A Scale-Invariance-Based Algorithm Application for Land Surface Temperature Downscaling in Denmark

Élio Pereira *, Manvel Khudinyan, Inês Girão, Bruno Marques, Vitor F. V. V. de Miranda, Hjalte Jomo Danielsen Sørup, Quentin Paletta, Ana Patrícia Oliveira

Posted Date: 8 April 2026

doi: 10.20944/preprints202604.0495.v1

Keywords: urban climate; downscaling; land surface temperature; machine learning; scaleinvariance; residual correction; Sentinel-3; landsat

 doi:10.20944/preprints202604.0495.v1

Article

A Scale-Invariance-Based Algorithm Application for Land Surface Temperature Downscaling in Denmark

Élio Pereira [1,*], Manvel Khudynian [1], Inês Girão [1], Bruno Marques [1], Vitor F. V. V. de Miranda [1], Hjalte Jomo Danielsen Sørup [2], Quentin Paletta [3,4] and Ana Patrícia Oliveira [1]

[1] +ATLANTIC CoLAB, Peniche, Portugal
[2] Danish Meteorological Institute, Copenhagen, Denmark
[3] Φ-Lab, European Space Agency, ESRIN, Frascati, Italy
[4] Climate Team, European Space Agency, ECSAT, Harwell, UK
* Correspondence: elio.pereira@ist.utl.pt

Highlights

- In coarse prediction, the multi-timestamp Machine Learning models, in particular Gradient Boosting, performed as good or better than the benchmarking single-timestamp Linear Regression model.
- In fine prediction, all multi-timestamp models (specially the tree-based ones) performed worse than the single-timestamp Linear Regression model which suggests that training with coarse data from multiple timestamps may deteriorate downscaling performance, compared to training timestamp-specific models.
- The single-timestamp Linear Regression model proved to be the best downscaling method, producing the smallest errors. Also, even though the single-timestamp Linear Regression model must be re-trained for every single timestamp, its architecture is remarkably simple, making it highly recommendable for operations.
- Although the assumed principle of scale-invariance was found to not hold, residual correction was quite effective in the reduction of its error (in average, by 0.33 K in RMSE), allowing to attest its fitness for applications where scalability is the most important feature.

Abstract

With an ever-growing recognition of Land Surface Temperature (LST) as a key Essential Climate Variable (ECV), it becomes utmost important to have such a variable at both the fine spatial and temporal scales of urban spaces and dynamics. Sentinel-3 provides coarse LST (1 km, daily) based on thermal imagery acquired by its Sea and Land Surface Temperature Radiometer (SLSTR) as well as fine Spectral Directional Reflectances (SDR, 300 m, every two days) synergically inferred from both SLSTR and the optical bands acquired through the Ocean and Land Colour Instrument (OLCI), which gives opportunity for using the latter as predictor in the downscaling of the former. Herein, two scale-invariance-based architectures were developed: a single-timestamp model, trained with the coarse data of the timestamp whose fine target it tries to infer; and a multi-timestamp one, trained with several timestamps and that can infer for any other. While for the case of the multi-timestamp architecture, Machine Learning (ML) models besides Linear Regression (LR) were trained, solely LR was considered for the single-timestamp architecture due to the smaller amount of data available, making it less suitable for hyperparameter tuning. The models were developed over four Danish Functional Urban Areas (FUAs) between 2020 and 2023 using SRD-derived indices, seasonal and geospatial predictors. From 112 Sentinel-3 scenes, 105 were used for training and 7 for validation against Landsat data. While Gradient Boosting (GB) achieved the best coarse-scale performance (test set Root Mean Square Error, RMSE, of 1.56 K), fine-scale predictions showed degraded performance, indicating scale-invariance breakdown. Tree-based models performed poorly due to extrapolation limitations, whereas Neural Net (NN) and LR proved more robust. After residual correction, single-timestamp LR achieved the best fine-scale performance (test set RMSE of 1.40 K), making it the most reliable and operationally recommended architecture.

 doi:10.20944/preprints202604.0495.v1

Keywords: Urban climate; downscaling; land surface temperature; machine learning; scale-invariance; residual correction; Sentinel-3; landsat

1. Introduction

The increasing accessibility of thermal satellite observations, combined with the recognition of Land Surface Temperature (LST) as a key Essential Climate Variable (ECV) [1], for climate, urban, and environmental applications, have intensified efforts to improve the spatial resolution of satellite-derived thermal information [2]. Indeed, LST plays a central role in the assessment of surface-atmosphere interactions, urban climate dynamics, and heat-related impacts. However, its operational use remains constrained by limitations inherent to current Earth Observation (EO) systems, such as sub-optimal acquisition time at high spatial resolution and vice-versa, the necessity of quasi-clear-sky conditions for reliable acquisition, and the degrees of uncertainty in the LST conversion algorithms, which are still dependent on assumptions/inputs.

Freely accessible thermal infrared imagery from satellite missions such as Landsat, Terra/Aqua, and Sentinel-3 has enabled widespread access to LST datasets at spatial resolutions ranging from approximately 100 m to 1 km, with global coverage and multi-decadal continuity. These characteristics have supported a broad range of urban climate and surface heat studies across diverse geographical contexts [3–8]. However, satellite-derived surface thermal products predominantly characterise the surface urban heat island (SUHI), which exhibits spatial structures, temporal behaviour, and magnitudes that differ substantially from those of atmospheric urban heat islands. Consequently, SUHI metrics should not be interpreted as a direct surrogate for near-surface atmospheric heat conditions, but rather as an ancillary information layer in UHI assessment [6,9–12].

A major limitation in the application of SUHI for urban analysis arises from the trade-off between spatial resolution and temporal sampling, often described as the granularity versus spatial resolution dilemma [13,14]. Urban-scale studies require sub-kilometre resolution to adequately represent the heterogeneity of urban morphology and surface thermal behaviour. While such spatial details can be provided by Low Earth Orbit (LEO) satellites, routinely available high-resolution thermal imagery is largely restricted to the Landsat missions [15], which provide long-term LST products but are characterised by revisit intervals of approximately 16 days per satellite (8 days if Landsat 8 and Landsat 9 are considered together). Such an interval severely constrains the systematic monitoring of rapidly evolving phenomena, such as urban thermal responses during extreme heat events.

Beyond revisit frequency, the practical exploitation of satellite thermal data is further limited by cloud contamination, uncertainties associated with atmospheric correction, the limited availability of in situ observations for LST validation, sensor-specific spatial resolution and viewing geometry, and orbital characteristics that determine overpass timing [6,13]. In addition, publicly available Landsat thermal imagery is predominantly acquired during daytime descending orbits, corresponding to mid-morning conditions at mid-latitudes. Such overpass timing is generally unfavourable for the SUHI detection, as nocturnally stored urban heat has largely dissipated, while incoming solar radiation has not yet generated strong urban–rural thermal contrasts [16].

In response to these combined spatial and temporal constraints, a substantial body of research has explored statistical and physically informed approaches to downscale LST from kilometre-scale satellite products to finer spatial resolutions. A wide range of LST downscaling and disaggregation approaches has been developed, spanning empirical/statistical thermal sharpening, physically based energy-balance formulations, and machine-learning-driven (often multi-source fusion) methods, each with distinct trade-offs in urban contexts [17]. Physically based methods, including energy-balance-driven or radiative transfer-informed approaches, aim to explicitly represent surface–atmosphere exchanges and urban thermal processes, offering stronger physical interpretability and temporal consistency [18]. However, these methods typically require detailed ancillary data (e.g. surface emissivity, aerodynamic resistance, urban morphology parameters) that are rarely available

 doi:10.20944/preprints202604.0495.v1

at adequate resolution and coverage for large-scale or operational urban applications [16,19,20]. More broadly, Machine Learning (ML) and deep learning methods can capture non-linear interactions between urban morphology, land cover, and thermal dynamics and often outperform linear models in reproducing fine-scale hotspots, but may suffer from limited transferability across cities/seasons and require explicit uncertainty and validation strategies to support operational use [21,22]. One of the most widely adopted techniques is thermal sharpening, which exploits empirical relationships between land surface temperature and vegetation-related indicators, such as the Normalised Difference Vegetation Index (NDVI) or Fractional Vegetation Cover (FVC) [23,24]. Commonly referred to as DisTrad and TsHARP, respectively, these approaches assume that the relationship between LST and vegetation indices observed at coarse spatial resolutions remains invariant when transferred to finer scales [25,26]. While thermal sharpening has demonstrated robust performance in vegetated and semi-vegetated landscapes, its underlying assumptions may be challenged in highly heterogeneous urban environments, where surface materials, building morphology, and anthropogenic heat sources introduce additional complexity. Nevertheless, it is an efficient method for replicating downscaling routines at scale, since it does not require additional data inputs than the co-registered optical and thermal imagery products.

This paper presents the results of employing DisTrad and TsHARP-like models in Denmark, developed to reveal urban patterns of the SUHI, and how these reflect land use/land cover features, to support of Danish cities in employing Nature-Base Solutions (NBS) to adapt to climate change. In this context, the present study investigates whether the integration of ML techniques into a scale-invariance-based LST downscaling framework outperforms conventional linear regression approaches, particularly when extending inference across multiple timestamps. By benchmarking single-timestamp and multi-timestamp architectures, and by evaluating linear and non-linear models both with and without residual correction, this work aims to clarify which options lead to improved LST downscaling model performance. The findings contribute to the ongoing debate on the suitability of data-driven methods for LST downscaling by highlighting the trade-offs between flexibility, robustness, and physical consistency, and by providing practical guidance for the design of reliable LST downscaling pipelines in urban environments, over four Danish Functional Urban Areas (Aalborg, Aarhus, Copenhagen and Odense).

2. Materials and Methods

2.1. Materials

2.1.1. Data Curation

Satellite imagery was used as both model input and reference data for validation. LST and optical-based indices (as predictors) were acquired from official satellite products provided by the Sentinel-3 and Landsat 8/9 missions. Sentinel-3 LST was obtained from the SLSTR Level-2 Non-Time Critical (NTC) product [27], delivering atmospherically corrected LST at approximately 1 km spatial resolution with a daily revisit frequency enabled by the dual-satellite constellation, making it suitable for spatio-temporal modelling and seasonal analysis. Landsat 8 and 9 LST scenes were acquired from the Collection 2 Level-2 Surface Temperature product through the Earth Explorer data portal of the United States Geological Survey [28] which provides physically corrected LST at 30 m (resampled from Thermal Infrared Radiometer Sensor (TIRS), with a native resolution of 100 m, to match the multispectral optical bands [15]) with a 16-day revisit cycle per satellite, offering fine-scale thermal detail for independent validation only.

In parallel, spectral indices were derived from the Sentinel-3 Synergy Level-2 (SYN) reflectance product [29]. This dataset combines reflectance measurements from the Ocean and Land Colour Instrument (OLCI) with atmospheric correction information derived from SLSTR, producing spatially and radiometrically coherent surface reflectance fields with 300 meters of resolution. The set of spectral indices calculated was:

 doi:10.20944/preprints202604.0495.v1

- The Normalised Difference Vegetation Index (NDVI), which corresponds to the difference between the surface directional reflectance in the near-infrared (NIR) and the one in the red ranges of the spectrum, divided by their sum:

$$\text{NDVI} = \frac{R_{\text{NIR}} - R_{\text{Red}}}{R_{\text{NIR}} + R_{\text{Red}}} \quad (1)$$

- The Normalised Difference Water Index (NDWI), which corresponds to the difference between surface directional reflectances in the green and in near-infrared ranges of the spectrum, divided by their sum:

$$\text{NDWI} = \frac{R_{\text{Green}} - R_{\text{NIR}}}{R_{\text{Green}} + R_{\text{NIR}}} \quad (2)$$

- The Fractional Vegetation Cover (FVC), which at some pixel in scene is obtained according to Agam et al. [30] and as suggested by Choudhury et al. [31] given by $\text{FVC} = 1 - \left(\frac{\text{NDVI}_{\text{max}} - \text{NDVI}}{\text{NDVI}_{\text{max}} - \text{NDVI}_{\text{min}}}\right)^{0.625}$, (3)

where NDVI_{min} and NDVI_{max} are the minimum and maximum NDVI values in the scene, respectively.

For the construction of spatio-temporal series of LST and SYN-derived indices, an automated pipeline was implemented to systematically download available imagery using OData API from Copernicus Data Space Ecosystem (CDSE) [32]. More precisely SLSRT LST and SYN products within designated Sentinel-3 orbits that cover Denmark, clipping data to the specified Area of Interest (AOI) and assessing its cloud cover. Only images with ≤ 5% cloud cover in the AOI were considered in the analysis. Also, solely data collected from year 2020 to 2023 were included. This filtering resulted in a total of 112 timestamps. Because of the cloud cover fraction criterion, most of the timestamps were obtained in summer and spring (55 and 40, respectively, in contrast with autumn's 14 and winter's 3) - as shown in **Figure 1**.

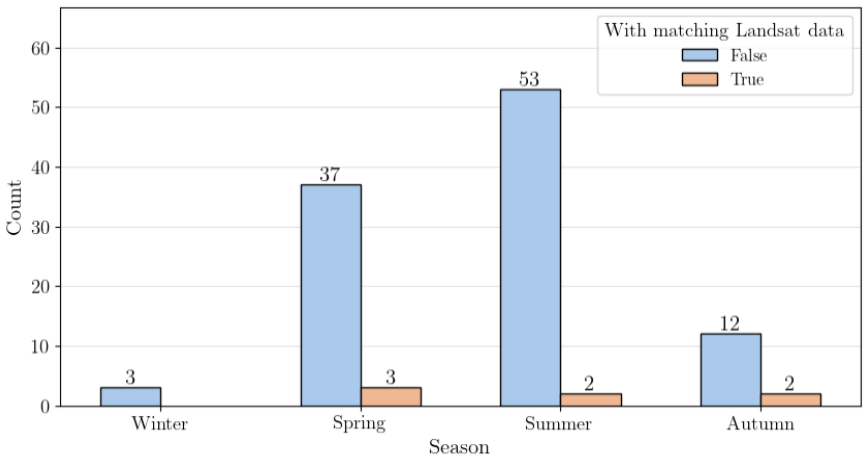


Figure 1. Seasonal distribution of the collected Sentinel-3 data.

Following the assembly of the Sentinel-3 collection, an automated process identified corresponding Landsat 8/9 images within a ±30-minute window of the Sentinel-3 acquisition times. This approach was designed to ensure temporal quasi-synchronisation, enhancing the matching quality of inputs for downscaling validation. As a result, a total of 7 unique matching dates – described in **Table 1** and **Table A1** of **Appendix A.1** – were identified: 3 in spring, 2 in summer, 2 in autumn and none in winter. This seasonal coverage, while not comprehensive, still enables testing in different seasons, in particular, the ones when heat is a more pressing concern.

Table 1. Matching dates between Sentinel-3 and Landsat 8/9 and corresponding time difference.

Sentinel Timestamp	Landsat Timestamp	Landsat Path/Row	Time difference in minutes

 doi:10.20944/preprints202604.0495.v1

5/30/2020 10:17	30/05/2020 10:19	L8 196/20-21	2.15
6/15/2020 10:02	15/06/2020 10:19	L8 196/20-21	17.3
4/19/2022 10:15	19/04/2022 10:13	L9 195/21-22	1.81
10/19/2022 10:10	19/10/2022 10:20	L9 196/20-21	10.59
5/8/2023 9:59	8/5/2023 10:13	L9 195/21-22	14.77
6/8/2023 9:55	8/6/2023 10:19	L8 196/20-21	24.49
9/4/2023 10:13	4/9/2023 10:20	L9 196/20-21	6.46

To make the multi-timestamp models account for seasonality, the year-season was considered as a possible predictor. This corresponds to a pure temporal categorical variable whose classes correspond to four meteorological groups: spring (March – April – May), summer (June – July – August), autumn (September – October – November), and winter (December – January – February).

Pure spatial variables were considered as possible predictors for all models. These refer to time-invariant geospatial predictors that describe the fixed physical characteristics of the study domain (**Table 2**). They provide essential information on topographic, geographic and surface-related controls that systematically modulate temperature fields and land–atmosphere interactions.

Regarding topography, Digital Elevation Models (DEMs) provide high-resolution representations of terrain height that enhance temperature modelling by explicitly accounting for elevation-dependent variability, particularly in regions with complex terrain or in proximity to coastlines and large water bodies. As noted by Oke et al. [9], such geographic controls exert a strong influence on the spatial organisation, diurnal and seasonal evolution of urban and regional thermal patterns.

To further characterise terrain-related influences, a Topographic Exposure Index (TOPEX) [33] was included to quantify spatial variations in terrain exposure and sheltering. TOPEX was computed using a Python-based workflow to ensure computational efficiency and scalability. For each grid cell in the DEM [34], terrain elevation angles were calculated along radial transects at regular distance increments (100 m) up to a maximum radius of 2 km for each directional sector. The maximum horizon (occultation) angle within each sector was retained as the TOPEX value, with positive values indicating relative topographic sheltering and negative values indicating exposure. TOPEX was computed for the eight cardinal and intercardinal directions (N, NE, E, SE, S, SW, W, NW), and a single composite index was obtained by averaging values across all directional sectors.

Proximity to large water bodies influences LST through the thermal inertia of water and associated land–sea thermal contrasts, which modulate surface heating and cooling rates. To represent this effect, Euclidean distance to the coastline was computed for each grid cell using a high-resolution European coastline dataset and included as a fixed spatial predictor using QGIS, [35].

Beyond large-scale geographic controls, spatial variability in LST is also influenced by persistent surface characteristics associated with artificial surfaces, which affect radiative properties, surface moisture availability, and the partitioning of energy fluxes. Several urban-related geospatial variables were therefore included as fixed predictors. Imperviousness Degree (IMD) was used to represent the proportion of sealed surfaces within each grid cell. Impervious surfaces are associated with reduced evapotranspiration, enhanced sensible heat storage, and altered radiative behaviour, making IMD a key determinant of spatial LST variability.

Vegetation-related surface properties were also included to account for spatial differences in evaporative cooling potential. Tree Cover Density (TCD) [36] was used to represent the fraction of vegetated surfaces.

