reported in the previous study. Recent PM modeling studies in other regions have also employed spatial cross-validation approaches (Hough et al., 2021; Shtein et al., 2020; Just et al., 2025), providing additional support for our validation scheme.

To the best of our knowledge, this is the first study applying feature engineering to enhance the performance of exposure models for PM estimation in Israel. Overall, few studies in epidemiology and environmental exposure have employed this approach, leaving considerable room for future advancements that integrate FE with emerging machine learning methods, such as artificial neural networks (Wang et al., 2025). For example, Just et al. (2025) used FE to create a new PM predictor based on interpolation of monitors within 500 km of the site of interest, enabling predictions at arbitrary locations in the U.S. Similarly, Wang et al. (2025) applied a KNN-based approach to generate PM features from the nearest monitors for each site, subsequently integrating these features with deep learning models in the U.S. Although feature engineering using distance-based approaches like KNN has been applied in other countries, this study is among the first frameworks to emphasize the optimization of key parameters, such as the number of nearest neighbors k, and to demonstrate its efficiency.

This methodology provides an innovative and adaptable approach for constructing new independent predictors (FEs) using the KNN model. While we did not extend FE generation to other attributes, our formulation can be easily modified through the distance function, allowing experts to apply their own criteria when constructing the distance matrix D. For the PM case, we focused solely on latitude and longitude coordinates; however, additional dimensions could be explored to further enhance the model's explanatory power. We encourage researchers investigating other pollutants (e.g., air temperature, CO_2, NO_2) to adopt our FE approach and develop new predictors based on factors relevant to their studies, such as traffic congestion, population density, NDVI, or proximity to watercourses. Incorporating these predictors into exposure modeling could lead to substantial improvements.

5. Conclusion

This study presents a novel six-stage framework for estimating PM (PM_{10} and $PM_{2.5}$) using satellite data from MODIS. By leveraging current computational capabilities, we enhanced traditional exposure modeling studies through the integration of feature engineering techniques, specifically utilizing distance matrix functions and the KNN model. Our approach

enables high-resolution PM estimates across both spatial and temporal dimensions. Furthermore, we provide a dataset that captures PM variability in Israel, from coarse particles to fine particles smaller than 2.5 μm, which can penetrate the human body. The methodology includes detailed steps for deriving predictors from the original collected data, ranging from basic to advanced features, which can be readily adapted to other research areas and objectives. Our models demonstrated strong performance based on RMSE metrics over yearly timeframes. This innovative approach leverages globally accessible geospatial datasets, allowing broad applicability. While our study focused on Israel, the methodology can be easily replicated in other contexts. We believe our findings will contribute to reducing bias in health-related studies and strengthen the reliability of scientific hypotheses.

Appendix 1

Table A1. (Appendix 1) Basic feature-engineered predictors calculated in stage 1 modeling.

Feature Name	Input Variables	Objective	Model Stages Used	Missing Value Treatment
Cyclical calendar transformations (sine, cosine)	Day of year, weekday, week of year, month	Represent cyclical temporal patterns	1–6	Not applicable (no missing values)
Weekly mean MODIS AOD (global)	MODIS AOD	Summarize weekly aerosol variability	1–2	Simple imputation
Weekly mean MODIS AOD (per grid cell)	MODIS AOD	Summarize weekly aerosol variability at each grid cell	1–2	Simple imputation
Weekly minimum MODIS AOD (per grid cell)	MODIS AOD	Identify low weekly aerosol values	1–2	Simple imputation
Weekly maximum MODIS AOD (per grid cell)	MODIS AOD	Identify high weekly aerosol values	1–2	Simple imputation

CAMS AOD ratio (550/670 nm)	CAMS AOD at 550 nm, 670 nm	Characterize differences between spectral bands	1–6	Not applicable (no missing values)
CAMS AOD ratio (550/1240 nm)	CAMS AOD at 550 nm, 1240 nm	Characterize differences between spectral bands	1–6	Not applicable (no missing values)
CAMS AOD difference (550–670 nm)	CAMS AOD at 550 nm, 670 nm	Quantify differences between bands	1–6	Not applicable (no missing values)
CAMS AOD difference (1240–469 nm)	CAMS AOD at 1240 nm, 469 nm	Quantify differences between bands	1–6	Not applicable (no missing values)
Normalized Difference Index (NDI 550–1240)	CAMS AOD at 550 nm, 1240 nm	Summarize spectral relationships	1–6	Not applicable (no missing values)
Minimum CAMS AOD (multi-band)	CAMS AOD at 469, 550, 670, 865, 1240 nm	Identify low values across bands	1–6	Not applicable (no missing values)
Mean CAMS AOD (multi-band)	CAMS AOD at 469, 550, 670, 865, 1240 nm	Summarize average values across bands	1–6	Not applicable (no missing values)
Maximum CAMS AOD (multi-band)	CAMS AOD at 469, 550, 670, 865, 1240 nm	Identify high values across bands	1–6	Not applicable (no missing values)
Standard deviation CAMS AOD (multi-band)	CAMS AOD at 469, 550, 670, 865, 1240 nm	Measure variability across bands	1–6	Not applicable (no missing values)

Appendix 2

Table A2. (Appendix 2) Basic feature-engineered predictors calculated in stage 3 and 5 modeling.

Feature Name	Input Variables	Objective	Model Stages Used	Missing Value Treatment

Categorical target encode of natural areas	Natural areas	Encode natural areas dinamically with their means per region and date	3-6	Not applicable (no missing values)
Categorical target encode of regions	Israel regions	Encode regions dynamically with their means per region and date	3-6	Not applicable (no missing values)
Distance to first PM nearest neighbor monitor	None (Distance matrix D is used)	Penalize the reliance on the FE for PM where first neighbor is far	3-6	Not applicable (no missing values)

Appendix 3

Table A3. (Appendix 3) Monitors measuring PM10 and PM2.5 in Israel yearly.

Year	Monitors measuring PM_{10}	Monitors measuring $PM_{2.5}$
2003	14	29
2004	15	28
2005	16	28
2006	18	33
2007	20	34
2008	23	33
2009	22	37
2010	22	30
2011	28	31
2012	28	30
2013	34	35
2014	37	40
2015	40	35
2016	44	29

2017	48	34
2018	45	34
2019	53	31
2020	47	28
2021	24	12
2022	25	13
2023	27	14
2024	34	17

Appendix 4

Distance is calculated with respect to stationary variables (l_1, l_2, l_3, etc.); for example, $l_i = x_p$ (longitude), $l_i = y_p$ (latitude), or $l_i = E_p$ (elevation). Once distances between points are computed, sets of nearest neighbors are determined by selecting the column arguments with the smallest scores.

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2} \text{ or } D_{pj} = \sqrt{(E_p - E_j)^2}, \text{or even}$$

$$D_{pj} = \sqrt{(x_p - x_j)^2 + (y_p - y_j)^2 + (E_p - E_j)^2} \quad \forall p, j \in P \qquad \text{Eq. 1}$$

kN_p : Set of first k nearest neighbors points of the point $p \in P$.

$$kN_p = [N_1 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}), N_2 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}) \mid D_{p,N1} \notin \{D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}\}, N_3 = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}) \mid D_{p,N1}, D_{p,N2} \notin \{D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}\} \ldots, N_k = \arg.\min(D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}) \mid D_{p,N1}, D_{p,N2}, D_{p,N3} \ldots, D_{p,NK-1} \notin \{D_{p,1}, D_{p,2}, D_{p,3 \ldots} D_{p,|P|}\}]$$

Eq. A2.1

To calculate one of the MODIS AOD FEs, we used the following expression:

$$\text{MODIS AOD nearest neighbors } k \text{ mean}_{p,t} = \sum_{j \in kN_p}^{k} AOD_{j,t}/k \; \forall p \in P, t \in T \qquad \text{Eq. A2.2}$$

Appendix 5

$P_{monitors}$: Set of points with monitors allocated in Israel.

S_i : Denotes a monitor $i \in P_{monitors}$; $P_{monitors} = \{S_1, S_2, S_3, .. S_s\}$; $P_{monitors} \in P$

$$D = \begin{vmatrix} D_{1,S_1} & D_{1,S_2} & \dots & D_{1,S_s} \\ D_{2,S_1} & D_{2,S_2} & \dots & D_{2,S_s} \\ \dots & \dots & \dots & D_{|P|-1,S_s} \\ D_{|P|,S_1} & D_{|P|,S_2} & D_{|P|,S_s-1} & D_{|P|,S_s} \end{vmatrix}$$

$$kS_p : \text{Set of first } k \text{ monitors to the point p} \in \text{P}.$$

$$k < s$$

If a point p is itself a monitor, then we change the value $D_{p=S_i,S_i} = \infty$, thus the point itself is not selected as one of its nearest neighbors.

Annex A

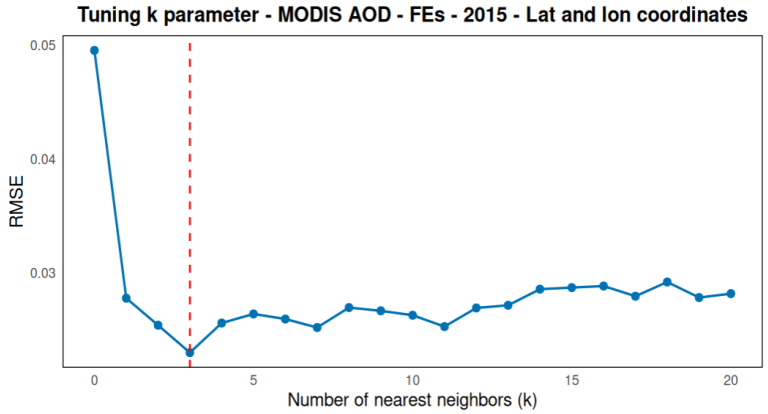


Fig A.1 (Annex A) Selection of the best AOD FE feature (latitude and longitude coordinates in distance matrix *D*) based on 2015 data. Identical results (*k*=3) were obtained for the 2024 dataset.

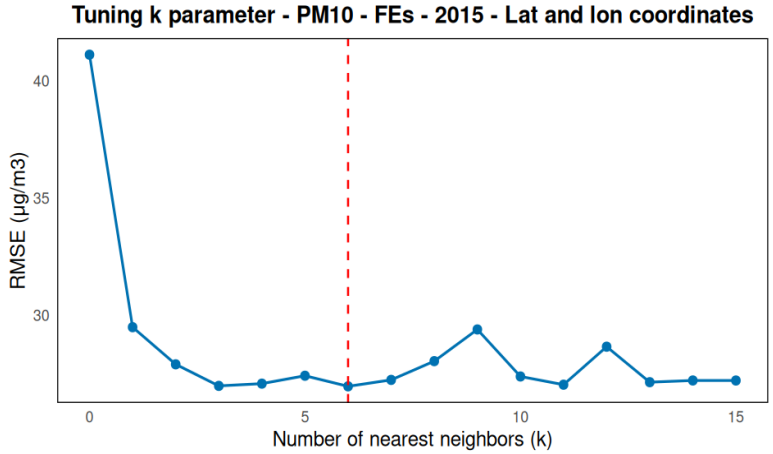


Fig A.2 (Annex A) Selection of the best PM_{10} FE feature (latitude and longitude coordinates in distance matrix *D*) based on 2015 data. Identical results (*k*=6) were obtained for the 2024 dataset.

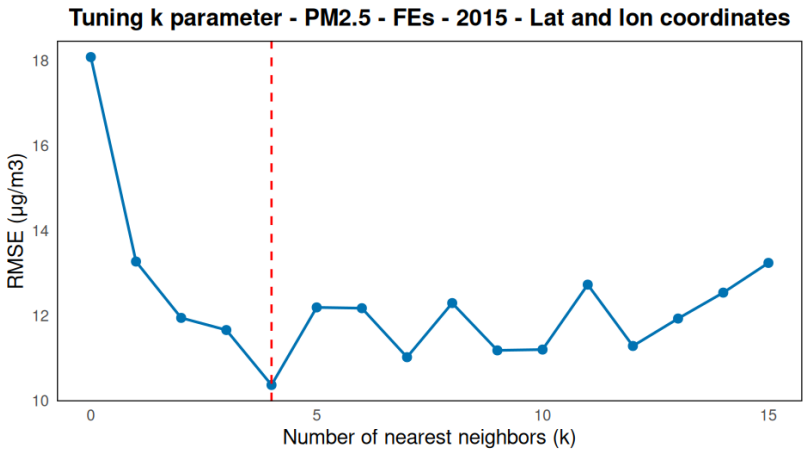


Fig A.3 (Annex A) Selection of the best $PM_{2.5}$ FE feature (latitude and longitude coordinates in distance matrix *D*) based on 2015 data. Identical results (*k*=4) were obtained for the 2024 dataset.