While individual surface variables describe specific physical properties, land surface temperature is often governed by the combined effect of multiple surface characteristics acting simultaneously. Local Climate Zones (LCZs) were developed to capture these combined effects by grouping areas that exhibit similar surface–atmosphere interaction behaviour under comparable

 doi:10.20944/preprints202604.0495.v1

atmospheric forcing. As such, LCZs provide a spatial framework for representing typical surface energy exchange regimes that emerge from the interaction of land cover, surface materials, and vegetation characteristics. In this study, LCZs were derived using a GIS-based tool provided by Oliveira et al. [37]. Rather than being introduced directly as categorical predictors, LCZ information was incorporated into the modelling framework through LCZ-specific Bowen-ratio values – named here "Urban Density" (UD) – which represent characteristic ratios between sensible and latent heat fluxes associated with different surface types [38].

To make the predictor variables suitable for modelling, data transformation was required. A consistent spatial resolution across all inputs was achieved by resampling each dataset to a common grid, using aggregation methods selected according to the characteristics of each variable. Specifically, predictor values were summarised within a regular 0.002 × 0.002° grid through zonal statistics. This procedure, implemented in QGIS (QGIS.org, 2025), calculates descriptive statistics for each grid cell—such as mean, median, sum, minimum, maximum, majority, or range. Unlike simple point-based sampling at cell centroids, zonal statistics reduce the likelihood of assigning non-representative extreme values, as they account for all pixels within each target cell. Consequently, the statistical measure applied to each predictor depended on its data type, with different approaches used for continuous and categorical variables, as summarised in **Table 2**.

Table 2. Final-predictors data sources and corresponding zonal statistics methods for resampling to the regular target-grid

Predictor	Source & References	Zonal Statistic Method
Digital Elevation Model (DEM)	[34]	Mean
Topographic Exposure Index (TOPEX)	[33,39]	
Distance to the Coast (DCOAST)	[35,40]	
Imperviousness Density (IMD)	[41]	
Tree Cover Density (TCD)	[36]	
Local Climate Zones in Bowen Ratio (UD)	[37]	Majority

2.1.2. Data Splitting and Training/Validation/Testing Strategy

Sentinel-3's 7 timestamps for which there were fine Landsat matched data were considered for testing (see **Table 1**). With them, it was decided to define three different kinds of tests: two for coarse and fine predictions using Landsat LST data reprojected to Sentinel-3's fine and coarse grids (these grids are associated with OLCI and SLSTR sensors, respectively), and one for coarse prediction using Sentinel-3's LST. The purpose of considering a coarse test using Landsat data, besides another using Sentinel's, was to assess how the two datasets differed in a common coarse grid and if such differences could influence the results in the fine test. Differences could indeed arise due to multiple factors such as lack of synchronicity in the acquisitions, different view angles, and LST computation algorithms.

The conception of the downscaling models would require hyperparameter tuning and subsequent training, besides testing. It was decided to perform the hyperparameter tuning by picking the remaining 105 timestamps and doing a 5-fold cross-validation scheme stratified by season. In practice, this means that each fold contained approximately the same proportion of data for each season as in the original data. In this way, the performance of the models is proportionally assessed with respect to seasonality. In such hyperparameter tuning, for each candidate set of hyperparameter

 doi:10.20944/preprints202604.0495.v1

values, the models were trained with 4 folds and validated with 1, further rotating the validating fold until all of them were used. The resulting cross-validation scores were then defined as the arithmetic mean of the respective 5 validations. And the set of hyperparameter values with the highest cross-validation score was the one taken. With tuning done, the models were retrained using the whole cross-validation data and subsequently tested on the test subset (i.e., the withheld Landsat-Sentinel pairs mentioned in **Table 1**).

The whole data architecture considered in this work is summarised in **Figure 2**.

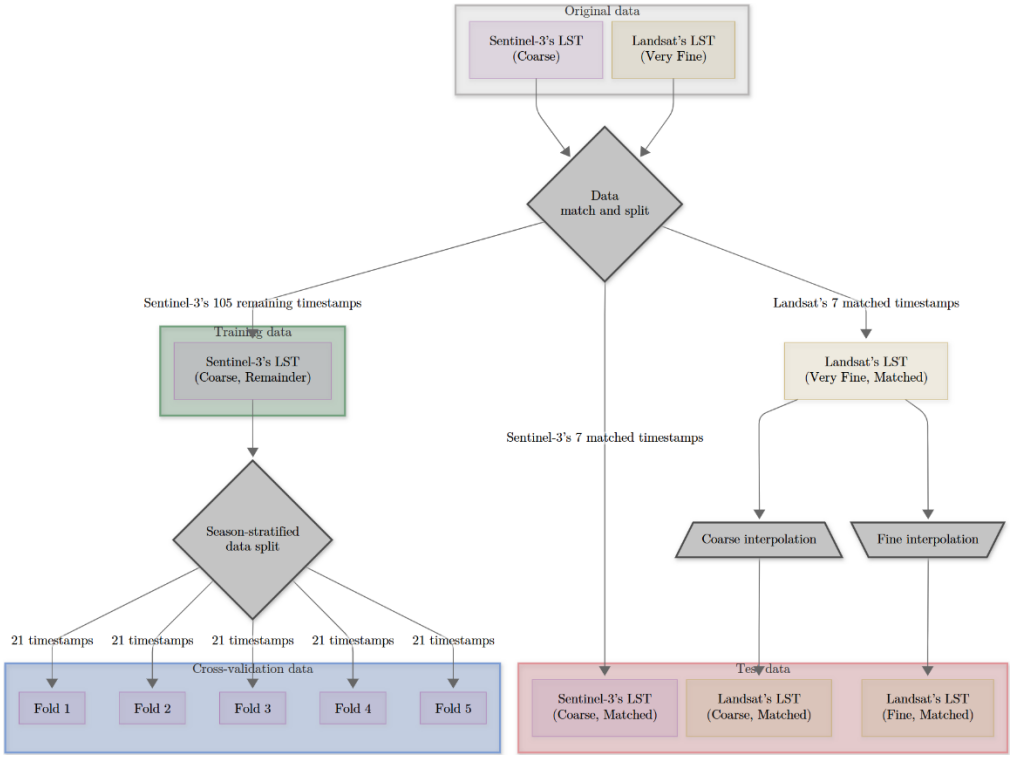


Figure 2. Data architecture considered in the present work.

2.2. Methods

2.2.1. Hypothesis of Scale Invariance

Scale invariance [42] corresponds to the conservation of relation between variables with respect to scale, or in practical terms, to grids of different resolution. Let f be the relation between predictors X and target LST. And let "coarse" and "fine" denote grids of coarse and fine resolution, respectively. Under the assumption of scale invariance, the relationship between land surface temperature and its predictors can be expressed as $\mathrm{LST}_{\mathrm{coarse}} = f(X_{\mathrm{coarse}})$ and $\mathrm{LST}_{\mathrm{fine}} = f(X_{\mathrm{fine}})$. This hypothesis was taken in the present work by training the base model f with predictors and target at Sentinel-3's coarse grid (X_{coarse} and $\mathrm{LST}_{\mathrm{coarse}}$) and inferring $\mathrm{LST}_{\mathrm{fine}}$ through f using predictors at Sentinel-3's fine grid (X_{fine}) as a first approximation.

It is important to exercise caution when assuming that a relation f is scale-invariant. It is well known that scale invariance does not hold in all cases [43,44]. For instance, if the coarse data are regarded as an area-weighted average of the fine data, scale invariance from a finer to a coarser grid would mean conservation of the relationship with respect to averaging. However, spatial averaging tends to shrink the distribution of values which may cause relationships observed at the coarse grid to break at fine-scale extremes.

2.2.2. Residual Correction

 doi:10.20944/preprints202604.0495.v1

Given the true Sentinel-3 coarse resolution values, $\mathrm{LST}_{\mathrm{coarse}}$, and the ones predicted by the trained model f, $\widehat{\mathrm{LST}}_{\mathrm{coarse}}$, the coarse-scale prediction residual ($\varepsilon_{\mathrm{coarse}}$) can be computed as

$$\varepsilon_{\mathrm{coarse}} = \mathrm{LST}_{\mathrm{coarse}} - \widehat{\mathrm{LST}}_{\mathrm{coarse}}\,, \tag{4}$$

and used to estimate the fine prediction residual ($\varepsilon_{\mathrm{fine}}$). In this work, the fine-resolution prediction residual was approximated by the bilinear interpolation of the coarse prediction residual onto the fine grid, that is,

$$\varepsilon_{\mathrm{fine}} = \mathrm{LST}_{\mathrm{fine}} - \widehat{\mathrm{LST}}_{\mathrm{fine}} \approx \mathrm{interp}_{\mathrm{fine}}(\varepsilon_{\mathrm{coarse}}) =: \hat{\varepsilon}_{\mathrm{fine}}\,. \tag{5}$$

Such an approximation would be exact, for example, if the true $\mathrm{LST}_{\mathrm{fine}}$ was a bilinear interpolation of the true $\mathrm{LST}_{\mathrm{coarse}}$ and if the predicted $\widehat{\mathrm{LST}}_{\mathrm{fine}}$ was obtained by a bilinear interpolation of $\widehat{\mathrm{LST}}_{\mathrm{coarse}}$ instead of the application of the model f on the fine predictors. Given the estimated fine-scale prediction residual, it is then possible to correct the $\widehat{\mathrm{LST}}_{\mathrm{fine}}$ values predicted by f through

$$\widehat{\mathrm{LST}}_{\mathrm{fine,corr}} = \widehat{\mathrm{LST}}_{\mathrm{fine}} + \hat{\varepsilon}_{\mathrm{fine}}\,. \tag{6}$$

This procedure is called “residual correction” [45].

2.2.3. Architecture of the Single-Timestamp Model

Figure 3 presents the architecture of the single-timestamp model considered in this work, which is based on the scale-invariance principle and the application of residual correction. In this architecture, and as mentioned before, the base model f is trained with coarse predictors (X_{coarse}) and coarse LST ($\mathrm{LST}_{\mathrm{coarse}}$). It predicts a coarse LST ($\widehat{\mathrm{LST}}_{\mathrm{coarse}}$) or fine one ($\widehat{\mathrm{LST}}_{\mathrm{fine}}$) whether the issued predictors are coarse (X_{coarse}) or fine (X_{fine}). The coarse predictors correspond to spatio-temporal and pure spatial predictors bilinearly interpolated onto Sentinel-3’s coarse grid ($X_{xt,\mathrm{coarse}}$ and $X_{x,\mathrm{coarse}}$, respectively).

Similarly, the fine predictors correspond to a combination of spatio-temporal predictors ($X_{xt,\mathrm{fine}}$) and pure spatial predictors bilinearly interpolated onto Sentinel-3’s fine grid ($X_{x,\mathrm{fine}}$).

The predicted fine LST ($\widehat{\mathrm{LST}}_{\mathrm{fine}}$) is then corrected using the estimation (5) for the fine residual ($\hat{\varepsilon}_{\mathrm{fine}}$). This architecture has been shown to be quite effective in the prediction of fine LST using a base model trained with coarse data from the very same timestamp [30,46,47] – hence the designation “single-timestamp model”.

 doi:10.20944/preprints202604.0495.v1

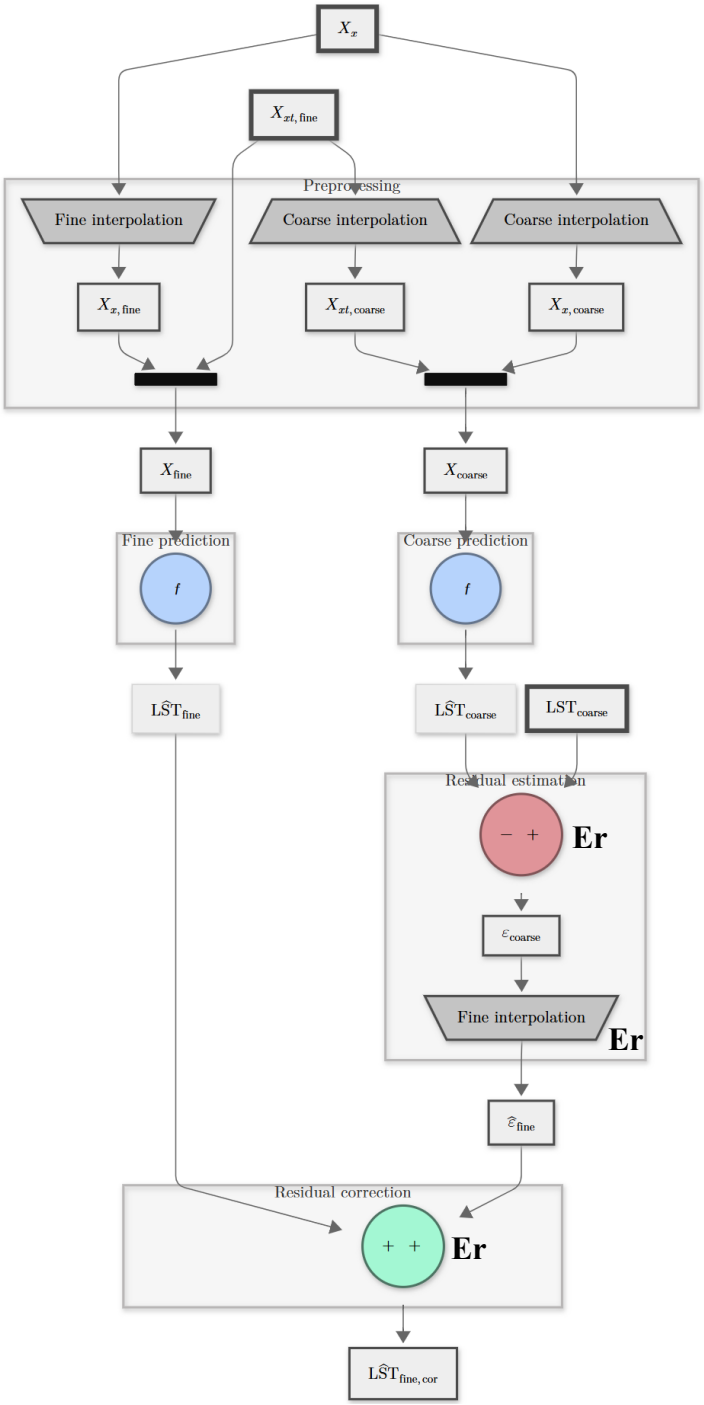


Figure 3. Architecture of the single-timestamp downscaling model, with reference to the equations of the involved residual handling processes.

2.2.4. Architecture of the Multi-Timestamp Model

There are several downsides of using a pure timestamp-specific model when compared to a general (multi-timestamp) one. The training data is quite limited – it solely concerns the coarse data of the timestamp whose fine target is to be inferred. The coarse data of the timestamp does not necessarily fully describe the respective fine one. For instance, when one obtains coarse data through area-weighted averaging of the fine data, the tails of the value distribution diminish, and some of the information is inevitably lost. Information may be added by considering coarse data from multiple other timestamps. The central question is whether the additional information is consistent with the fine-resolution data of the target timestamp. It should also be noted that considering ML timestamp-specific models beyond LR would be highly impractical. Not only ML models usually

 doi:10.20944/preprints202604.0495.v1

require larger amounts of data in their training (so that they become general enough, avoiding overfitting), but their hyperparameters must also be tuned. Performing such tuning for each timestamp would be infeasible: the already few coarse data of a sole timestamp would need to be batched into different folds for cross-validation or early stopping, the results could be too biased, and the task would need to be repeated for every single timestamp. A more convenient approach is to conceive a multi-timestamp ML model, tune and train it with coarse data from several timestamps, enabling inference for new ones without the need for further hyperparameter re-tuning and re-training. Note that the authors do acknowledge that single-timestamp architectures using ML base models have been widely reported in the literature, but a higher interest in operationality would inevitably imply discarding such option.

As it will be shown later in the Results section, timestamp-specific standardisation of the target in a multi-timestamp architecture is highly recommended since, in contrast with the raw default case, it better conserves the predictor-target correlation in each timestamp. The multi-timestamp model with timestamp-specific standardisation can be further extended by incorporating spatio-temporal (X_{xt}), purely spatial (X_x) and purely temporal (X_t) predictors - as well as by replacing LR with an ML model. The flowchart of **Figure 4** describes this general architecture. In this architecture, the base model f is trained with coarse predictors (X_{coarse}) and standardised coarse LST ($\delta\text{LST}_{\text{coarse}}$). Therefore, it predicts a standardised coarse LST ($\delta\widehat{\text{LST}}_{\text{coarse}}$) or fine one ($\delta\widehat{\text{LST}}_{\text{fine}}$) whether the issued predictors are coarse (X_{coarse}) or fine (X_{fine}). The coarse predictors correspond to pure temporal predictors (X_t), standardised coarse spatio-temporal predictors ($\delta X_{xt,\text{coarse}}$) and coarse pure spatial predictors ($X_{x,\text{coarse}}$). The standardised coarse spatio-temporal variables are obtained from the raw ones through

$$\delta\text{LST}_{\text{coarse}} = \frac{\text{LST}_{\text{coarse}} - \overline{\text{LST}}_{\text{coarse}}}{s_{\text{LST}_{\text{coarse}}}}, \tag{7}$$

$$\delta X_{xt,\text{coarse}} = \frac{X_{xt,\text{coarse}} - \bar{X}_{xt,\text{coarse}}}{s_{X_{xt,\text{coarse}}}}, \tag{8}$$

where $s_{\text{LST}_{\text{coarse}}}$, $s_{X_{xt,\text{coarse}}}$, $\overline{\text{LST}}_{\text{coarse}}$ and $\bar{X}_{xt,\text{coarse}}$ correspond to the sample standard deviations and arithmetic means of the coarse LST and spatio-temporal predictors associated with the given timestamp.