ACKNOWLEDGEMENTS

This work was funded by the U.S.-Israel Binational Science Foundation (grant number 2021197) and the French National Research Agency (ANR-15-IDEX-02). Juan David Briceno is supported by a French public grant under the "Investissements d'avenir" program and a scholarship from the Kreitman School for Advanced Graduate Studies at Ben-Gurion University of the Negev. We would like to thank the EUMETSAT Satellite Application Facility on Land Surface Analysis (LSA SAF) for providing SEVIRI LST data, NASA LP DAAC for MODIS LST data, and IMS for remotely sensed weather data.

AUTHOR CONTRIBUTIONS

Juan David Briceno: Data curation; formal analysis; investigation; methodology; visualization; writing – original draft; writing – review and editing.

Itai Kloog: Formal analysis; Methodology; writing – review and editing.

Johanna Lepeule: Formal analysis; methodology; writing – review and editing.

Guillaume Barbalat: writing – Review and editing.

Elena Colicino: Writing – Review and editing.

Massimo Stafoggia: Methodology; conceptualization; writing – review and editing.

Jakob Alderson: Data curation; visualisation.

Data Availability Statement

The data supporting the findings of this study are not publicly available due to confidentiality restrictions. Access to the data can be requested from the IAB (Institut pour l'Avancée des Biosciences), UGA, where the data is managed.

Funding information

U.S.-Israel Binational Science Foundation (BSF), Grant/Award Numbers: 2021197; and French National Research Agency in the framework of the "Investissements d'avenir" program, Grant/Award Numbers: ANR-15-IDEX-02 IRGA 2021 PAPex project.

ORCID

Juan David Briceno https://orcid.org/0009-0006-3771-528X

References

Achilleos, S., Eleftheriadis, K., Solomos, S., Kalogiros, J., Moussiopoulos, N., Kallos, G., & Kassomenos, P. (2020). Spatio-temporal variability of desert dust storms in the Eastern Mediterranean (Crete, Cyprus, Israel) between 2006 and 2017 using a uniform methodology. Science of the Total Environment, 713. Retrieved from https://doi.org/10.1016/j.scitotenv.2020.136585Central Bureau of Statistics. (2018). *Statistical Abstract of Israel 2018*.

Cohen, N. (2011). Ramat Hovav Industrial Park in the Negev Desert, Israel. *PESD Journal*, *1*. Retrieved from https://pesd.ro/articole/nr.1/Nathan%20Cohen.pdf

De Blesser, P., Hooghe, B., Vlieghe, D., & Van Roy, F. (2007). A distance difference matrix approach to identifying transcription factors that regulate differential gene expression. *Genome Biology*, *8*(5), R83. https://doi.org/10.1186/gb-2007-8-5-r83

Dudhia, J. (2008). Fundamental issues in numerical weather prediction. In *Large-Scale Disasters Prediction, Control, and Mitigation* (pp. 447–452). Cambridge University Press. https://doi.org/10.1017/CBO9780511535963.021

Hough, I., Just, A. C., Zhou, B., Dorman, M., Lepeule, J., & Kloog, I. (2020). A multi-resolution air temperature model for France from MODIS and Landsat thermal data. *Environmental Research*, *183*, 1–11. https://doi.org/10.1016/j.envres.2020.109244

Israel Airports Authority. (n.d.). Ben Gurion International Airport – Carbon Management Plan. Retrieved from https://www.iaa.gov.il/en/environment-and-sustainability/treatingwaste/ben-gurion-international-airport-carbon-management-plan/

Israel Meteorological Service (IMS). (n.d.). Israel Meteorological Service website. Israeli Ministry of Transport and Road Safety. https://ims.gov.il/en (accessed 20 February 2022).

Kishcha, P., et al. (2016). Modelling of a strong dust event in the complex terrain of the Dead Sea valley. *Tellus B: Chemical and Physical Meteorology, 68*(1), 29751. https://doi.org/10.3402/tellusb.v68.29751

Mukid, M. A., Widiharih, T., Rusgiyono, A., & Prahutama, A. (2018). Credit scoring analysis using weighted k-nearest neighbor. *Journal of Physics: Conference Series, 1025*, 012114. https://doi.org/10.1088/1742-6596/1025/1/012114

NASA. MOD13Q1 - MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid.

OECD. (2023). *OECD Environmental Performance Reviews: Israel 2023*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/05/oecd-environmental-performance-reviews-israel-2023_7d20073c/0175ae95-en.pdf

Ochsenwald, W. L., Sicherman, H., et al. (2019). *Israel*. Encyclopaedia Britannica. Encyclopaedia Britannica, Inc.

Shtein, A., Karnieli, A., Katra, I., Raz, R., Levy, I., Lyapustin, A., Dorman, M., Broday, D. M., & Kloog, I. (2020). Estimating daily and intra-daily PM10 and PM2.5 in Israel using a spatio-temporal hybrid modeling approach.

Stein, A., Sorek-Hamer, M., Lyapustin, A., Coull, B., Wang, Y., Just, A., Schwartz, J., & Broday, D. M., Kloog, I. (2015). Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data.

Swathi, N., Bhat, M., Reddy, N., & Rao, A. (2022). Study of distance metrics on k-nearest neighbor algorithm for star categorization. *Journal of Physics: Conference Series, 2161*, 012004. https://doi.org/10.1088/1742-6596/2161/1/012004

Tøttrup, C. (2014). *EU-DEM Statistical Validation Report*.

Watson, G., Telesca, D., Reid, C., Pfister, G., & Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution, 254*, 112792. https://doi.org/10.1016/j.envpol.2019.06.088

Wang, Z., Crooks, J. L., Regan, E. A., & Karimzadeh, M. (2024). High-resolution estimation of daily PM2.5 levels in the contiguous US using Bi-LSTM with attention. *[Journal info if available]*

WHO. (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization. https://www.who.int/publications/i/item/9789240034228

Ng, A. (2017). *Machine Learning Yearning*. Available at: https://www.mlyearning.org

Zhang, T., Zhou, Y., Zhao, K., Zhu, Z., Chen, G., Hu, J., & Wang, L. (2022). A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003–2020). *Earth System Science Data, 14*, 5637–5649. https://doi.org/10.5194/essd-14-5637-2022

Zhou, K., Liu, H., Deng, X., Wang, H., Zhang, S., & Bonafoni, S. (2020). Comparison of machine-learning algorithms for near-surface air-temperature estimation from FY-4A AGRI data. *Advances in Meteorology*. https://doi.org/10.1155/2020/8887364

Zhou, M., et al. (2021). Changes in the concentration of air pollutants before and after the COVID-19 blockade period and their correlation with vegetation coverage. *Environmental Science and Pollution Research, 28*(10), 12345–12356. https://doi.org/10.1007/s11356-021-11964-5

Zhou, X., et al. (2024). Longitudinal association between urban blue-green space exposure and respiratory mortality: A nationwide study. *The Lancet Regional Health – Europe, 13*, 100276. https://doi.org/10.1016/j.lanepe.2024.100276

Just, A. C., Arfer, K. B., Rush, J., Lyapustin, A., & Kloog, I. (2025). XIS-PM2.5: A daily spatiotemporal machine-learning model for PM2.5 in the contiguous United States. *Environmental Research, 251*, 120948. https://doi.org/10.1016/j.envres.2025.120948

Conclusions and Future Work

Throughout this dissertation, we have provided a comprehensive overview of the evolution of exposure modeling approaches, tracing their development from early physically based models to modern, data-driven frameworks. Over time, the availability of large environmental datasets and advances in computation have enabled a paradigm shift in modeling strategies. Land Use Regression (LUR) with satellite data marked the first step toward hybrid models, while current developments increasingly rely on sophisticated machine learning techniques, including neural networks and feature engineering, which represent the future of the field.

Numerous studies have investigated the modeling of environmental variables such as air temperature (Ta) and particulate matter (PM) because of their relevance for health and environmental implications. This thesis focused on France and Israel, two regions with contrasting environmental conditions and dense monitoring networks, highlighting the adaptability of hybrid modeling approaches across diverse settings. A key milestone in this field has been the integration of satellite-derived covariates, which has transformed exposure modeling by enabling large-scale, high-resolution estimations. As satellite technology continues to improve, it is expected that the accuracy and spatial coverage of exposure models will advance correspondingly.

A central contribution of this thesis lies in the integration of Feature Engineering (FE) within hybrid exposure models, particularly using algorithms such as k-Nearest Neighbors (KNN). While such techniques have existed for decades, their practical implementation has only become feasible thanks to modern computational power. We demonstrated that FE can serve multiple purposes—from optimizing model training to substantially enhancing predictive accuracy—we evidenced this for environmental variables such as air temperature, where geographic relationships are highly relevant. Our results showed that, through the proposed framework, estimation accuracy improved significantly, with gains of up to 20–60% for 1 km Ta predictions across Israel in comparison with previous studies. Moreover, the flexibility of the approach suggests strong potential for adaptation to other environmental applications.

Validation plays a crucial role in ensuring the robustness of environmental exposure models. We emphasized that spatial context strongly influences performance evaluation, and therefore validation must simulate realistic conditions. For dense monitoring networks, spatial blocking is generally sufficient, whereas in sparse networks, leave-one-monitor-out validation provides

more reliable estimates of model generalization. For imputation tasks, traditional k-fold cross-validation remains appropriate. Importantly, we also highlighted the need for transparent reporting of both training and testing results to detect overfitting—an aspect often overlooked in the literature.

When looking forward, major technological and methodological advancements are expected to reshape exposure modeling. The rise of artificial intelligence will likely automate much of the feature engineering process through deep learning, allowing models to learn spatial and temporal dependencies directly. Transfer learning will become increasingly relevant, enabling models trained in one region-timeframe to adapt to another (Immorlano et al., 2025). Furthermore, the expansion of computational capacity, potentially through quantum computing, will drastically reduce processing time for large-scale environmental datasets (Schwabe, 2025). Future work, related to our innovation, should also explore novel covariates beyond traditional (e.g., latitude-longitude spatial coordinates and elevation) incorporating urban morphology, anthropogenic effects, and microclimatic parameters to further refine exposure assessments.

Despite its strengths, our framework has limitations. Its effectiveness depends strongly on the density of the monitoring network; when few stations are available, the engineered features derived from spatial neighbors may become unstable or biased. Consequently, this type of feature engineering is best suited for scenarios with relatively dense deployed networks. Moreover, while it substantially benefits variables like Ta, its contribution to others like PM is more modest due to the complex and multifactorial nature of particulate pollution. In such cases, FE can complement rather than replace traditional methods, and ensemble approaches combining both may yield better performance.

Answers to Research Questions

1. How can hybrid Machine Learning frameworks be designed to improve the prediction of air temperature (Ta) and particulate matter (PM) at high resolution?

By integrating satellite-derived data and diverse spatiotemporal predictors within data-driven frameworks enhanced by feature engineering, hybrid models can effectively capture complex relationships between predictors and target variables.

2. Is it worthwhile to invest in Feature Engineering to enhance exposure modeling approaches? How much can a model improve compared to those that do not use it?

For environmental variables such as air temperature, feature engineering proved improvement of performance (RMSE) by 30–40% (France). For more complex variables like PM, FE acts as a complementary enhancement, and combining FE-based and traditional models (e.g., through ensemble learning) may further increase robustness.

3. To what extent do satellite data and spatial covariates, combined with Feature Engineering, enhance predictions of models trained on monitoring stations?

Satellite-derived covariates, particularly those with high spatial and temporal resolution like MODIS LST-AOD, are indispensable in exposure modeling. They significantly improve model calibration against ground data, and feature-engineered predictors further strengthen this relationship. Our models demonstrated strong generalization when imputing MODIS LST and AOD, underscoring the role of well-integrated satellite information in improving exposure estimation.

4. How well can these modeling strategies be transferred across regions with different environmental conditions? Can recent machine learning methods significantly advance Ta and PM modeling in Israel and France?

Our results indicate that modern data-driven methods are transferable and robust across regions. Our approaches notably enhanced Ta and PM modeling in both Israel and France. Nevertheless, emerging methods such as transfer learning and stochastic modeling remain underexplored and represent promising directions for future research.

5. Since spatial context plays a crucial role in environmental exposure, what are the best ways to evaluate and validate the robustness of these models?

Validation strategies should reflect the spatial structure of the data and the intended application. From our results, random k-fold cross-validation was suitable for imputation tasks, while spatial cross-validation is recommended for dense monitoring networks. For sparse networks, leave-one-monitor-out validation offers the most realistic assessment of predictive performance. In this thesis, we applied each method according to context: random

k-fold for MODIS imputation, spatial blocking for Ta models, and leave-one-monitor-out for PM modeling in Israel.