Similarly, the fine predictors correspond to a combination of pure temporal variables (X_t), standardised fine spatio-temporal predictors ($\delta X_{xt,\text{fine}}$) and fine pure spatial predictors ($X_{x,\text{fine}}$). Note that since the base model f is trained with standardised coarse spatio-temporal predictors, the fine ones are expected to be also standardised using the coarse statistics, that is,

$$\delta X_{xt,\text{fine}} = \frac{X_{xt,\text{fine}} - \bar{X}_{xt,\text{coarse}}}{s_{X_{xt,\text{coarse}}}}. \tag{9}$$

The predicted standardised fine LST ($\delta\widehat{\text{LST}}_{\text{fine}}$) is corrected using an estimation for the fine residual ($\hat{\varepsilon}_{\text{fine}}$). As previously mentioned, this estimation is simply the bilinear interpolation of the coarse residual ($\varepsilon_{\text{coarse}}$) onto the fine grid, and the coarse residual is in turn the difference between standardised coarse LST ($\delta\text{LST}_{\text{coarse}}$) and the predicted standardised coarse LST ($\delta\widehat{\text{LST}}_{\text{coarse}}$). The corrected predicted standardised fine LST ($\delta\widehat{\text{LST}}_{\text{fine,corr}}$) must then be de-standardised. Since the base model was trained with coarse data, it predicts values that are standardised using the coarse statistics. Therefore, the process of de-standardisation for obtaining the actual corrected fine LST also involves the coarse statistics:

$$\widehat{\text{LST}}_{\text{fine,corr}} = s_{\text{LST}_{\text{coarse}}} \cdot \delta\widehat{\text{LST}}_{\text{fine,corr}} + \overline{\text{LST}}_{\text{coarse}}. \tag{10}$$

In contrast with the single-timestamp architecture, the base model f of the multi-timestamp architecture is trained with combined coarse data from multiple timestamps. It predicts a standardised LST instead of a raw one. And the regarded spatio-temporal predictors are also

 doi:10.20944/preprints202604.0495.v1

standardised. In both architectures, f is pixel-agnostic, or, in other words, value-specific, that is, it is a function that does not depend on space but purely on the values of the predictors in that space. There is then one and only one function f for all pixels. This means that the positioning of the variable values in the pixel matrix is irrelevant to f. The predictor and target values in a pixel constitute a single observation, and the training dataset associated with a single timestamp becomes a collection of such observations, regardless of their position. And the training dataset associated with multiple timestamps naturally corresponds to a concatenation of the data collections of the multiple timestamps.

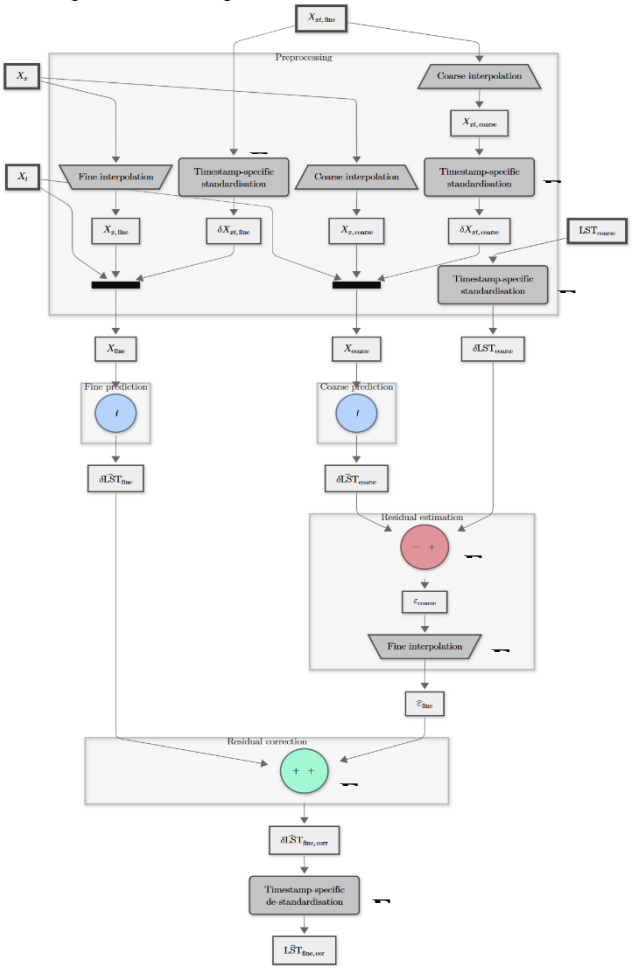


Figure 4. Architecture of the multi-timestamp downscaling model, considering timestamp-specific standardisation of the spatio-temporal variables (X_{xt}, LST). Equations associated with standardisation, de-standardisation and residual handling processes are also herein referred to.

2.2.5. Candidate Base Models

Besides LR, three different ML models were considered as possible candidates for the base model f of the multi-timestamp downscaling architecture: Feed-Forward Neural Network (MLPRegressor from scikit-learn [48]), Random Forest and Gradient Tree Boosting (XGBRFRegressor and XGBRegressor, respectively, from XGBoost [49]). For further benchmarking, Dummy Mean Regression (DMR) model will also be considered as base model of the multi-timestamp architecture. Note that a DMR base model always predicts the arithmetic mean of the target values seen in training, and one may show that, when considering residual correction, the resulting downscaling model is equivalent to pure bilinear interpolation of the true coarse Sentinel-3's LST onto Sentinel-3's fine grid. In fact, this is true for any residual-downscaling model with a base model f that predicts a constant c (see proof (A4) provided in Appendix A.2).

3. Results

 doi:10.20944/preprints202604.0495.v1

3.1. Exploratory Data Analysis

The left-hand side subfigure of **Figure 5** (a), presents the arithmetic mean of the Pearson correlation matrix of the numerical coarse data for each test timestamp. It reveals that Sentinel-3 and Landsat LST data do differ, having a correlation coefficient corresponding to 0.84. One should be aware that this could possibly introduce some bias when comparing the downscaling results of the different models with the fine Landsat data. Furthermore, if one assumed that the relation between coarse Sentinel-3 and Landsat data coincided with the one between downscaled Sentinel-3 and fine Landsat data, a coefficient of determination of just $R^2 \sim 0.71$ would be (on average) expected.

Subfigure (a) of **Figure 5** also shows that the predictors that have the highest (in absolute value) correlation coefficient with respect to Sentinel-3's LST corresponded to the spatio-temporal ones: FVC (-0.48), immediately followed by NDVI (-0.47) – which is almost collinear with FVC – and NDWI (0.43) – which is also highly correlated with the other two. Predictors with a smaller but moderate correlation coefficient corresponded to TCD (-0.32), IMD (0.31), UD (0.29) – which is highly correlated with IMD – and TOPEX (-0.26) – which is significantly correlated with TCD. Other not so correlated predictors corresponded to DEM (-0.13) and d_{coast} (-0.09) – which is significantly correlated with DEM.

Figure 5 further presents in its right-hand side figure, (c), the Pearson correlation matrix that is obtained when considering the coarse data combined. By comparing it against the arithmetic mean of Pearson correlation matrices for each timestamp, one finds that all correlation coefficients between predictors and Sentinel-3's LST significantly decrease (in absolute value) when considering data combination. This effect is particularly pronounced for the predictors that originally showed the strongest correlation values (i.e. FVC, NDVI and NDWI) – the values not only diminish but also exhibit sign reversal. This may be explained by the fact of FVC, NDVI and NDWI not corresponding to actual physical quantities but normalised indexes (FVC is within the range $[0, 1]$, NDVI and NDWI are within $[-1, 1]$). Although within a given same timestamp it is uncommon to have largely different LST values associated with the same FVC, NDVI or NDWI values, this relationship does not persist across different timestamps, and linearity is, therefore, reduced.

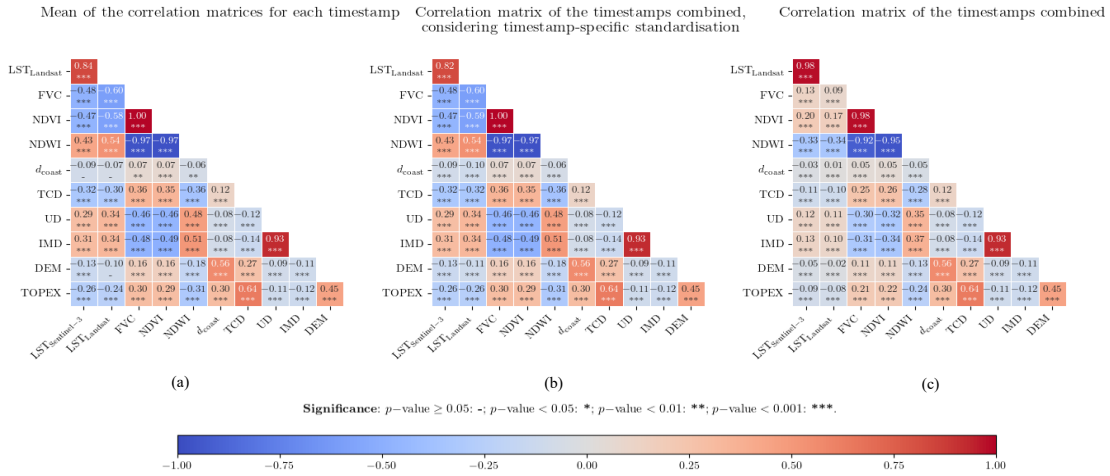


Figure 5. Pearson correlation matrices (and respective statistical significances) of the numerical coarse data for the test timestamps: as a mean of the correlation matrices for each timestamp (left-hand side, (a)), as the correlation matrix of the combined data considering timestamp-specific standardisation of the spatio-temporal variables (centre, (b)) and not considering (right-hand side, (c)). Coarse Landsat data had been obtained from original one by reprojecting the latter onto Sentinel-3's coarse grid.

Figure 6 shows how significantly different Sentinel-3's LST values from different timestamps can correspond to the same FVC value. It also shows that for each timestamp alone, LST varies with FVC in a quasi-linear relationship. Therefore, in the absence of data transformations, a timestamp-specific LR model would be expected to perform more accurately than a multi-timestamp model. The line on the left-hand side subfigure of **Figure 6**, (a), corresponds to the predicted LST using a multi-

 doi:10.20944/preprints202604.0495.v1

timestamp LR model with FVC as a predictor. This line does not agree with the actual data and can only roughly estimate the overall average, producing a R^2 value of just 0.02. The subfigure further shows that the data mostly differs in offset between each timestamp. This suggests that a multi-timestamp model based on centred variables (raw variables with their timestamp-specific means subtracted) may perform better than one on the raw variables. When using (A1) (see **Appendix A.2**) as model and training it with the combined coarse data of the test timestamps, the result presented by the central subfigure of **Figure 6**, (b), is obtained. The R^2 score abruptly increases from 0.02 to 0.89, RMSE decreases from 5.89 to 1.93 K and the predicted LST lines much better agree with the actual data. However, note how all these lines not only have their own distinct offset but a common slope whereas the actual data shows some slope variance. A reasonable approximation for the slope of the actual data for timestamp t could be defined as $s_{\mathrm{LST}_t}/s_{\mathrm{FVC}_t}$, where s_{LST_t} and s_{FVC_t} are the sample standard deviations of the LST and FVC values for timestamp t. The exact slopes may be approximately achieved by considering timestamp-specific standardisation of the variables (division of the centred variables by the timestamp-specific sample standard deviations of the respective raw ones) instead of centring.

With the timestamp-specific standardisation LR model (A2) (see **Appendix A.2**) trained with the coarse data of all test timestamps, the right-hand side subplot in **Figure 6** (c), was obtained. The R^2 score and RMSE improved just slightly, from 0.89 to 0.90, and from 1.93 to 1.91 K, respectively. The subplot shows that although some of the predicted LST lines better agree with the actual data, others do not. The similarity between the single and multi-timestamp LR models when considering timestamp-specific standardisation of the spatio-temporal variables while using a single predictor is further highlighted by the close agreement between the resulting Pearson correlation matrices, shown in the left-hand and centre subplots of **Figure 5**, (a) and (b), respectively. The curious reader may access **Appendix A.2** to understand how such proximity may be mathematically justified.

The current results justify the usage of timestamp-specific standardisation in a multi-timestamp downscaling mode.

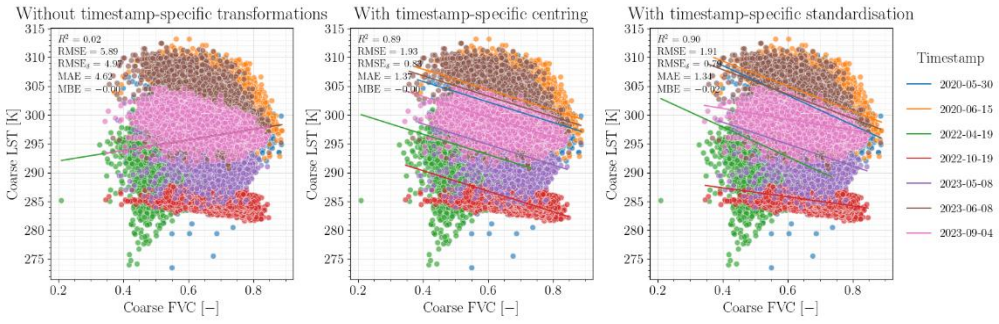


Figure 6. Predicted (lines) and actual (markers) coarse Sentinel-3's LST without timestamp-specific transformation (left-hand side, (a)), with centring (centre, (b)) and with standardisation (right-hand side, (c)) for test timestamps. Prediction is done using a multi-timestamp LR model with FVC as predictor.

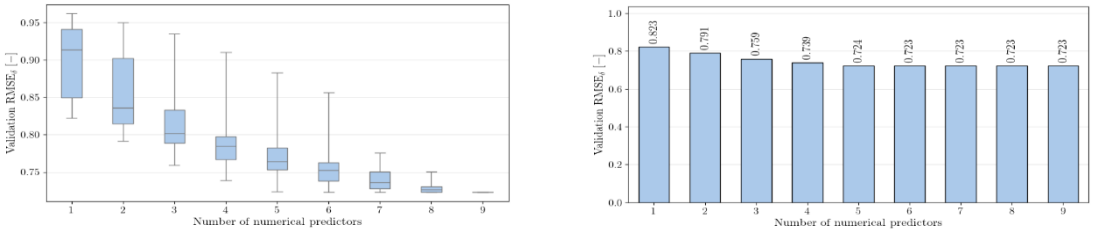
3.1. Hyperparameter Tuning

3.1.1 Selection of Numerical Predictors for a Multi-Timestamp Linear Regression Model

The numerical predictors for a multi-timestamp LR model may be selected as the combination that yields the highest cross-validation score. There are 9 possible numerical predictors: FVC, NDVI, NDWI, IMD, TCD, DEM, TOPEX, UD and d_{coast}. This results in $2^9 - 1 = 511$ possible combinations. **Figure 7** shows the obtained cross-validation RMSE values for the predicted standardised coarse target (RMSE_δ) for each number of numerical predictors – as a distribution (at the left-hand side, (a)) and as the best value (at the right-hand side, (b)) for each number. As expected, and in overall, the RMSE_δ values tend to decrease with the number of numerical predictors. When

 doi:10.20944/preprints202604.0495.v1

examining the best-performing models for each predictor count, one finds the $RMSE_\delta$ values to swiftly decrease with number of predictors but then to stagnate from 6 to 9.



(a) (b)

Figure 7. Cross-validation $RMSE_\delta$ (RMSE associated with the predicted standardised coarse target) of the LR models for each number of numerical predictors as a distribution (a) and as the best value (b), obtained with a multi-timestamp LR model considering timestamp-specific standardisation of the spatio-temporal variables.

Table 3 presents the best combinations of predictors for each of their numbers sorted from worst to best. By defining the best compromising overall combination as the set with the smallest number of predictors that achieves the lowest $RMSE_\delta$ value to the second decimal place, the combination FVC, d_{coast}, IMD, NDWI and TCD (5 predictors) is obtained. The other predictors – DEM, TOPEX, UD and NDVI – were ultimately found to be of lesser importance. This could be explained by the strong correlation of these other predictors with the ones already included in the set of 5. Indeed, and as shown in the centre subfigure of **Figure 5**, (b), DEM is highly correlated with d_{coast}, UD with IMD, TOPEX with TCD, and NDVI with FVC and NDWI.

Table 3. Cross-validation $RMSE_\delta$ (RMSE associated with the predicted standardised coarse target) of best combinations of numerical predictors for each of their number, from worst to best.

Numerical predictors	Number of numerical predictors	$RMSE_\delta$ [–]
FVC	1	0.823
FVC, d_{coast}	2	0.791
FVC, IMD, NDWI	3	0.759
FVC, d_{coast}, IMD, NDWI	4	0.739
FVC, d_{coast}, IMD, NDWI, TCD	5	0.724
FVC, d_{coast}, IMD, NDWI, TCD, DEM	6	0.723
FVC, d_{coast}, IMD, NDWI, TCD, DEM, TOPEX, UD, NDVI	9	0.723
FVC, d_{coast}, IMD, NDWI, TCD, DEM, TOPEX, UD	8	0.723
FVC, d_{coast}, IMD, NDWI, TCD, DEM, TOPEX	7	0.723

3.1.2. Selection of Categorical Predictors for a Multi-Timestamp Linear Regression Model

To incorporate seasonal effects in the multi-timestamp LR model, season was considered as a potential predictor in addition to the most relevant numerical variables identified in the previous section. Season is a categorical variable whose inclusion in a linear model may be done in different ways. Four linear model configurations were tested: (i) a model with no seasonal dependence; (ii) a model with season-specific intercepts; (iii) a model with season-specific intercepts and slopes applied exclusively to the spatio-temporal predictors (FVC and NDWI); and (iv) a model with season-specific intercepts and slopes applied to all numerical predictors. All these models may be conveniently expressed using Wilkinson formulas [50].

In similarity to what was done in the tuning of numerical predictors, the four different formulas were cross-validated, and the best one was defined as the simplest formula yielding the smallest $RMSE_\delta$ value to the second decimal place. **Figure 8** shows how the increasing complexity of the

 doi:10.20944/preprints202604.0495.v1

formulas did result in lower RMSE_δ values. Still, by rounding RMSE_δ to two decimal places it was possible to confirm that solely the most complicated formula (iv), that is,

$$\mathrm{LST} \sim (\mathrm{TCD} + d_{\mathrm{coast}} + \mathrm{IMD} + \mathrm{FVC} + \mathrm{NDWI}) * \mathrm{Season}\,, \tag{11}$$

could produce a non-negligible reduction of error (of 0.014). This formula was, therefore, the one taken.

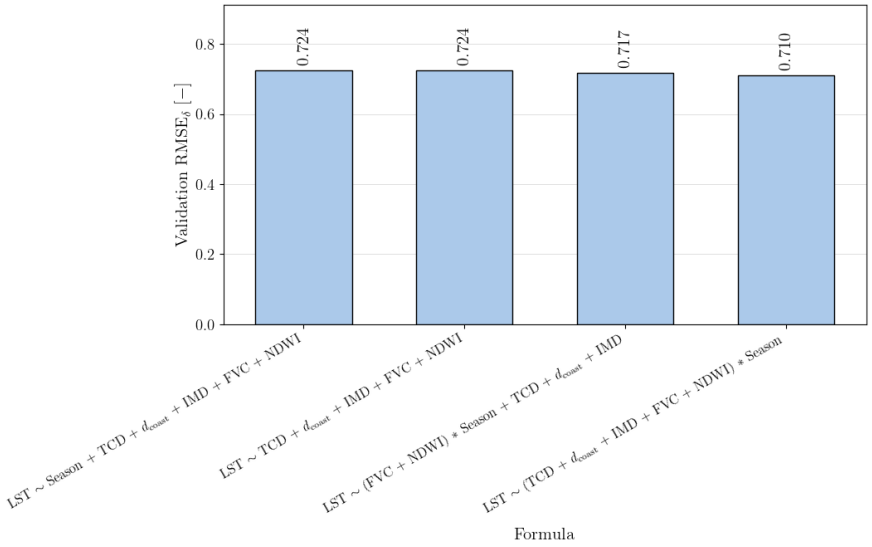


Figure 8. Cross-validation RMSE_δ (RMSE associated with the predicted standardised coarse target) for each season-dependent Wilkinson formula obtained with a multi-timestamp LR model.

- **Hyperparameter Tuning of Multi-Timestamp ML Models**

As mentioned before, three different ML models were considered as possible candidates for the base model f of the multi-timestamp downscaling architecture: NN, RF and GB. In the case of these models, all predictors except NDVI (which is almost collinear with FVC) were considered. The hyperparameters were tuned using Optuna [51] based on the cross-validated RMSE_δ score.

In the setting of the NN architecture for tuning, it the following specifications were considered: dummy encoding of the categorical predictors (Season), Rectified Linear Unit (ReLU) activation functions and training with Adaptive Moment Estimation (ADAM) Gradient Descent using mini-batches of size 1024, with a maximum number of 1000 epochs. One had also considered Early Stopping with 20 % of data for validation using R^2 as validation score, a patience of 10 epochs and tolerance of 0.001 in the validation score. **Table 4** shows the obtained tuned values for some of the notable hyperparameters of the three ML models.

Table 4. Tuned hyperparameter values for the Neural Net, Random Forest, and Gradient Boosting base models within multi-timestamp architectures.

Model	Hyperparameter	Tuned value
Neural Network (NN)	Hidden layers	Two hidden layers (31 and 25 units)
	Initial learning rate	1.3×10^{-3}
	L2 regularisation term (α)	1.8×10^{-5}
Random Forest (RF)	Categorical encoding	Dummy encoding (Season)
	Number of trees	758
	Maximum tree depth	14
	Minimum loss reduction for split (γ)	0.15
	Fraction of data records for each split ("subsample")	0.64

 doi:10.20944/preprints202604.0495.v1

	Fraction of features for each split ("colsample_bytree")	1.00
Gradient Boosting (GB)	Categorical encoding	Dummy encoding (Season)
	Number of trees	848
	Maximum tree depth	14
	Minimum loss reduction for split (γ)	0.51
	Fraction of data records for each split ("subsample")	0.77
	Fraction of features for each split ("colsample_bytree")	0.74
	Learning rate	0.041
	L1 regularisation term (α)	1.3×10^{-3}
	L2 regularisation term (λ)	2.90

3.2. Training, Cross-Validation and Test Overall Scores

Figure 9 presents the RMSE values of all tuned ML downscaling models obtained in training and testing as well as the values of (i) the multi-timestamp LR model described in the previous section, (ii) the single-timestamp LR model using the same numerical predictors as this multi-timestamp LR model, and (iii) the benchmarking multi-timestamp DMR model. The figure shows that in contrast with the multi-timestamp LR model, all ML models performed as good or better than the single-timestamp LR model in the case of coarse prediction when considering Sentinel-3's coarse LST as a true target. GB was found to be the best coarse predictor, however, with significant overfitting – the figure shows a training RMSE value for GB (1.33 K) that is significantly smaller than the coarse test one (1.56 K). In contrast, the single-timestamp LR model produced a coarse test RMSE of 1.62 K.

When considering coarsely interpolated Landsat's LST as true target, all RMSE values become smaller than the ones obtained with Sentinel-3's LST as true target, however, now with the single-timestamp LR model having the smallest error of them all. Still, the performance difference between the single-timestamp LR model and GB becomes quite marginal. These results show that Landsat's coarsened LST agrees better with the training Sentinel-3' coarse LST data than the test Sentinel-3 one. However, it differs significantly enough to change the ranking of the models in coarse prediction. Again, these differences make one to conclude that the Sentinel-3 and Landsat datasets do not completely match in a common coarse grid, and that this may influence the performance of the models in the fine test. When comparing fine prediction not considering residual correction with coarse prediction having Landsat's coarsened LST as true target, one finds all RMSE values to significantly increase, which may lead one to conclude that the hypothesis of scale-invariance does not hold well. However, the increase of RMSE could be associated with a greater dispersion of the fine LST values. A fairer comparison would be with respect to $RMSE_{\delta}$ values, that is, to the RMSE of the standardised predicted LST using the true LST statistics in the standardisation. **Figure 10** presents such $RMSE_{\delta}$ values, which also significantly increase from coarse to fine prediction not considering residual correction. Therefore, it can be unequivocally concluded that the hypothesis of scale-invariance leads to a large error. Furthermore, one may note that the tree-based models (RF and GB) are the worst fine predictors. This could be justified by the fact that the models have been trained with coarse data which do not contain the distribution tails of the fine one, underestimating extreme values in fine prediction. And since tree-based models are equivalent to piecewise functions, they perform sub-optimally when extrapolating.

 doi:10.20944/preprints202604.0495.v1

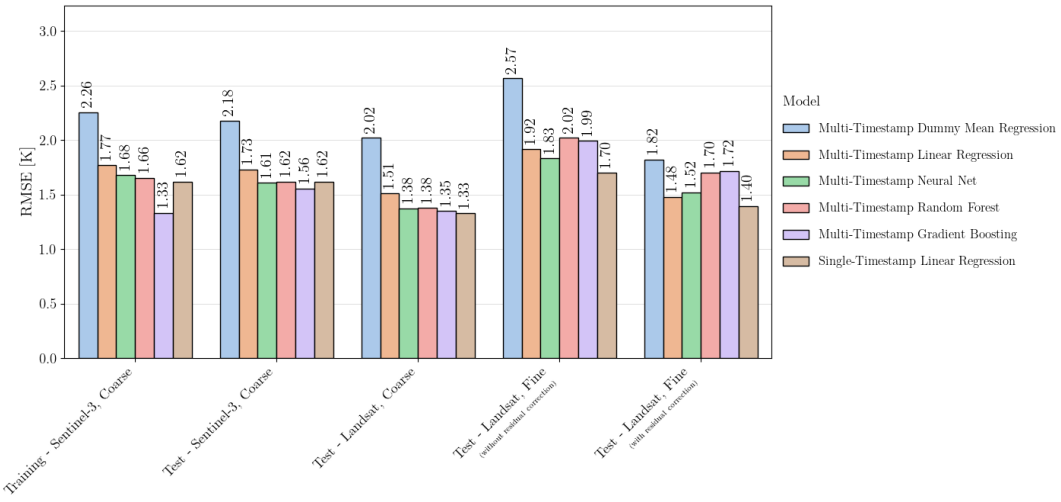


Figure 9. RMSE of all tuned downscaling models obtained in training and testing. Note that Landsat's coarse and fine data corresponds to the original one reprojected to Sentinel-3's coarse and fine grid, respectively.

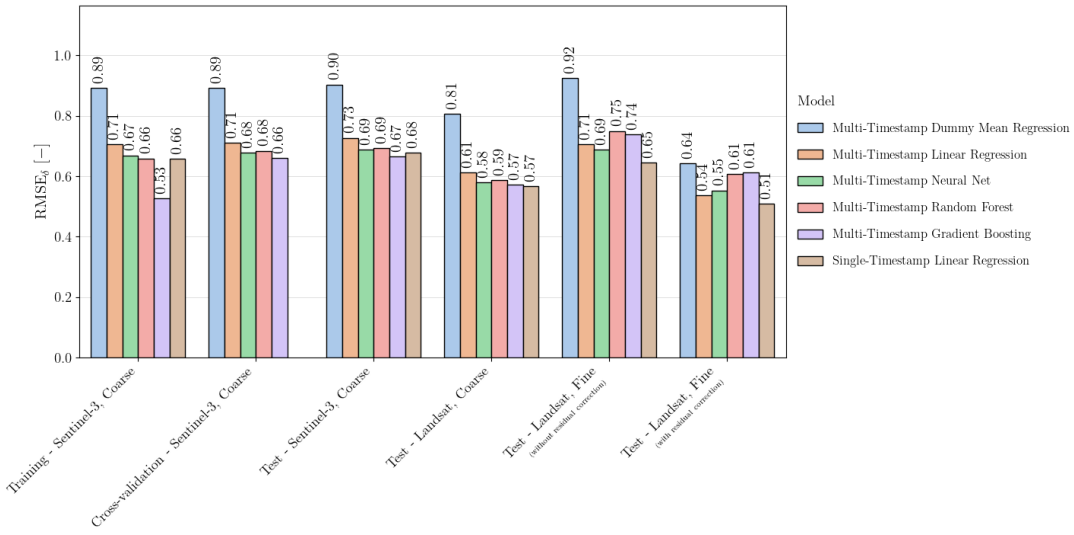


Figure 10. RMSE$_\delta$ (RMSE associated with the standardised predicted target using the statistics of the respective true target in the standardisation) of all tuned downscaling models obtained in training, cross-validation and testing. Note that Landsat's coarse and fine data corresponds to the original one reprojected to Sentinel-3's coarse and fine grid, respectively. Also note that there is no cross-validation RMSE$_\delta$ value for the single-timestamp LR model since this model can only infer for the same timestamp it is trained with, and cross-validation considers different timestamps for training and inference.

Breakage of scale invariance may also reveal itself through the values of the mean bias error (MBE) of the tuned downscaling models. **Figure 11** shows that the ML models, especially the tree-based ones, produce significantly negative MBE values in fine inference, revealing tendency for underprediction. This clearly evidences that the coarse relation learnt by the most complex downscaling models is actually different from the true fine one, or, that such coarse relation lacks support for the extreme values of the fine target.

 doi:10.20944/preprints202604.0495.v1

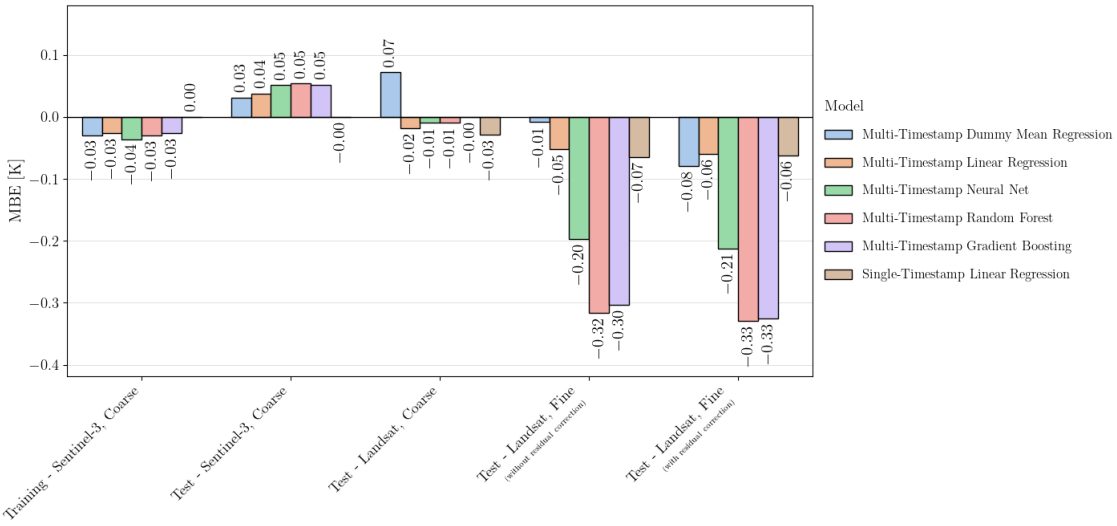


Figure 11. MBE (mean bias error) of all tuned downscaling models obtained in training and testing. Note that Landsat's coarse and fine data corresponds to the original one reprojected to Sentinel-3's coarse and fine grid, respectively.

By returning to **Figure 9** one additionally finds that residual correction in fine prediction make all RMSE values to abruptly decrease. In this case, the RF and GB models still do not perform well (with RMSE values of 1.70 and 1.72 K, respectively) – not too far from the result of pure interpolation (DMR produced a RMSE value of 1.82 K) – NN now performs slightly worse than the multi-timestamp LR model (with a RMSE value of 1.52 against 1.48 K) and the timestamp-specific LR model is still the best of them all (with a RMSE value of 1.40 K). Such results make the authors to ascertain that the timestamp-specific LR model is the downscaling model of choice when considering a scale-invariance-based architecture with residual correction.

3.3. Resultant Target Distributions and Respective Errors

To further prove how scale invariance does not hold particularly well in the current problem, the probability density functions (PDF) of the actual and predicted fine LST without residual correction for all test timestamps were plotted in **Figure 12**.

Overall, one finds the PDF of the predicted values to be much thinner than the PDF of the actual values for all models, which shows that these cannot predict the tails of the true distribution. The figure also shows that, from all models, the single-timestamp model is the one that better tends to encompass such tails. This evidence indicates that training with data from multiple timestamps instead of from solely one may actually deteriorate performance of the downscaling models, as common generalities (averages) between the different timestamps seem be more emphasised than their particularities (extremes). Also note that in the case of the DMR model, the PDF corresponds to a Dirac delta function centred on the mean of Sentinel-3's coarse LST for each timestamp.

 doi:10.20944/preprints202604.0495.v1

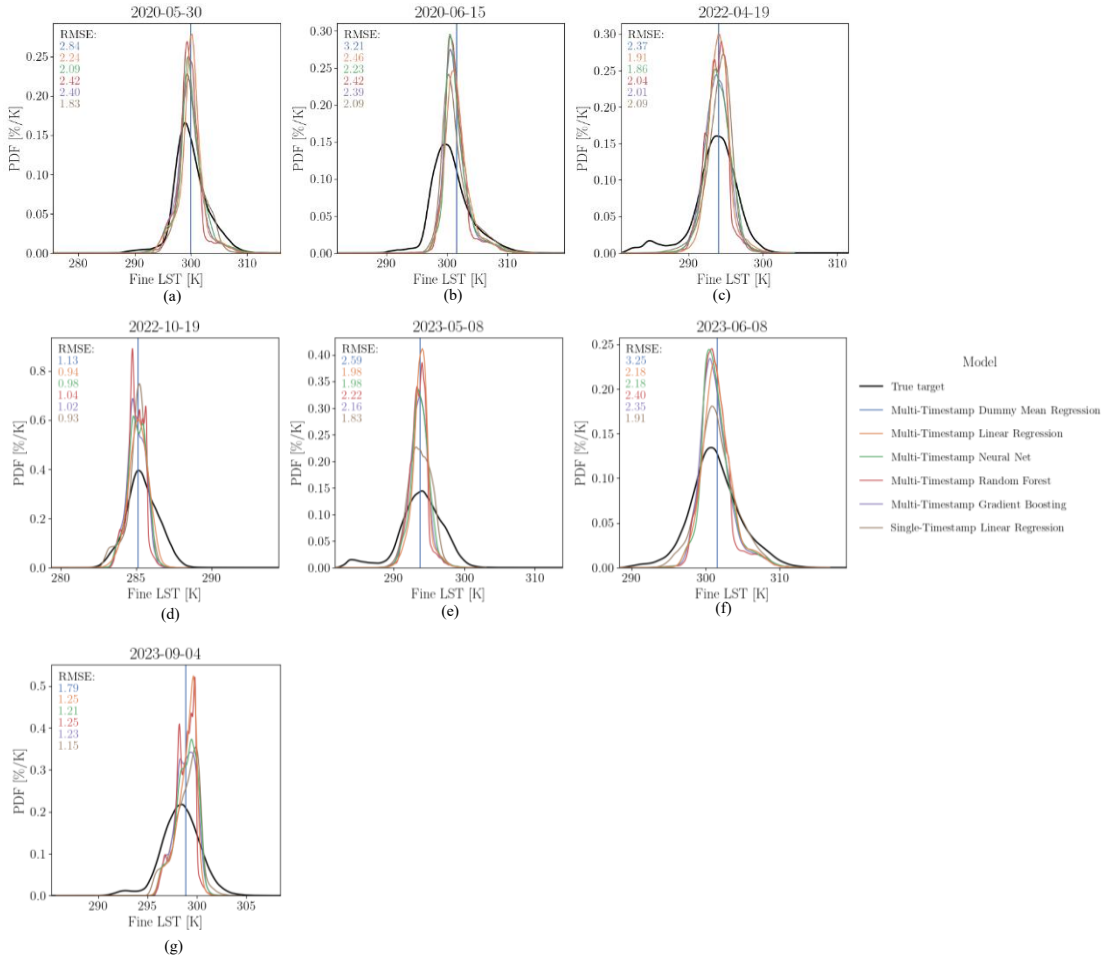


Figure 12. Probability density function (PDF) of the actual and predicted fine LST without residual correction for all test timestamps: 2020-05-30 (a), 2020-06-15 (b), 2022-04-19 (c), 2022-10-19 (d), 2023-05-08 (e), 2023-06-08 (f) and 2023-09-04 (g). Note how for the case of the DMR model, the PDF corresponds to a Dirac delta function centred on the mean of Sentinel-3's coarse LST for each timestamp.