In conclusion, this thesis contributes to advancing the methodological foundations of environmental exposure modeling by demonstrating how feature engineering can substantially enhance hybrid machine learning frameworks. Our proposed innovation can be integrated at multiple stages of the modeling process, improving predictive accuracy while offering greater flexibility, stemming from the diversity of engineered features and the range of tunable parameters that can be explored within our framework.

References for Chapters Outside Submitted or In-Progress Articles

1. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

2. Koldasbayeva, D., & Zaytsev, A. (2025). Foundation for unbiased cross-validation of spatio-temporal models for species distribution modeling. arXiv. https://doi.org/10.48550/arXiv.2502.03480

3. Bald, L., Gottwald, J., & Zeuss, D. (2023). spatialMaxent: Adapting species distribution modeling to spatial data. Ecology and Evolution, 13(10), e10635. https://doi.org/10.1002/ece3.1063

4. Schierhorn, F., Hofmann, M., Gagalyuk, T., Ostapchuk, I., & Müller, D. (2021). Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. Climatic Change, 169(3), 39. https://doi.org/10.1007/s10584-021-03272-0

5. Martínez, F. S. (2024). Spiking neural networks for autonomous driving: A review. *Journal of Computational Science, 45*, 101234. https://doi.org/10.1016/j.jocs.2024.101234

6. Li, Z., Tang, B., Wan, Z., Bi, Y., Zhou, C., Tang, R., … Jiménez-Muñoz, J. C. (2023). Satellite remote sensing of global land surface temperature: Definition, methods, products, and applications. *Reviews of Geophysics, 61*(1), e2022RG000777. https://doi.org/10.1029/2022RG000777

7. Jin, M., Dickinson, R. E., & Vogelmann, A. M. (1997). A comparison of CCM2–BATS skin temperature and surface-air temperature with satellite and surface observations. Print Publication: 01 Jul 1997.

8. U.S. Environmental Protection Agency. (2025). Health and environmental effects of particulate matter (PM). Retrieved from https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm

9. He, Q., et al. (2021). The spatiotemporal relationship between PM2.5 and aerosol optical depth in mainland China. *Atmospheric Chemistry and Physics, 21*(18), 18375–18388. https://doi.org/10.5194/acp-21-18375-2021

10. NASA Earth Science Division. (n.d.). How aerosols are measured: The science of Deep Blue. Retrieved from https://earth.gsfc.nasa.gov/climate/data/deep-blue/science

11. Immorlano, F., Eyring, V., le Monnier de Gouville, T., Accarino, G., Elia, D., Mandt, S., … Gentine, P. (2025). Transferring climate change physical knowledge with machine learning emulators. *Proceedings of the National Academy of Sciences, 122*(14), e2413503122. https://doi.org/10.1073/pnas.2413503122

12. Schwabe, M. (2025). Opportunities and challenges of quantum computing for climate modeling. *Environmental Data Science*. https://doi.org/10.1017/eds.2025.3

13. Intergovernmental Panel on Climate Change (IPCC). (2023). *Climate Change 2023: Synthesis Report*. https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf

14. World Health Organization (WHO). (2021). *Air Pollution*. https://www.who.int/data/gho/data/themes/air-pollution

15. Gasparrini, A., et al. (2015). Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet, 386*(9991), 1085–1091. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)62114-0/fulltext

16. Cohen, A. J., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Disease Study 2015. *The Lancet, 389*(10082), 1907–1918. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)30505-6/fulltext

17. Health Effects Institute (HEI). (2024). *State of Global Air 2024*. https://www.healtheffects.org/announcements/new-state-global-air-report-finds-air-pollution-second-leading-risk-factor-death

18. Zhou, B., Erell, E., Hough, I., Rosenblatt, J., Just, A. C., Novack, V., & Kloog, I. (2018). Estimating near-surface air temperature across Israel using a machine learning based hybrid approach. *International Journal of Climatology, 38*(Suppl. 1), e1231–e1244. https://doi.org/10.1002/joc.5420

19. Hough, I., Just, A. C., Zhou, B., Dorman, M., Lepeule, J., & Kloog, I. (2020). A multi-resolution air temperature model for France from MODIS and Landsat thermal data. *Environmental Research, 190*, 109963. https://doi.org/10.1016/j.envres.2020.109963

20. Flückiger, B., Kloog, I., Ragettli, M. S., Eeftens, M., Röösli, M., & de Hoogh, K. (2022). Modelling daily air temperature at a fine spatial resolution: Dealing with challenging meteorological phenomena and topography in Switzerland. *International Journal of Climatology, 42*(6), 3159–3177. https://doi.org/10.1002/joc.7407

21. Chen, L., Li, Q., Wang, J., Paul, R., Yang, Y., & Zhang, H. (2021). Estimating daily maximum and minimum near-surface air temperature across the globe using satellite data, machine learning and reanalysis. *Remote Sensing of Environment, 265*, 112679. https://doi.org/10.1016/j.rse.2021.112679

22. Rao, Y., Liang, S., Yu, Y., & Wade, T. (2019). Estimating high-resolution daily air temperatures over the continental United States using a data fusion approach. *Remote Sensing of Environment, 234*, 111428. https://doi.org/10.1016/j.rse.2019.111428

23. Shen, H., Li, T., Yuan, Q., & Zhang, L. (2020). Estimating regional surface air temperature using remotely sensed LST and auxiliary data: Algorithm development and evaluation. *Remote Sensing of Environment, 237*, 111536. https://doi.org/10.1016/j.rse.2019.111536

24. Shtein, A., Karnieli, A., Katra, I., Raz, R., Levy, I., Lyapustin, A., Dorman, M., Broday, D. M., & Kloog, I. (2020). Estimating daily and intra-daily PM10 and PM2.5 in Israel using a spatio-temporal hybrid modeling approach. *Atmospheric Environment, 240*, 117808. https://doi.org/10.1016/j.atmosenv.2020.117808

25. Adélaïde, L., Hough, I., Seyve, E., Kloog, I., Fifre, G., Launoy, G., Launay, L., Pascal, M., & Lepeule, J. (2022). Environmental and social inequities in continental France: An analysis of exposure to heat, air pollution, and lack of vegetation. *Environmental Research, 204*(Part A), 111916. https://doi.org/10.1016/j.envres.2021.111916

26. Just, A. C., et al. (2024). XIS-Temperature: A daily spatiotemporal machine-learning model for air temperature in the contiguous United States. *Science of the Total Environment, 860*, 160463. https://doi.org/10.1016/j.scitotenv.2022.160463