Fortunately, residual correction quite effectively compensates for the error involved in the scale invariance assumption, as it makes the predicted and actual PDFs reasonably agree with each other – as shown by **Figure 13**. The DMR PDF remains distinctively thinner than the ones for any other model revealing that pure fine-interpolation more difficultly estimates extreme values. The figure also presents the RMSE values of the fine predictions (with residual correction) obtained by each model for each test timestamp, revealing that the single-timestamp LR model produces the smallest values for all timestamps except 2022-04-19. For this timestamp, multi-timestamp LR and NN do perform better. This shows that a single-timestamp model does not necessarily perform better than a multi-timestamp one for every timestamp, and that this could be explained by the possibility of the coarse data of a single timestamp differing too much from the respective fine one, even more than the coarse data from the combination of many other timestamps.

 doi:10.20944/preprints202604.0495.v1

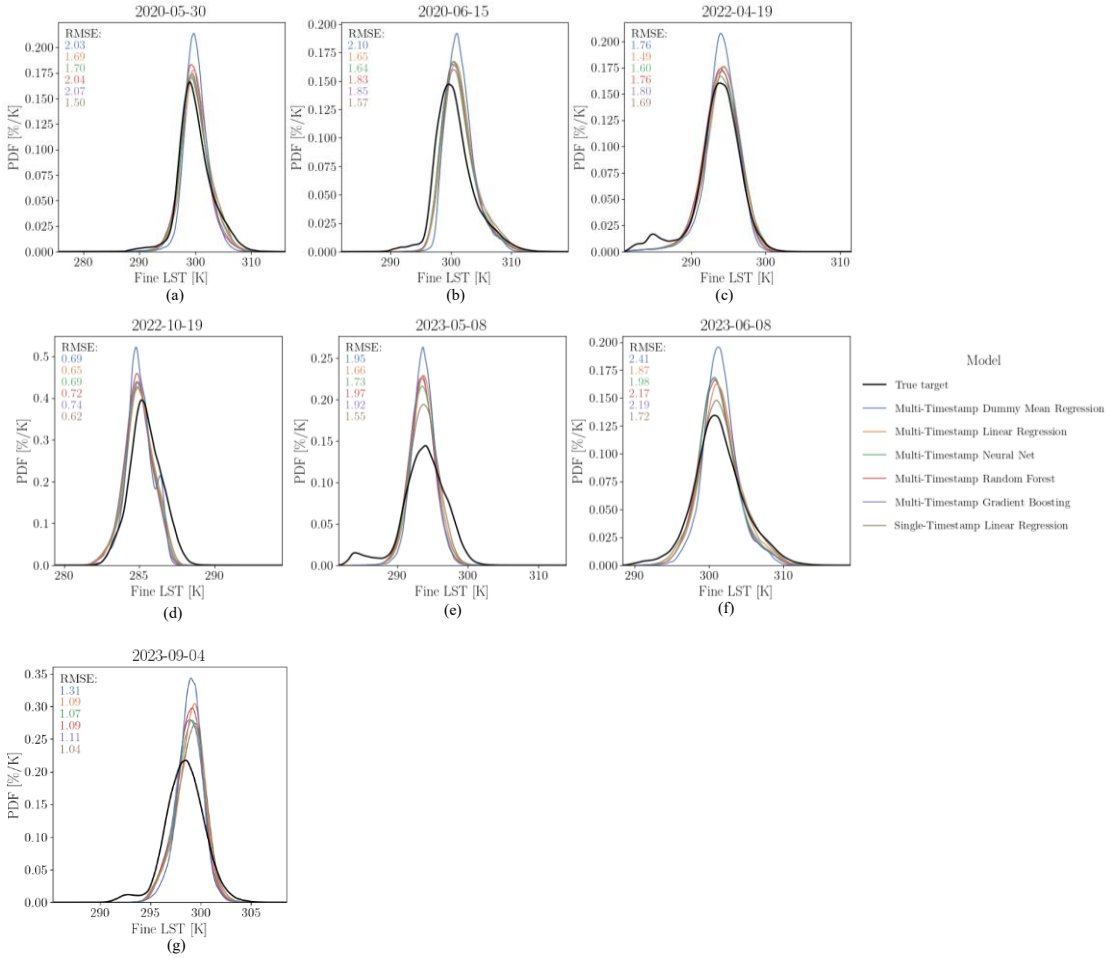


Figure 13. Probability density function (PDF) of the actual and predicted fine LST with residual correction for all test timestamps: 2020-05-30 (a), 2020-06-15 (b), 2022-04-19 (c), 2022-10-19 (d), 2023-05-08 (e), 2023-06-08 (f) and 2023-09-04 (g).

To better ascertain how the models perform at the distribution tails of the fine LST, **Figure 14** presents boxplots of the test fine prediction error (with residual correction) obtained by each model for different interquantiles of the true fine LST: between 0 and 10-th, 10 and 90-th, and 90 and 100-th percentiles. Note that since the extreme values of each timestamp and not of all of them combined are wanted, the computed percentiles are timestamp-specific. To better visualise the distributions, the outliers were removed in the subfigure on the left-hand side (a) and not in the subfigure on the right-hand side (b). In both subfigures, the boxplot whiskers are constrained to the 2-nd and 98-th percentiles of the error distributions. The subfigure on the left-hand side (a) shows that the single-timestamp LR model is the one that best performs at the extreme interquantiles. The tree-based models are visibly worse than all others at the higher extreme interquantile. The subfigure also shows a significantly higher tendency for the models to underpredict at the higher extreme interquantile, which could be explained by the already mentioned breakage of scale-invariance. The subfigure on the right-hand side of **Figure 14**, (b), reveals the presence of quite significant outliers in the error distributions, especially at the lower extreme interquantile. Some of the outlying errors even surpass 30K of magnitude.

 doi:10.20944/preprints202604.0495.v1

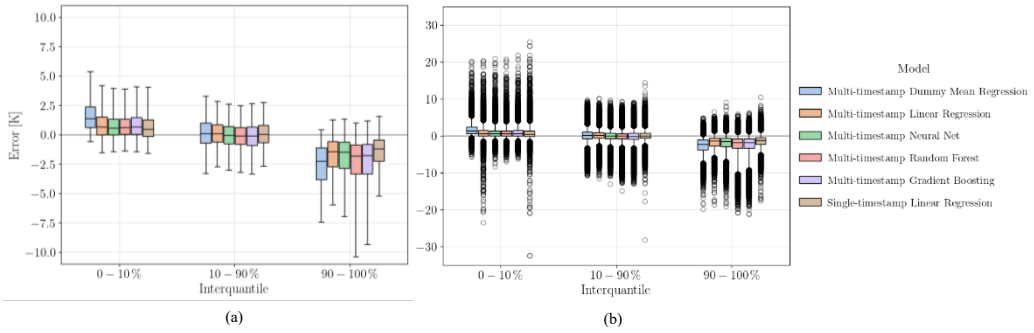


Figure 14. Boxplot distributions of the test fine prediction error (with residual correction) of each model for low, intermedium and high value interquantiles of the true fine LST. For both (a) and (b) plots, the whiskers are constrained to the 2-nd and 98-th percentiles of the error distributions. The outliers were removed from the plot (a) and kept in the plot (b).

One may visualise the errors within more detail through a predicted-vs-actual plot for the fine LST hued by Local Climate Zone (LCZ). **Figure 15** presents such a plot for the case of the single-timestamp LR model. With it, one concludes that the outlying large errors at small LST values in each timestamp mostly occur at points around water bodies. These are, overall, associated with underpredictions, with the exception of timestamp 2020-06-15, in which overprediction is instead quite predominant. Significant overprediction at small LST values also occur in some cases for open mix-rise (that is, lower built-up densities intertwined by green areas), dense trees and low plants' regions. Regarding outliers at high LST values, these seem to be associated with underprediction at some low plants, large low-rise (openly arranged buildings of 1 to 3 stories tall within paved soil) and heavy industry areas. For the case of the multi-timestamp ML models such as NN, with the respective plot being presented in **Figure 16**, all these patterns except the large errors in the water bodies seem to occur. One finds the distribution of points in the predicted-vs-actual plots to be, overall, wider – therefore, being associated with associated with a larger error – for the multi-timestamp models than for the single-timestamp LR one. And as mentioned before, solely results for timestamp 2022-04-19 are unequivocally better for the multi-timestamp LR and NN model than for the single-timestamp LR model.

 doi:10.20944/preprints202604.0495.v1

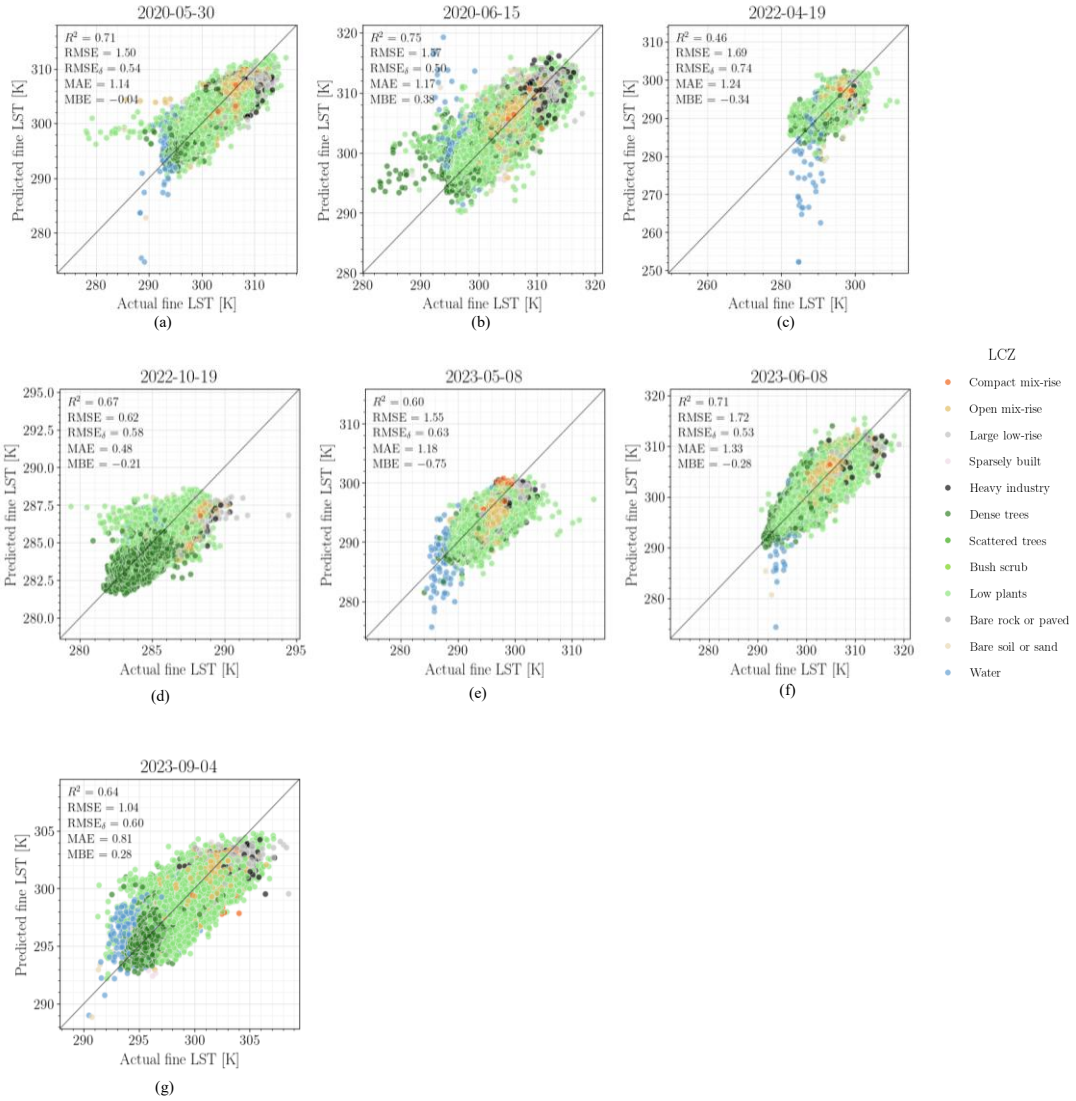


Figure 15. Predicted versus actual fine LST (with residual correction) using the single-timestamp LR model for each test timestamp: 2020-05-30 (a), 2020-06-15 (b), 2022-04-19 (c), 2022-10-19 (d), 2023-05-08 (e), 2023-06-08 (f) and 2023-09-04 (g). The data is hued by Local Climate Zone (LCZ).

 doi:10.20944/preprints202604.0495.v1

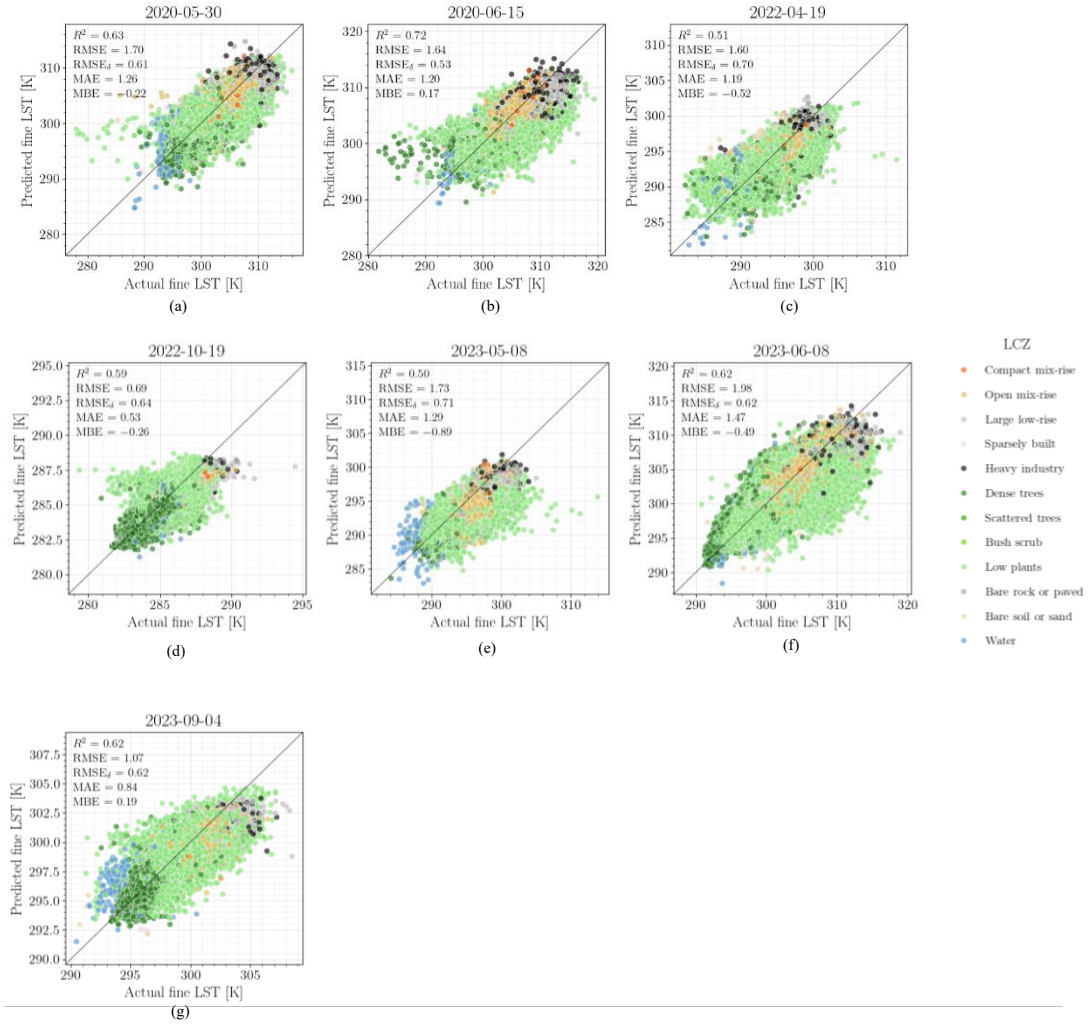


Figure 16. Predicted versus actual fine LST (with residual correction) using the multi-timestamp NN model for each test timestamp: 2020-05-30 (a), 2020-06-15 (b), 2022-04-19 (c), 2022-10-19 (d), 2023-05-08 (e), 2023-06-08 (f) and 2023-09-04 (g). The data is hued by Local Climate Zone (LCZ).

Figure 17 and **Figure 18** present boxplot distributions of test's fine prediction error (with residual correction) obtained by the downscaling models for each LCZ class, without and with outliers, respectively. The figures show that most classes have a tendency for being underpredicted rather than overpredicted. In, fact, solely water bodies have a clear tendency for being overpredicted (as these are usually associated with the smallest true temperatures). Dense trees, low plants and bare soil or sand (in the case of NN and single-timestamp LR) neither tend to be under or overpredicted. All the other, which are mostly associated to built-up, or open areas, with less vegetation, tend to be underpredicted (as these are usually associated with the highest true temperatures). **Figure 18** shows that more dispersed outliers occur for open mix-rise, dense trees, low plants, bare soil or sand, and, in the case of linear models, also water bodies. This latter shows how non-linear models may, in some cases, better "capture" particularities of the data than the linear ones.

 doi:10.20944/preprints202604.0495.v1

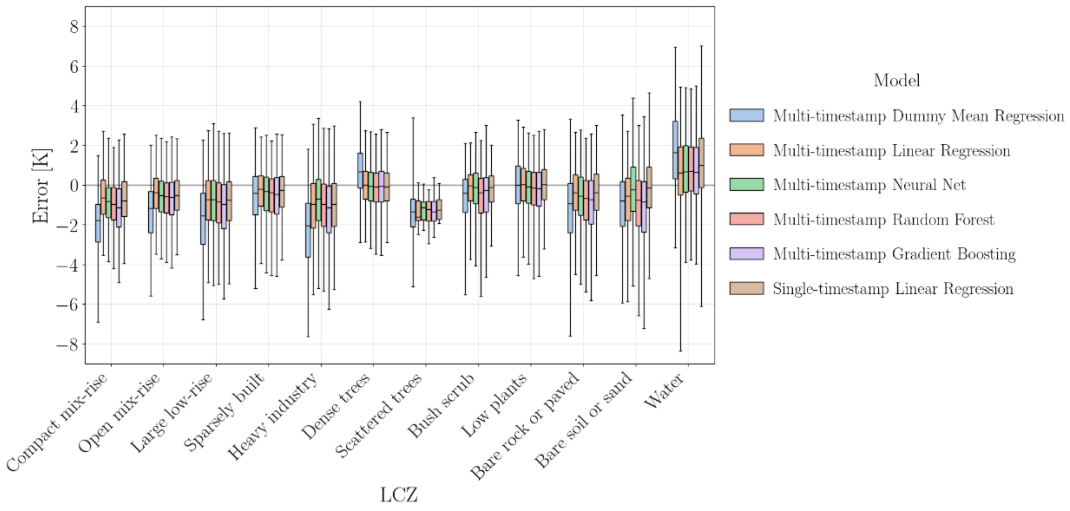


Figure 17. Boxplot distributions of the test fine prediction error (with residual correction) of each model for each LCZ class, with the outliers removed. The whiskers are constrained to the 2nd and 98th percentiles of the error distributions.

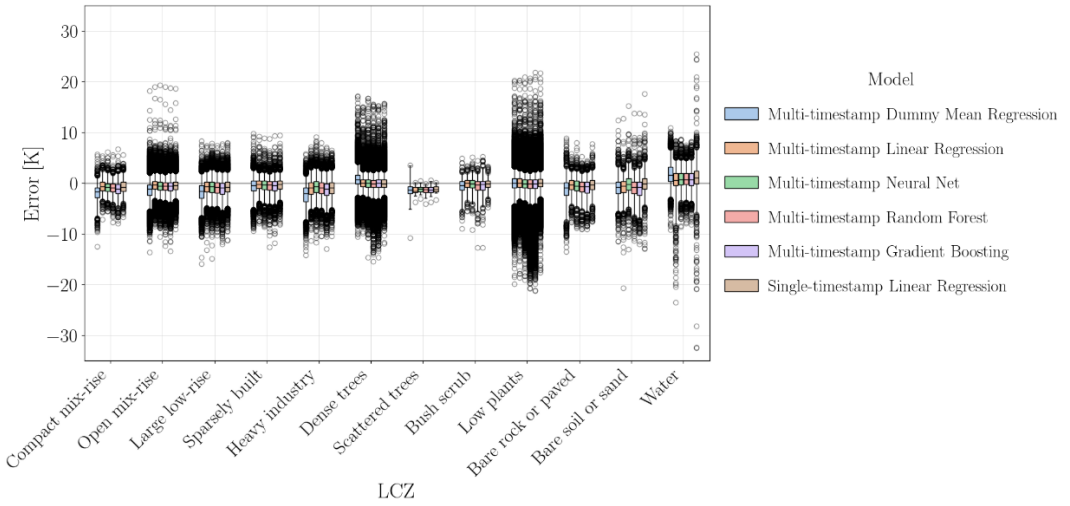


Figure 18. Boxplot distributions of the test fine prediction error (with residual correction) of each model for each LCZ class, with outliers. The whiskers are constrained to the 2nd and 98th percentiles of the error distributions.

Figure 18 further reveals that, although extremely rare, errors of very large magnitude may occur, reaching as much 30 K. Indeed, the outlying error values obtained by the different models are even comparable to the ones obtained by DMR. This makes one to suspect that such errors are probably originated from highly incorrect fine estimations which simultaneously contrast with reasonable coarse ones, not allowing the former to be corrected by the residuals of the latter. While such large fine errors may result from outlying fine predictor values (or from punctual disruption of the established predictor-target relationship), the simultaneous small coarse errors should result from moderate coarse predictor values – as these are obtained from a weighted average or mode (therefore, involving smoothness) of the fine ones. To better assess the phenomenon, one had produced **Figure 19** which presents boxplot distributions of the estimated fine residuals (that is, the finely interpolated coarse residuals), $|\hat{\varepsilon}_{\text{fine}}|$, for different ranges of the fine prediction errors obtained when considering residual correction: between -5 and 5, 5 and 15, 15 and 40 K and their symmetric. The figure reveals that the interquartiles of the residuals are consistently small (< 5 K) with median around 0 except for the case of the single-timestamp LR model at the extremes of the fine prediction errors (between -40 and -15, and 15 and 40 K). In this case, the interquartiles become slightly larger than 5 K but still smaller than 10 K and the median remains close enough to 0. One may safely say that most of the obtained residuals are, overall, small throughout the whole range of errors, which further shows that the main cause for the extremely large fine prediction errors is an incorrect estimation of the fine target whose finely interpolated coarse prediction residual cannot compensate

 doi:10.20944/preprints202604.0495.v1

for. This makes one to also conclude that residual correction is highly ineffective for cases of extreme fine estimation error.

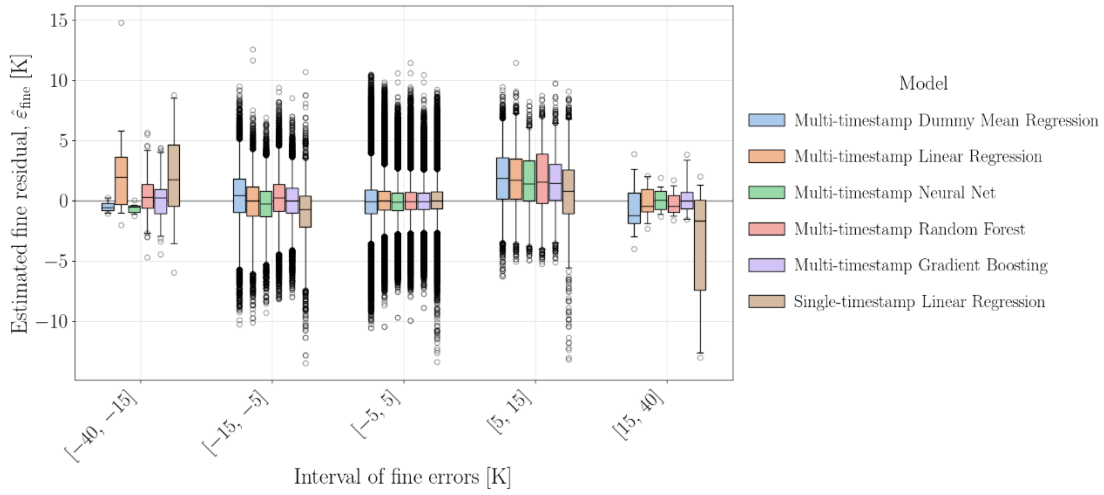


Figure 19. Boxplot distributions of the estimated fine residuals (that is, the finely interpolated coarse residuals), $|\hat{\varepsilon}_{\text{fine}}|$, for different ranges of the fine prediction errors obtained when considering residual correction. The whiskers are constrained to the 2nd and 98th percentiles of the estimated fine residuals.

It is also possible that the extreme fine prediction errors derive from punctual high discrepancies between Sentinel-3 and Landsat data, with the latter being regarded as the supplier of ground-truth fine LST values. To assess this, boxplot distributions of the differences between Sentinel-3 and Landsat's coarse LST values were obtained for the different timestamps and LCZ classes – as shown in **Figure 20**. The figure reveals that all interquartiles of the differences are not too significant, being encompassed by the interval $[-5, 5]$ K. This means that most of the data acquired by Sentinel-3 and Landsat should not disagree too much from each other. However, the distributions do also show a considerable number of outliers that even surpass the -10 and 10 K limits, especially in open mix-rise, low plants and water regions. Moreover, note that coarsening involves smoothing, and, therefore, one could expect even larger differences in the fine grid. Such discrepancies seem to be transcendent to all timestamps except 2022-10-19 for which they were found to be consistently small (being encompassed by the interval $[-5, 5]$ K). Curiously, this is not the timestamp with the smallest difference between Sentinel-3 and Landsat's sensing times but the one with the fourth smallest difference. This means that the large discrepancies obtained in the other timestamps may not be related to also large temporal differences. The current results reveal that the differences between acquisitions platforms can indeed lead to punctual large errors in the fine predictions.

 doi:10.20944/preprints202604.0495.v1

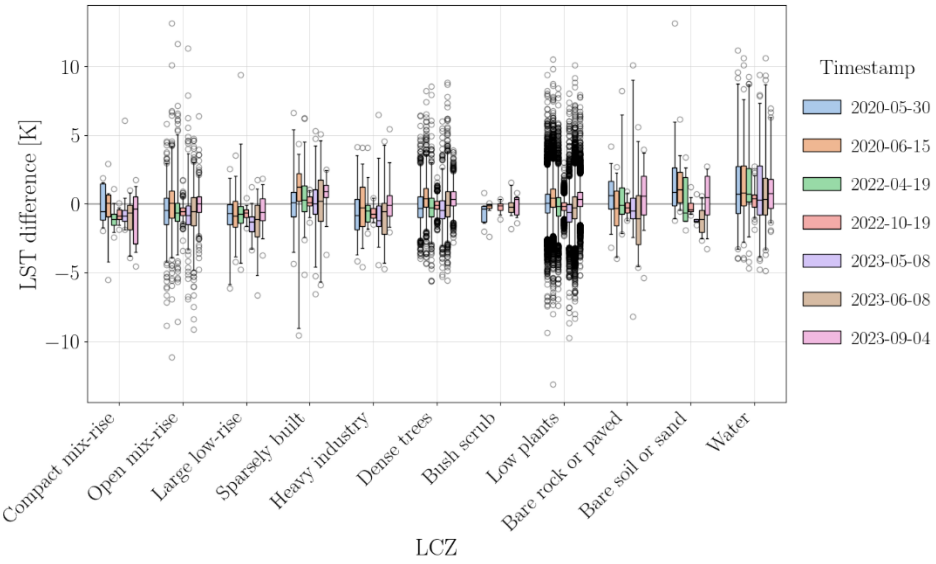


Figure 20. Boxplot distributions of the differences between Senitnel-3 and Landsat's true coarse LST for each timestamp and LCZ class. The whiskers are constrained to the 2nd and 98th percentiles of the LST differences.

3.4. Test Score Distributions

Figure 21 presents boxplot distributions of the fine metrics obtained by the models for the test timestamps when considering residual correction. Regarding R^2, one finds single-timestamp LR to have the highest upper limit (0.75), with the multi-timestamp LR and NN models tightly coming second (both with 0.72). The lower limits of these two last models (0.54 and 0.50, respectively) are, however, greater than the one of the single-timestamp LR model (0.46) (since the latter performs quite poorly for timestamp 2022-04-19). The lower R^2 limits of the RF and GB models (0.35 and 0.38, respectively) are equal or even smaller than the one of DMR (0.38) (and note that DMR in fine prediction with residual correction is equivalent to pure interpolation). And the upper R^2 limits (both with 0.65) are only greater than the one of DMR (0.59) by 0.06, evidencing how the tree-based models are suboptimal fine predictors.

As previously stated, one should note that the test coarse Sentinel-3 and Landsat LST were correlated with a score of $R^2 \sim 0.71$, and, therefore, that not much higher values could be expected for the fine test score of the downscaling models. However, even though the R^2 lower limits obtained with the LR and NN models are significantly smaller than this reference value, the distributions do encompass it. Considering the current conditions, one may safely state that LR and NN models are indeed good performers.

One should point out that the tree-based models are the only models – together with DMR – whose upper RMSE limit surpasses 2 K. Regarding the MBE metric, one may say that all upper limits are positive and all lower ones are negative, showing that both under and overprediction tend to occur. The single and multi-timestamp LR models are the ones with the greatest upper limit (0.38 and 0.39 K, respectively), being similar to DMR's (0.36 K). The lower limit of the multi-timestamp LR model (−0.31 K) is the one closest to 0, therefore, being the model with the smallest tendency for underprediction. Conversely, the lower limits of the RF and GB models (−1.06 and −1.00 K, respectively) are the most negative ones. Clearly, these are the ones with the highest tendency for underprediction.

 doi:10.20944/preprints202604.0495.v1

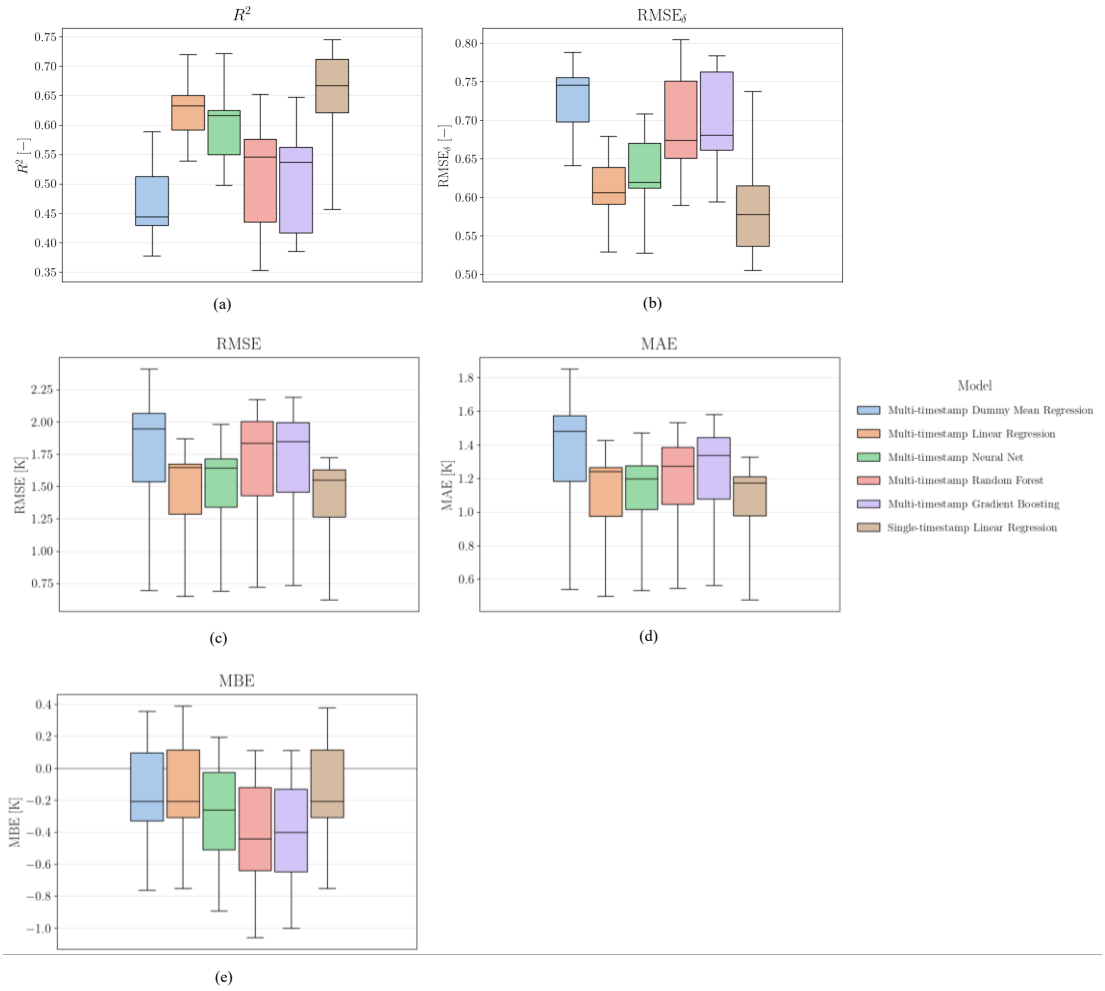


Figure 21. Boxplot distributions of the test scores in fine prediction (with residual correction): R^2 (a), RMSE$_\delta$ (Root Mean Squared Error of the standardised predicted LST using the statistics of the true target in the standardisation) (b), RMSE (Root Mean Squared Error) (c), MAE (Mean Absolute Error) (d) and MBE (Mean Bias Error) (e). The whiskers are constrained to the 0nd and 100th percentiles of the error distributions.

3.5. Resultant Maps

Due to the large extent of the AOI, a full visual comparison between the true and predicted LST maps would be unfeasible. A visual comparison may rather be done instead at a FUA level. **Figure 22** presents the coarse Sentinel-3 and fine Landsat LST values for Odense on 2023-06-08 while **Figure 23** presents the fine values predicted by the downscaling models considering residual correction. Analogous figures for all the other regions are found in **Appendix A.3**, whose remarks and conclusions could be said to be identical to the ones which are herein presented for Odense. To facilitate comparison, all maps of each region are classified and displayed using a common colour scale, allowing the LST values at the pixel level to be directly interpreted and compared across the different images based on their colour representation. Odense's maps show that all predicted fine-resolution LST products (except dummy model's) exhibit similar spatial patterns in terms of distribution, colour tones, and texture features associated with different land-cover types (e.g., urban areas, bare land, and vegetation). Specifically, distinct highway patterns can be identified in all predicted LST maps, with hues closely matching those observed in the Landsat-derived fine-resolution LST. Such fine-scale linear features are not distinguishable in the original coarse-resolution LST data, nor are they captured in the fine-grid output of the dummy model. Although all models substantially enhance the visual quality relative to the coarse-resolution LST by effectively reducing the mosaic (tiling) artifacts, the single-timestamp model clearly outperforms the

 doi:10.20944/preprints202604.0495.v1

others in capturing extreme high-temperature signals in urban areas. This indicates that single-timestamp LR is more effective at highlighting thermal contrasts between land-cover features, making it particularly suitable for urban heat island analysis. These observations are consistent with the quantitative evaluation, in which the single-timestamp model achieves significantly lower errors in general and at the high extremes when compared to the other methods.