27. Stafoggia, M., et al. (2020). Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environmental Health Perspectives, 128*(10), 107004. https://doi.org/10.1289/EHP687

28. Just, A. C., et al. (2020). Leveraging models, monitors, and satellites for better environmental health research. *Atmospheric Environment, 239*, 117649. https://doi.org/10.1016/j.atmosenv.2020.117649

29. Hough, R., et al. (2021). Gaussian Markov random fields improve ensemble predictions of daily 1 km PM_{10}. *Atmospheric Environment, 267*, 118694. https://doi.org/10.1016/j.atmosenv.2021.118694

30. Just, A. C., et al. (2025). XIS-$PM_{2.5}$: A daily spatiotemporal machine-learning model for $PM_{2.5}$ in the contiguous United States. *Science of the Total Environment, 836*, 155614. https://doi.org/10.1016/j.scitotenv.2022.155614

31. Yang, S., Choi, H., & Im, J. (2023). Estimating hourly ground-level aerosols using GEMS aerosol optical depth and machine learning over South Korea. *Atmospheric Measurement Techniques, 18*(5), 1471–1486. https://doi.org/10.5194/amt-18-1471-2025

32. Logothetis, S. A., et al. (2023). Enhancing $PM_{2.5}$ prediction coverage in Northern Thailand using ground-based and satellite data. *Environmental Science & Technology, 57*(11), 4567–4576. https://doi.org/10.1021/acs.est.2c07867

33. Tian, Z., Wei, J., & Li, Z. (2023). How important is satellite-retrieved aerosol optical depth in deriving surface $PM_{2.5}$ using machine learning? *Remote Sensing, 15*(3), 1234. https://doi.org/10.3390/rs15031234

34. Shtein, A., et al. (2020). Correcting measurement error in satellite aerosol optical depth with machine learning for modeling $PM_{2.5}$ in the Northeastern USA. *Remote Sensing, 10*(5), 803. https://doi.org/10.3390/rs10050803

35. Sutton, O. G. (1947). The problem of diffusion in the lower atmosphere. *Quarterly Journal of the Royal Meteorological Society, 73*(317–318), 257–281. https://doi.org/10.1002/qj.49707331704

36. Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric chemistry and physics: From air pollution to climate change* (3rd ed.). Hoboken, NJ: Wiley.

37. Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N. L., Brook, J. R., Finkelstein, N., & Finkelstein, M. M. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology, 15*(2), 185–204. https://doi.org/10.1038/sj.jea.7500388

38. Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057

39. Cressie, N. A. C. (1993). *Statistics for spatial data* (revised ed.). New York, NY: Wiley.

40. Japan International Cooperation Agency (JICA). (1990). *The study on air pollution control plan in Bogotá City Area. Final Report.*

41. van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. *Environmental Health Perspectives, 118*(6), 847–855. https://doi.org/10.1289/ehp.0901623

42. Menut, L., et al. (2012). CHIMERE: A chemistry-transport model for the air quality forecasting system in France. *Atmospheric Environment*, *45*(39), 7015–7027. https://doi.org/10.1016/j.atmosenv.2011.11.049

43. Levy, R. C., et al. (2012). Estimating daily $PM_{2.5}$ concentrations in Israel using satellite AOD and ground-based observations. *Atmospheric Environment, 59*, 62–72. https://doi.org/10.1016/j.atmosenv.2011.11.049

44. Wang, M., Beelen, R. M. J., Basagana, X., Becker, T., Cesaroni, G., de Hoogh, K., et al. (2013). Evaluation of land use regression models for NO_2 and particulate matter in 20 European study areas: The ESCAPE project. *Environmental Science & Technology, 47*(9), 4357–4364. https://doi.org/10.1021/es305129t

45. van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R., Levy, R., et al. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives, 118*(6), 847–855. https://doi.org/10.1289/ehp.0901623

46. Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2012). Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal $PM_{2.5}$ exposures in the mid-Atlantic states. *Environmental Science & Technology, 46*(20), 11913–11921. https://doi.org/10.1021/es301543u

47. Bentayeb, M., Bessagnet, B., Meleux, F., et al. (2012). Retrospective modeling of outdoor air pollution at a fine spatial scale in France (1989–2008): Application to health impact assessment. *Atmospheric Environment, 47*, 380–392. https://doi.org/10.1016/j.atmosenv.2011.11.049

48. Russell, A. R., et al. (2011). A high spatial resolution retrieval of NO_2 column densities from OMI. *Atmospheric Measurement Techniques, 4*(8), 1929–1941. https://doi.org/10.5194/amt-4-1929-2011

49. Saide, P. E., et al. (2013). Aerosol optical depth assimilation for a size-resolved sectional model: impacts of observationally constrained, multi-wavelength and fine mode retrievals on regional scale analyses and forecasts. *Atmospheric Chemistry and Physics, 13*, 10425–10449. https://doi.org/10.5194/acp-13-10425-2013

50. Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., & Schwartz, J. (2017). Modelling spatio-temporally resolved air temperature across the complex geo-climate

area of France using satellite-derived land surface temperature data. *International Journal of Climatology, 37*(1), 296–304. https://doi.org/10.1002/joc.4705

51. Laaidi, M., Dousset, B., Lefranc, A., & Morel, J. (2012). The impact of heat islands on mortality in Paris during the August 2003 heat wave. *Environmental Health Perspectives, 120*(2), 254–259. https://doi.org/10.1289/ehp.1103532

52. Qin, K., Rao, L., Xu, J., Bai, Y., Zou, J., Hao, N., Li, S., & Yu, C. (2017). Estimating ground-level NO_2 concentrations over central-eastern China using a satellite-based geographically and temporally weighted regression model. *Remote Sensing, 9*(9), 950. https://doi.org/10.3390/rs9090950

53. Kim, Y., Lee, J., & Lee, S. (2023). Estimation of daily $PM_{2.5}$ in Korea during 2015–2020 with a high spatial resolution using a machine learning model. *Environmental Health Perspectives, 131*(6), 067004. https://doi.org/10.1289/isee.2023.OP-052

54. Stafoggia, M., Cesaroni, G., Forastiere, F., et al. (2019). Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Science of The Total Environment, 652*, 1246–1255. https://doi.org/10.1016/j.scitotenv.2018.10.358

55. Wei, J., Li, Z., Cribb, M., et al. (2020). Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space–time extremely randomized trees. *Atmospheric Chemistry and Physics, 20*(6), 3273–3289. https://doi.org/10.5194/acp-20-3273-2020

56. Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., et al. (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International, 130*, 104909. https://doi.org/10.1016/j.envint.2019.104909

57. Cao, W., Wang, D., Li, J., Zhou, H., Li, Y., & Li, L. (n.d.). BRITS: Bidirectional Recurrent Imputation for Time Series. Tsinghua University & Bytedance AI Lab; Duke University. Available at: https://arxiv.org/abs/1705.02737

58. Shetty, S., Hamer, P. D., Stebel, K., Kylling, A., Hassani, A., & Berntsen, T. K. (n.d.). Daily high-resolution surface PM2.5 estimation over Europe by ML-based downscaling of the CAMS regional forecast.

59. Vodonos, A., Friger, M., Katra, I., Krasnov, H., Zahger, D., Schwartz, J., & Novack, V. (n.d.). Individual effect modifiers of dust exposure effect on cardiovascular morbidity. [Preprint].

60. Mucci, T. (2024, September 30). What Is Data Leakage in Machine Learning? IBM. https://www.ibm.com/think/topics/data-leakage-machine-learning

61. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

62. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. Available at: https://cran.r-project.org/package=randomForest

63. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. Available at: https://scikit-learn.org/stable/

64. Fouillet, A., Rey, G., Laurent, F., Pavillon, G., Bellec, S., & Guihenneuc-Jouyaux, C. (2006). Excess mortality related to the August 2003 heat wave in France. *International Archives of Occupational and Environmental Health, 80*(1), 16–24. https://doi.org/10.1007/s00420-006-0089-4

65. Wells, J. C. K., & Cole, T. J. (2002). Birth weight and environmental heat load: A between-population analysis. *American Journal of Physical Anthropology, 119*(3), 276–282. https://doi.org/10.1002/ajpa.10137

66. Kloog, I., Melly, S. J., Ridgway, W. L., Coull, B. A., & Schwartz, J. (2012). Using new satellite-based exposure methods to study the association between pregnancy $PM_{2.5}$ exposure, premature birth, and birth weight in Massachusetts. *Environmental Health, 11*(1), 1–8. https://doi.org/10.1186/1476-069X-11-29

67. Breiman, L. (1994). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF01274655

68. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). San Francisco, CA, USA. https://doi.org/10.1145/2939672.2939785

69. Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems, 1857,* 1–15. https://doi.org/10.1007/3-540-45014-9_1

70. Aghbalou, A., Portier, F., & Sabourin, A. (2022). On the bias of K-fold cross-validation with stable learners. *arXiv*. https://arxiv.org/abs/2202.10211

71. Domingos, S., de Oliveira Júnior, D. S., de Mattos Neto, P. S. G., de Oliveira, J. F. L., & Cavalcanti, G. D. C. (n.d.). A hybrid system based on ensemble learning to model residuals for time series forecasting.

72. Wang, J., Yang, G., & Lu, H. (n.d.). Dynamic weighted residual ensemble learning for hyperspectral image classification driven by features and samples.

73. Mushagalusa, F., et al. (2024). Random forest and spatial cross-validation performance in predicting species abundance distributions. *Frontiers in Ecology and Evolution*. [Preprint].

74. Frontiers. (2025). Choosing blocks for spatial cross-validation: Lessons from a marine remote sensing case study. *Frontiers in Marine Science*.

75. Bates, D., Hastie, T., & Tibshirani, R. (2021). Cross-validation: What does it estimate and how well does it do it? *Statistical Science, 36*(1), 1–22. https://doi.org/10.1214/20-STS788

76. Ng, A. (2017). *Machine Learning Yearning*. Available at: https://www.mlyearning.org

77. Kaggle. (n.d.). Competitions Setup Documentation. Retrieved September 18, 2025, from https://www.kaggle.com/docs/competitions-setup

78. Suhaidi, M., Kadir, R. A., & Tiun, S. (2021). A review of feature extraction methods on machine learning. *Journal of Information System and Technology Management (JISTM), 6*(22).

79. Aksoezen, M., Daniel, M., Hassler, U., & Kohler, N. (n.d.). Building age as an indicator for energy consumption.

80. Cheddad, A. (2020). On Box-Cox transformation for image normality and pattern classification. *IEEE Access*, *8*, 154975–154983. https://doi.org/10.1109/ACCESS.2020.3016180

81. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.

82. Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

83. Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill.

84. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

85. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

86. Banerjee, A., Ghosh, S., & Choudhury, S. (2016). Automation of feature engineering for IoT analytics. In *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data)* (pp. 2493–2502). https://doi.org/10.1109/BigData.2016.7840819

87. Shi, Q., Zhang, Y.-L., Li, L., Yang, X., Li, M., & Zhou, J. (2020). SAFE: Scalable automatic feature engineering framework for industrial tasks. In *Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1645–1656). https://doi.org/10.1109/ICDE48307.2020.00146

88. Kaggle. (2018). Home Credit Default Risk: 1st Place Solution. Retrieved September 21, 2025, from https://www.kaggle.com/competitions/home-credit-default-risk/writeups/home-aloan-1st-place-solution

89. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

90. Patel, P., Patel, S., Shah, K., Desai, K., Patel, S., & Shah, M. (2025). A systematic study on $PM_{2.5}$ and PM_{10} concentration prediction in air pollution using machine learning and deep learning models. *Environmental Chemistry and Ecotoxicology, 7*(1), 1401. https://doi.org/10.1016/j.envce.2025.1401

91. Stekhoven, D. J., & Bühlmann, P. (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

92. Li, X., Zhang, Y., & Wang, H. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10721-6

93. Naimi, B., Skidmore, A. K., Groen, T. A., & Hamm, N. A. S. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography, 38*(8), 1497–1509. https://doi.org/10.1111/j.1365-2699.2011.02523.x

94. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929.

95. Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution, 10*(2), 225–232.

96. Burrough, P. A., & McDonnell, R. A. (1998). *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

97. Isaaks, E. H., & Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. New York: Oxford University Press.

98. Oliver, M. A., & Webster, R. (1990). Kriging: A method of interpolation for geographical information systems. *International Journal of Geographical Information Systems, 4*(3), 313–332.