Note that in spite of the lowest LST values being consistently observed over water bodies in both the Landsat and Sentinel-3 LST products, these areas are absent from the predicted LST maps. This is due to the fact of the original Sentinel-3 SYN products (whose derived variables are used as predictors of the downscaling models) have their water pixels masked out.

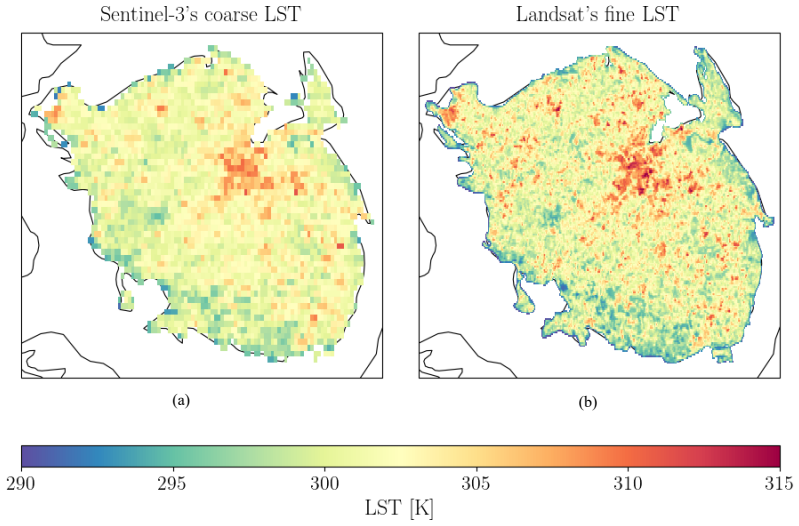


Figure 22. Actual coarse (a) and fine (b) LST (with residual correction) for Odense on 2023-06-08.

 doi:10.20944/preprints202604.0495.v1

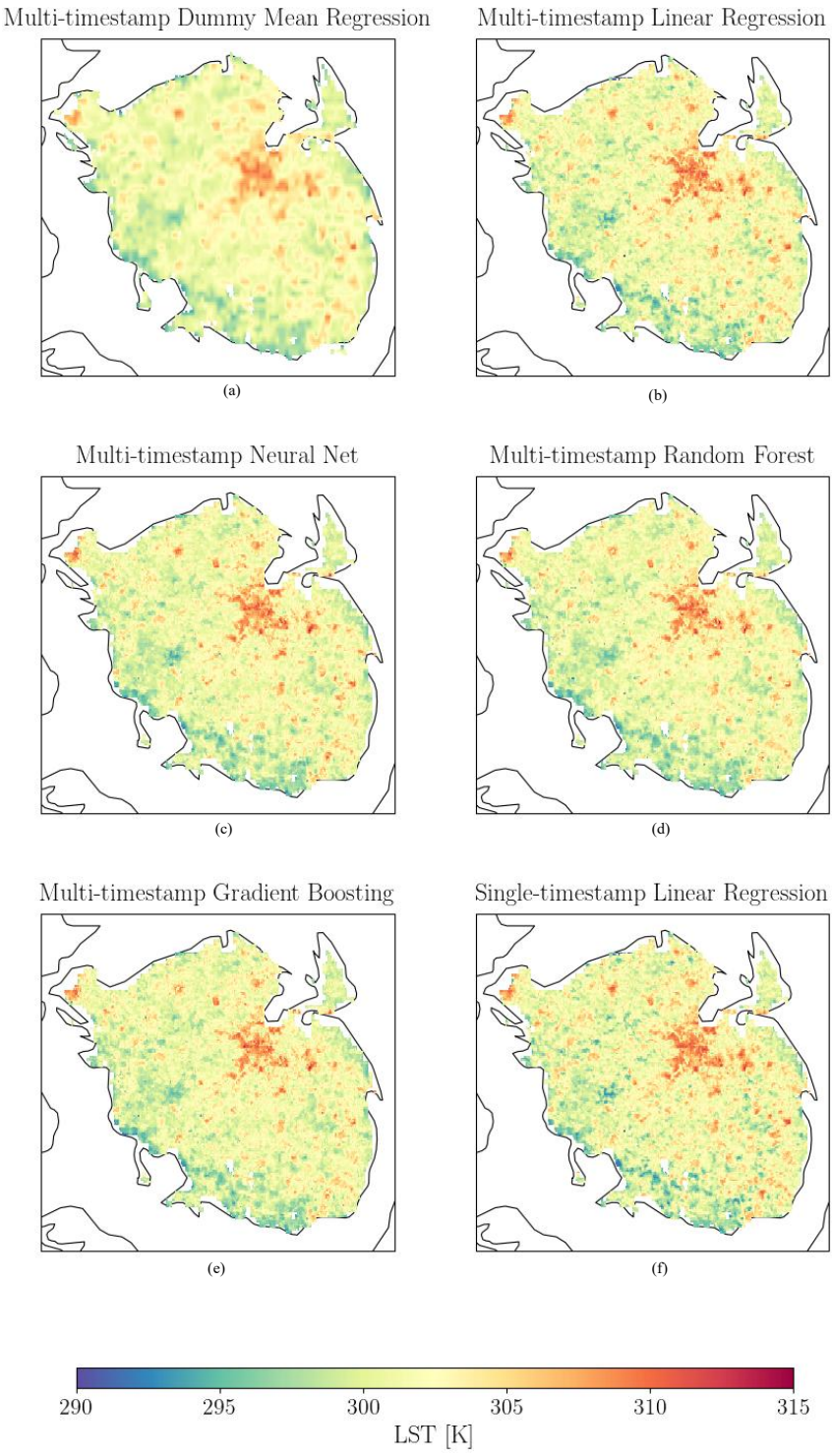


Figure 23. Predicted fine LST (with residual correction) for Odense on 2023-06-08: multi-timestamp Dummy Mean Regression (a), multi-timestamp Linear Regression (b), multi-timestamp Neural Net (c), multi-timestamp Random Forest (d), multi-timestamp Gradient Boosting (e) and single-timestamp Linear Regression (f).

Figure 24 presents the error obtained by downscaling models for Odense on 2023-06-08. One should note here that the colourmap limits were truncated to avoid presenting sparse outliers that would overwhelm the colour spectrum. Not surprisingly, the dummy model is the one with highest amount of extreme error values. The ML architectures are the ones with a greater population of negative errors, (emphasising underestimation) and that the single-timestamp LR is the one with the smallest amount of extreme error values.

 doi:10.20944/preprints202604.0495.v1

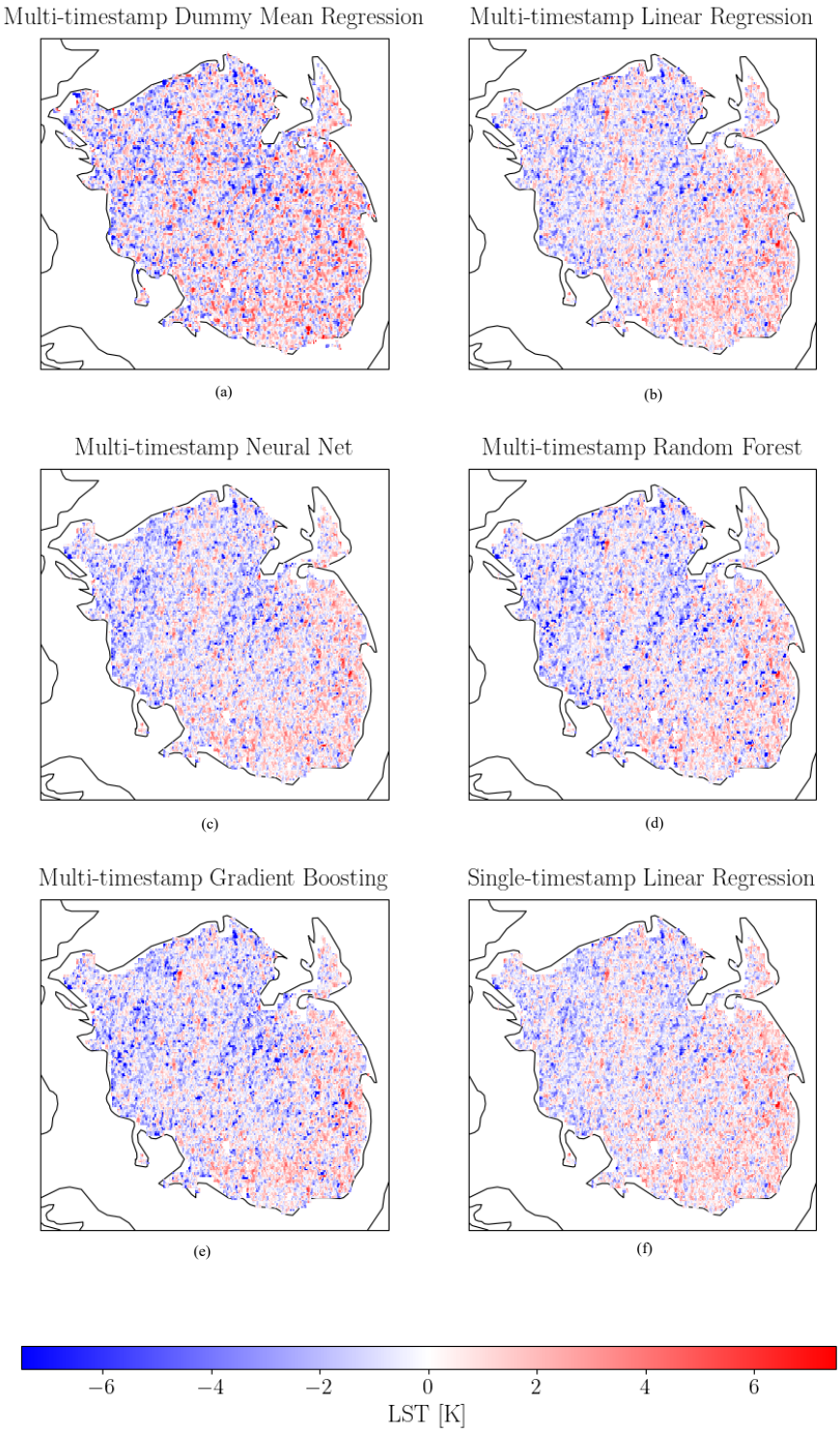


Figure 24. Fine prediction error (with residual correction) obtained by the downscaling models for Odense on 2023-06-08: multi-timestamp Dummy Mean Regression (a), multi-timestamp Linear Regression (b), multi-timestamp Neural Net (c), multi-timestamp Random Forest (d), multi-timestamp Gradient Boosting (e) and single-timestamp Linear Regression (f).

3.6. Feature Importance according to Best Coarse Predicting Model

The feature selection study which was previously performed with the multi-timestamp LR model gave a hint about the relative importance of each feature in a linear model for the relation between feature and target in the coarse grid. However, it would also be pertinent to numerically quantify these importances using the model that best captures the coarse relation: the multi-timestamp GB model. Fortunately, the XGBoost package already provides a built-in function for

 doi:10.20944/preprints202604.0495.v1

computing such feature importances. In this work, XGBoost's feature importance was defined as the average information gain across all tree splits the feature is used in. **Figure 25** shows that the most important features for GB are the winter season indicator (that is, whether the season corresponds to winter (returning 1) or not (returning 0)), followed by FVC, IMD, summer season indicator, spring season indicator, NDWI, TCD, UD, d_{coast}, DEM and TOPEX. The quite large value for the winter season indicator evidences how the data associated with winter strongly differs from the one of all the other seasons and how the inclusion of seasonality is remarkably important in a multi-timestamp ML model. The results of the previous feature selection study, which are presented in **Table 3** already emphasised the importance of FVC, IMD, NDWI and TCD. However, even though it regarded d_{coast} as the second-best numerical feature, this variable was relegated to sixth best numerical predictor in the current study. Still, one should note that the results of the two analyses are not completely comparable since they rely on different methods.

The curious reader may be intrigued by the non-consideration of an autumn season indicator as a feature of the GB model. One may show that the autumn season indicator is redundant with the presence of all other season indicators: if all the latter do not indicate the respective season, then autumn is necessarily the actual season.

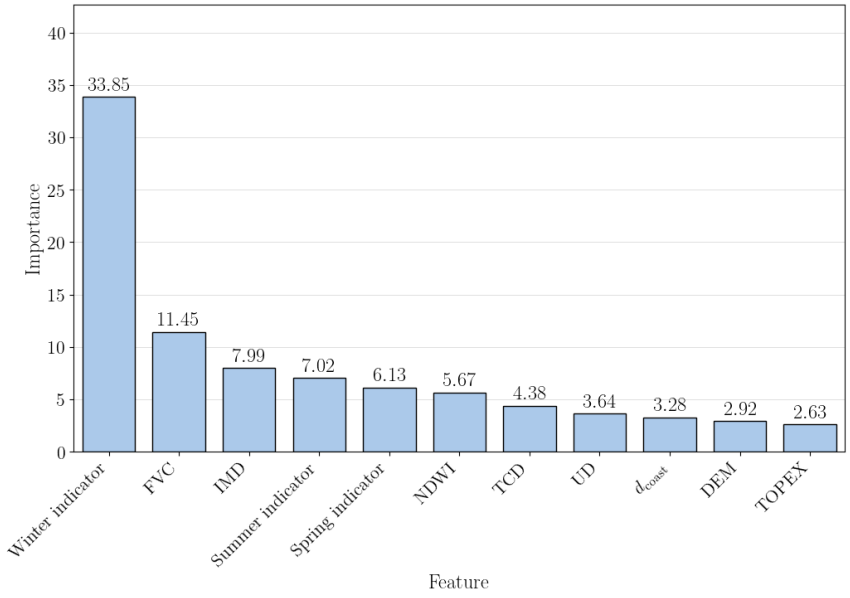


Figure 25. Importance of each feature in the multi-timestamp GB model. Feature importance is herein defined using XGBoost package as the average information gain across all tree splits the feature is used in.

4. Discussion

Numerous studies in the literature have discussed and benchmarked single-timestamp scale-invariance models. However, studies addressing multi-timestamp models, such as the ones considered in this work, remain scarce. This had made the scrutiny of multi-timestamp models to be a difficult exercise within the current state-of-the-art. Still, the results and learnings from other works regarding single-timestamp models can guide one in the judgement of all models employed in the present study.

Suboptimal performance of the ML models in fine-scale prediction when trained with coarse-scale data, as found in the present work, had been emphasised in other studies. Ding et al. [52] simulated coarse LST from Landsat fine data through spatial averaging and downscaled it with different single-timestamp models, one of them corresponding to RF. As in the present work, they did observe that RF could not predict the extremes of the original fine LST data due to the loss of such information when simulating the coarse data. They termed the phenomenon "boundary effect". Furthermore, the results of the experiment done by Hernanz et al. [53] on the extrapolation of NN and Support Vector Regression models for the prediction of maximum surface temperature made

 doi:10.20944/preprints202604.0495.v1

them to state that the "ML techniques can perform wrong under extrapolation" and that "their suitability for SD of climate change projections should be seriously questioned". They further mention that "experiments which validate over spatially/temporarily aggregated data might hide extrapolation problems in finer spatial/temporal scales" – as are the scale-invariance architectures considered in the present work. Moreover, underperformance may additionally be exacerbated by the non-verification of the scale invariance principle. Gao et al. [54] concluded that although "this assumption is reasonable for a uniformly (homogeneous) vegetated area and worked well in rainfed agricultural areas", the "LST-NDVI relationship is not well-defined over many complex heterogeneous landscapes" as demonstrated in previous studies [55–60]. This is, in fact also the case of the Danish Functional Urban Areas considered in the present work, which are described by a large variety of LCZ classes (as evidenced by **Figure 15** and **Figure 16**).

A vast amount of works report significant improvements when using a single-timestamp RF model in place of the ubiquitous TsHARP or DisTrad models. However, while these RF models employed multiple predictors, TsHARP and DisTrad solely regard one (FVC and NDVI, respectively), which would make one wonder if such relative improvements would still occur if the multiple predictors were also used in an LR model. Furthermore, most works issue general scores for the downscaling of the whole data, but not particular ones for the data extremes (which are notably important in the context of urban planning). Works such as the ones of Li et al. [61] and Wu and Li [62], who downscaled coarse MODIS LST and validated the result with fine ASTER data, showed that the single-timestamp RF model can perform, overall, significantly better than TsHARP. However, the resultant maps revealed that, in contrast with TsHARP, RF could not predict the low and high extremes of the true fine LST. Moreover, Wu and Li [62], showed that for impervious surfaces, which usually contain the highest LST values in the scene, a multiple predictor RF model may perform even worse than the single predictor TsHARP model. These findings suggest that underperformance of the tree-based models in the extremes, as observed in the present work, is a common limitation of this type or architecture.

Strong evidence had been found in the literature for the benchmarking results of the downscaling models being highly dependant on the LST coarse and fine resolution products (e.g., instruments, bands used, processing algorithm) used in training and validation. Hutengs and Vohland [22] showed that while RF was able to perform better than TsHARP when training with simulated coarse data from a validating fine Landsat one, it did it marginally when training with coarse MODIS data and validating with the same Landsat data. And it should be noted again that while RF used several predictors in the downscaling, TsHARP used solely one (FVC). The authors emphasised the fact that using coarsened and fine LST data from the same instrument for training and validation, respectively, was a "best case scenario", which in contrast to the case in which different instruments were used, did not suffer from mismatches in radiometric processing, geo-referencing (as well as acquisition time). Because of this, many downscaling models have been tested within such framework [30,52,54,63,64]. However, one cannot guarantee that these models would perform identically well in practical downscaling applications which use "real" coarse LST instead of a simulated one. In the case of using different training and validating instruments, and as previously concluded, the authors stated that the not much better results that they obtained when downscaling with a RF model instead of TsHARP could be partially justified by the absence of extremes in the training coarse data (i.e., the MODIS LST values do not encompass the high extreme ones of Landsat's). Because of this, the authors stated that "for RF regression (…) one has to be aware that the predictive range of LSTs is restricted to those covered by the training data". With all these reflections, one may more confidently postulate that limitations of the downscaling performance of the models obtained in the present work could indeed be partially explained by the mismatch between the two instruments (and the different LST processing algorithms for each one) issuing the training and validating data. And this has been further emphasised by the differences which were herein found in coarse-prediction performance when using either Sentinel-3 or Landsat's LST as validating data.

 doi:10.20944/preprints202604.0495.v1

It is worthy to highlight the study of Wang et al. [65], which considered LST data simulation in their downscaling models: instead of the ubiquitous area-weighted spatial average, the authors employed a Planck's law-based method to obtain training coarse LST data from a validating fine Landsat one. This method seems to conserve much more the extremes of the fine data than the traditional one, as presented by the resultant maps shown in their work. Subsequently, the trained tree-based models were found not to suffer from the "boundary effect" and were able to predict for the tails of the fine LST distribution. However, one should point out that for the case of the coarsening of Landsat-derived predictors, the traditional area-weighted spatial average was the method used, which could create a mismatch between the predictor and target value distributions – something that merits future study. Wang et al. revealed that all ML models performed better than the LR ones, though, again, by considering multiple predictors for the former and just one for the latter (since TsHARP and DisTrad were tried). Nonetheless, even though the LST coarsening method employed by the authors only makes sense in the realm of LST data simulations, it may motivate one to consider identical approaches for "real" world practices, and this does not only include transformations of the coarse data but also data augmentations.