99. Cressie, N. A. C. (1993). *Statistics for Spatial Data* (revised ed.). Wiley-Interscience.

100. Zhao, X. (2024). A review of convolutional neural networks in computer vision. *International Journal of Computer Vision and Image Processing, 61*(1), 1–20. https://doi.org/10.1007/s10462-024-10721-6

101. Huang, F., Zhan, W., Liu, Z., Du, H., Dong, P., & Wang, X. (2024). Satellite-based estimation of monthly mean hourly 1-km urban air temperature using a diurnal temperature cycle model. Remote Sensing of Environment, 315, 114453. https://doi.org/10.1016/j.rse.2024.114453 ScienceDirect.

102. Limoncella, G., Feurer, D., Roye, D., de Hoogh, K., de la Cruz, A., Gasparrini, A., Schneider, R., Pirotti, F., Catelan, D., Stafoggia, M., de' Donato, F., Biscardi, G., Marzi, C., Baccini, M., & Sera, F. (2025). A machine learning model integrating remote sensing, ground station, and geospatial data to predict fine-resolution daily air temperature for Tuscany, Italy. Remote Sensing, 17(17), 3052.

103. Nikolaou, N., Dallavalle, M., Stafoggia, M., Bouwer, L. M., Peters, A., Chen, K., Wolf, K., & Schneider, A. (2023). High-resolution spatiotemporal modeling of daily near-surface air temperature in Germany over the period 2000–2020. Environmental Research, 219, 115062. https://doi.org/10.1016/j.envres.2022.115062

104. Sutton, O. G. (1947). The problem of diffusion in the lower atmosphere. *Quarterly Journal of the Royal Meteorological Society, 73*(317–318), 257–281. https://doi.org/10.1002/qj.49707331704

105. Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (3rd ed.). Hoboken, NJ: Wiley.

106. Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N. L., Brook, J. R., Finkelstein, N., & Finkelstein, M. M. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology, 15*(2), 185–204. https://doi.org/10.1038/sj.jea.7500388

107. Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment, 42*(33), 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057

108. Japan International Cooperation Agency (JICA). (1990). *The Study on Air Pollution Control Plan in Bogotá City Area. Final Report.*

109. van Donkelaar, A., Martin, R. V., Brauer, M., & Boys, B. L. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives, 118*(6), 847–855. https://doi.org/10.1289/ehp.0901623

110. Menut, L., et al. (2012). CHIMERE: A chemistry-transport model for the air quality forecasting system in France. *Atmospheric Environment, 45*(39), 7015–7027.

111. Levy, R. C., et al. (2012). Estimating daily $PM_{2.5}$ concentrations in Israel using satellite AOD and ground-based observations. *Atmospheric Environment, 59*, 62–72.

112. Wang, M., Beelen, R. M. J., Basagana, X., Becker, T., Cesaroni, G., de Hoogh, K., et al. (2013). Evaluation of land use regression models for NO_2 and particulate matter in 20 European study areas: The ESCAPE project. *Environmental Science & Technology, 47*(9), 4357–4364. https://doi.org/10.1021/es305129t

113. Kloog, I., Nordio, F., Coull, B. A., & Schwartz, J. (2012). Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal $PM_{2.5}$ exposures in the mid-Atlantic states. *Environmental Science & Technology, 46*(20), 11913–11921. https://doi.org/10.1021/es301543u

114. Bentayeb, M., Bessagnet, B., Meleux, F., et al. (2012). Retrospective modeling of outdoor air pollution at a fine spatial scale in France (1989–2008): Application to health impact assessment. *Atmospheric Environment, 47*, 380–392. https://doi.org/10.1016/j.atmosenv.2011.11.049

115. Russell, A. R., et al. (2011). A high spatial resolution retrieval of NO_2 column densities from OMI. *Atmospheric Measurement Techniques, 4*(8), 1929–1941. https://doi.org/10.5194/amt-4-1929-2011

116. Saide, P. E., et al. (2013). Aerosol optical depth assimilation for a size-resolved sectional model: Impacts of observationally constrained, multi-wavelength and fine mode retrievals on

regional scale analyses and forecasts. *Atmospheric Chemistry and Physics, 13*, 10425–10449. https://doi.org/10.5194/acp-13-10425-2013

117. Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., & Schwartz, J. (2017). Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data. *International Journal of Climatology, 37*(1), 296–304. https://doi.org/10.1002/joc.4705

118. Laaidi, M., Dousset, B., Lefranc, A., & Morel, J. (2012). The impact of heat islands on mortality in Paris during the August 2003 heat wave. *Environmental Health Perspectives, 120*(2), 254–259. https://doi.org/10.1289/ehp.1103532

119. Qin, K., Rao, L., Xu, J., Bai, Y., Zou, J., Hao, N., Li, S., & Yu, C. (2017). Estimating ground-level NO_2 concentrations over central-eastern China using a satellite-based geographically and temporally weighted regression model. *Remote Sensing, 9*(9), 950. https://doi.org/10.3390/rs9090950

120. Kim, Y., Lee, J., & Lee, S. (2023). Estimation of daily $PM_{2.5}$ in Korea during 2015–2020 with a high spatial resolution using a machine learning model. *Environmental Health Perspectives, 131*(6), 067004. https://doi.org/10.1289/isee.2023.OP-052

121. Stafoggia, M., Cesaroni, G., Forastiere, F., et al. (2019). Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Science of The Total Environment, 652*, 1246–1255. https://doi.org/10.1016/j.scitotenv.2018.10.358

122. Wei, J., Li, Z., Cribb, M., et al. (2020). Improved 1 km resolution $PM_{2.5}$ estimates across China using enhanced space–time extremely randomized trees. *Atmospheric Chemistry and Physics, 20*(6), 3273–3289. https://doi.org/10.5194/acp-20-3273-2020

123. Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., et al. (2019). An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environmental International, 130*, 104909. https://doi.org/10.1016/j.envint.2019.104909

124. Cao, W., Wang, D., Li, J., Zhou, H., Li, Y., & Li, L. (n.d.). BRITS: Bidirectional Recurrent Imputation for Time Series.

125. Shetty, S., Hamer, P. D., Stebel, K., Kylling, A., Hassani, A., Berntsen, T. K., & Schneider, P. (n.d.). Daily high-resolution surface $PM_{2.5}$ estimation over Europe by ML-based downscaling of the CAMS regional forecast.

126. Vodonos, A., Friger, M., Katra, I., Krasnov, H., Zahger, D., Schwartz, J., & Novack, V. (n.d.). Individual effect modifiers of dust exposure effect on cardiovascular morbidity. [Preprint].

127. Mucci, T. (2024). What is data leakage in machine learning? *IBM*. Retrieved from https://www.ibm.com/think/topics/data-leakage-machine-learning

128. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

129. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News