Other possible steps for future work are to extend the list of predictors by including all the other bands of the Sentinel-3 SYN products as well as LCZ and area-specific variables (such as FUAs and sub-tiles of the scenes). Moreover, to overcome the issue of scale-invariance breakage, an approach such as the one implemented by Ait-Bachir et al. [42], which does not rely on such principle, can be considered. The strategy regarded by the authors consisted in the interpolation of coarse LST onto the fine grid, and training of a model that predicts a fine target with it as well as fine NDVI so that the degradation of the prediction (its interpolation onto the coarse grid) gets as close as possible to the original coarse LST. Note that this, however, relies on another hypothesis: that the inverse transformation of the coarsening of the predicted fine target concomitantly also gets as close as possible to the true fine LST data. Indeed, coarse interpolation results in loss of information, and one may show that it is possible for the degradation of different fine scenes to result in a common coarse scene – the inverse transformation can then have multiple solutions. This means that even though the degraded predicted target can get close enough to the true coarse one, the predicted fine target can only get close to the true fine one within some irreducible tolerance.

Finally, where urban applications are envisaged, some studies have looking into different downscaling approaches that look into energy balance equations and well-known urban morphology indicators, proving its usefulness in downscaling LST, albeit offering more case-specific results which may be difficult to generalise to other locations [8,66]. Furthermore, such approaches require additional data sources, certain parametrisations, and more complex data processing workflows, which may hinder their fitness for operational purposes.

5. Conclusions

The primary objective of this study was to assess whether the employment of practical ML models can provide better results than LR in a downscaling pipeline based on the scale invariance principle and residual correction. To do this, multi-timestamp ML models were benchmarked against a single-timestamp LR alternative. The performed exploratory data analysis revealed that to preserve the strong within-timestamp correlation of the spatio-temporal predictors (FVC, NDVI and NDWI), the multi-timestamp models would need to be trained to infer for a standardised LST using also standardised spatio-temporal predictors. The subsequent tuning, training and testing of the resultant models revealed that the multi-timestamp architecture is able to achieve better results for coarse prediction than the single-timestamp one when using GB or NN. However, the improvement was found to not be significantly large. RF performed as well as the single-timestamp model while the multi-timestamp LR performed significantly worse. This demonstrates that ML models can outperform LR in the inference of a target with the same resolution as the training one. Using Landsat's coarsened LST as the true target in the comparison made all RMSE values to significantly change, with the multi-timestamp ML models performing marginally worse than the single-

 doi:10.20944/preprints202604.0495.v1

timestamp architecture. Such results evidence that the distributions of the LST values retrieved by Sentinel-3 and Landsat do have differences as already suggested by the obtained not too high correlation coefficient (0.84).

Regarding fine predictions without residual correction, all scores got significantly worse than in coarse prediction, which evidences the breakage of scale invariance. GB and RF herein corresponded to the worst fine predicting models. And solely NN was able to perform better than multi-timestamp LR. As found in previous works, the suboptimal performance of the tree-based models could be justified by the so-called "boundary effect", in which the absence of the fine extremes in the training coarse data makes the downscaling models to extrapolate in fine-prediction. In such conditions the tree-based models tend to behave poorly. When considering residual correction, all produced RMSE values abruptly decreased. However, the single-timestamp model remained the best model and the multi-timestamp NN model ceased to outperform the multi-timestamp LR one (with a RMSE that was worse by 0.04 K).

The present work also revealed that training with data from multiple timestamps instead of from solely one may actually deteriorate performance of the downscaling models, as common generalities (averages) between the different timestamps tend to be more emphasised than their particularities (extremes). This is evidenced by the greater capacity for the single-timestamp model to predict the tails of the target value distribution when compared to the other models. Still, it is important to note that the single timestamp model was found to be better than the multi-timestamp models for almost all timestamps, but not all, suggesting that the coarse data that is available for a given timestamp may be less suitable for fine-scale inference at that same timestamp compared to data from other timestamps.

This study demonstrated that in the realm of scale-invariance-based models, the simplicity and overall performance of the single-timestamp LR position it as the best candidate for operational LST downscaling applications.

Author Contributions: Conceptualization, E.P. and M.K.; methodology, E.P., M.K. and Q.P.; software, E.P. and M.K.; validation, E.P. and M.K.; formal analysis, E.P. and M.K.; data curation, I.G. and B.M.; writing—original draft preparation, E.P.; writing—review and editing, E.P., I.G., V.F.V.V.M., H.J.D.S., Q.P. and A.P.O.; visualization, E.P. and M.K.; supervision, A.P.O.; project administration, H.J.D.S., Q.P. and A.P.O.; funding acquisition, H.J.D.S. and A.P.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Space Agency (ESA) under Contract No. 4000143628/24/I-DT (AI Trustworthy Applications for Climate).

Data Availability Statement: The upsampled LST data produced in this study using the benchmarking single-timestamp LR downscaling model (which was found to be the ultimate best performer) are openly available in Zenodo at: https://doi.org/10.5281/zenodo.19220975 (accessed on 26 of March of 2026).

Acknowledgments: This activity was carried out in the context of the CLIM4cities project, led by the Danish Meteorological Institute (DMI) in collaboration with +ATLANTIC CoLAB. The authors gratefully acknowledge the support and guidance of the ESA Φ-lab and Climate Office throughout this work. The views expressed in this publication are those of the authors and do not reflect the official opinion of the European Space Agency.

This publication has been prepared using European Union's Copernicus Land Monitoring Service information; https://doi.org/10.2909/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6 (Urban Atlas 2018), https://doi.org/10.2909/3bf542bd-eebd-4d73-b53c-a0243f2ed862 (Imperviousness Density 2018 — 10 m) , https://doi.org/10.2909/e677441e-fb94-431c-b4f9-304f10e4dfd8 (Tree Cover Density 2018 — 10 m), https://doi.org/10.2909/82f93572-9888-47ef-97a1-5cac5985a26a (Dominant Leaf Type 2018), https://doi.org/10.2909/0b6254bb-4c7d-41d9-8eae-c43b05ab2965 (Grasslands 2018 – raster 10 m), https://doi.org/10.2909/71c95a07-e296-44fc-b22b-415f42acfdf0 (Corine Land Cover 2018 – raster).

Conflicts of Interest: The authors declare no conflicts of interest. A scientific employee of the funding body took part in analyzing and interpreting the results, as well as in reviewing the manuscript.

 doi:10.20944/preprints202604.0495.v1

Appendix A

Appendix A.1

- **Unique identifiers of the matched satellite products**

Table A1. Unique identifiers of satellite imagery of the matching dates between Sentinel-3 and Landsat 8/9.

Sentinel-3	Landsat- 8/9
S3A_SL_2_LST____20200530T101738_20200530T102038_20200531T155247_0180_059_008_1980_LN2_O_NT_004.SEN3	LC08_L2SP_196020_20200530_20200820_02_T1 LC08_L2SP_196021_20200530_20200820_02_T1
S3A_SL_2_LST____20200615T100239_20200615T100539_20200616T162427_0179_059_236_1980_LN2_O_NT_004.SEN3	LC08_L2SP_196020_20200615_20200823_02_T1 LC08_L2SP_196021_20200615_20200823_02_T1
S3B_SL_2_LST____20220419T101543_20220419T101843_20220420T063118_0179_065_065_1980_PS2_O_NT_004.SEN3	LC09_L2SP_195021_20220419_20230421_02_T1 LC09_L2SP_195022_20220419_20230421_02_T1
S3A_SY_2_SYN____20221019T101010_20221019T101310_20221021T074604_0180_091_122_1980_PS1_O_NT_002.SEN3	LC09_L2SP_196020_20221019_20230325_02_T1 LC09_L2SP_196021_20221019_20230325_02_T1
S3A_SL_2_LST____20230508T095900_20230508T100200_20230509T191603_0180_098_293_1980_PS1_O_NT_004.SEN3	LC09_L2SP_195021_20230508_20230510_02_T1 LC09_L2SP_195022_20230508_20230510_02_T1
S3A_SL_2_LST____20230608T095513_20230608T095813_20230609T190024_0179_099_350_1980_PS1_O_NT_004.SEN3	LC08_L2SP_196020_20230608_20230614_02_T1 LC08_L2SP_196021_20230608_20230614_02_T1
S3A_SL_2_LST____20230904T101349_20230904T101649_20230905T191154_0180_103_065_1980_PS1_O_NT_004.SEN3	LC09_L2SP_196020_20230904_20230906_02_T1 LC09_L2SP_196021_20230904_20230906_02_T1

Appendix A.2

- **Timestamp-specific centring in a single-predictor LR model**

Timestamp-specific centring in an LR model considering FVC as sole predictor would be such that the LST value at some timestamp t and some pixel is related to the respective FVC value – let these be denoted by LST_t and FVC_t – through

$$\underbrace{\mathrm{LST}_t - \overline{\mathrm{LST}}_t}_{=:\,\Delta\mathrm{LST}_t} = a\,\underbrace{\left(\mathrm{FVC}_t - \overline{\mathrm{FVC}}_t\right)}_{=:\,\Delta\mathrm{FVC}_t} + b\,, \tag{A1}$$

where $\overline{\mathrm{LST}}_t$ and $\overline{\mathrm{FVC}}_t$ correspond to the spatial arithmetic means of LST and FVC at timestamp t, and a and b are the parameters of the LR model. Let $\Delta\mathrm{LST}_t$ and $\Delta\mathrm{FVC}_t$ denote the timestamp-specific centred LST and FVC values, respectively, for timestamp t. If $\Delta\mathrm{LST}_t$ was plotted against $\Delta\mathrm{FVC}_t$, the value distributions for each timestamp would be centred (vertically and horizontally) at the origin.

 doi:10.20944/preprints202604.0495.v1

- **Timestamp-specific standardisation in a single-predictor LR model**

Timestamp-specific standardisation in an LR model considering FVC as sole predictor would be such that the LST value at some timestamp t and some pixel is related to the respective FVC value through

$$\underbrace{\frac{\mathrm{LST}_t - \overline{\mathrm{LST}}_t}{s_{\mathrm{LST}_t}}}_{=:\,\delta\mathrm{LST}_t} = a \underbrace{\frac{\mathrm{FVC}_t - \overline{\mathrm{FVC}}_t}{s_{\mathrm{FVC}_t}}}_{=:\,\delta\mathrm{FVC}_t} + b\,. \tag{A2}$$

where s_{LST_t} and s_{FVC_t} are the sample standard deviations of the LST and FVC values for timestamp t.

Let $\delta\mathrm{LST}_t$ and $\delta\mathrm{FVC}_t$ denote the timestamp-specific standardised LST and FVC values for timestamp t.

- **Similarity between a multi-timestamp single-predictor LR model when considering timestamp-specific standardisation and a single-timestamp single-predictor LR model**

A multi-timestamp LR model considering FVC as sole predictor with timestamp-specific standardisation would be equivalent to a single-timestamp one if the actual LST data were perfectly linear with respect to FVC and the line slopes of the raw data had all the same sign for all timestamps. Indeed, the solution of a LR problem using the raw data of a sole timestamp t corresponds to

$$\delta\mathrm{LST}_t = R_t \cdot \delta\mathrm{FVC}_t\,, \tag{A3}$$

where R_t is the Pearson correlation coefficient between LST and FVC values at timestamp t. If the raw data were perfectly linear for all timestamps and all line slopes shared the same sign, one would expect values of $R_t = 1$ if the slopes were positive, or $R_t = -1$ if the slopes were negative. On the other hand, if $\delta\mathrm{LST}_t$ was plotted against $\delta\mathrm{FVC}_t$ for all timestamps, the resultant lines would be centred at the origin – implying $b = 0$ in (A2) (as in (A3)) – and because in the case of perfect linearity the line slopes of the raw data coincide with $R_t \cdot s_{\mathrm{LST}_t}/s_{\mathrm{FVC}_t}$, the line slopes of the standardised data would correspond to R_t – implying $a = R_t$ in (A2) (as in (A3)).

- **Proof that, when residual correction is considered, a downscaling model whose base one predicts a constant c is equivalent to bilinear interpolation**

According to the flow chart of **Figure 4**,

$$\begin{aligned}
\widehat{\mathrm{LST}}_{\mathrm{fine,\,corr}} &= \left(\underbrace{\delta\widehat{\mathrm{LST}}_{\mathrm{fine}}}_{=c} + \hat{\varepsilon}_{\mathrm{fine}}\right) \cdot S_{\mathrm{LST}_{\mathrm{coarse}}} + \overline{\mathrm{LST}}_{\mathrm{coarse},\,i} \\
&= \left(c + \mathrm{interp}_{\mathrm{fine}}(\varepsilon_{\mathrm{coarse}})\right) \cdot S_{\mathrm{LST}_{\mathrm{coarse}}} + \overline{\mathrm{LST}}_{\mathrm{coarse}} \\
&= \left(c + \mathrm{interp}_{\mathrm{fine}}\left(\delta\mathrm{LST}_{\mathrm{coarse}} - \underbrace{\delta\widehat{\mathrm{LST}}_{\mathrm{coarse}}}_{=c}\right)\right) \cdot S_{\mathrm{LST}_{\mathrm{coarse}}} \\
&\quad + \overline{\mathrm{LST}}_{\mathrm{coarse}} \\
&= \left(c + \mathrm{interp}_{\mathrm{fine}}\left(\frac{\mathrm{LST}_{\mathrm{coarse}} - \overline{\mathrm{LST}}_{\mathrm{coarse}}}{S_{\mathrm{LST}_{\mathrm{coarse}}}} - c\right)\right) \cdot S_{\mathrm{LST}_{\mathrm{coarse}}} \\
&\quad + \overline{\mathrm{LST}}_{\mathrm{coarse}} \\
&= \left(c - c + \frac{\mathrm{interp}_{\mathrm{fine}}(\mathrm{LST}_{\mathrm{coarse}}) - \overline{\mathrm{LST}}_{\mathrm{coarse}}}{S_{\mathrm{LST}_{\mathrm{coarse}}}}\right) \cdot S_{\mathrm{LST}_{\mathrm{coarse}}} \\
&\quad + \overline{\mathrm{LST}}_{\mathrm{coarse}} \\
&= \mathrm{interp}_{\mathrm{fine}}(\mathrm{LST}_{\mathrm{coarse}}) - \overline{\mathrm{LST}}_{\mathrm{coarse}} + \overline{\mathrm{LST}}_{\mathrm{coarse}} \\
&= \mathrm{interp}_{\mathrm{fine}}(\mathrm{LST}_{\mathrm{coarse}}).
\end{aligned}$$

□

 doi:10.20944/preprints202604.0495.v1

(A4)

In the mathematical proof above, one has used the fact that bilinear interpolation (the method regarded in the fine interpolation of the coarse residual) is a linear operator, therefore, having the homogeneity property (the interpolation of a map multiplied by some factor is this factor multiplied the interpolation of the map) and addition property (the interpolation of the sum of maps is the sum of the interpolations of the maps).

Appendix A.3

- **True and predicted LST maps (remainder)**

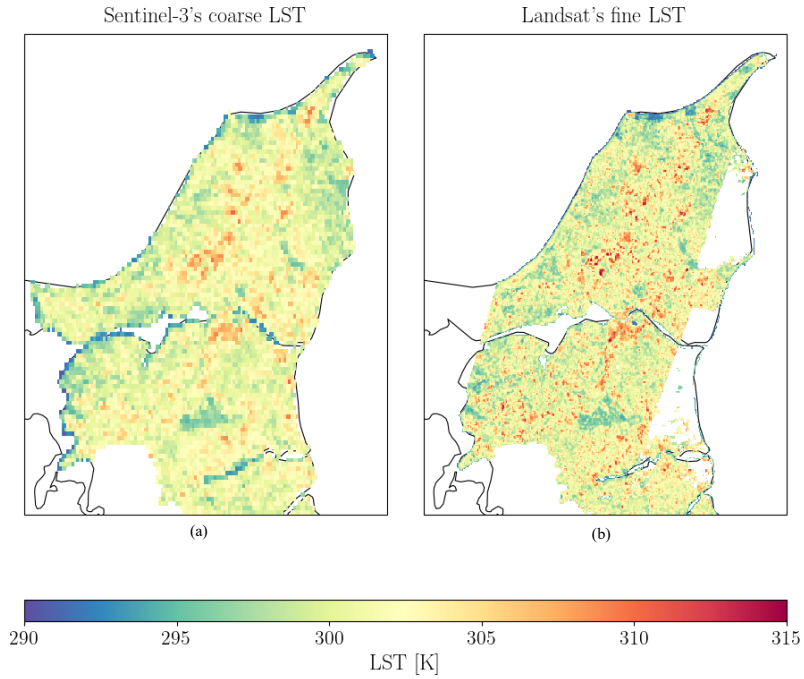


Figure A1. - Actual coarse (a) and fine (b) LST (with residual correction) for Aalborg on 2023-06-08.

 doi:10.20944/preprints202604.0495.v1

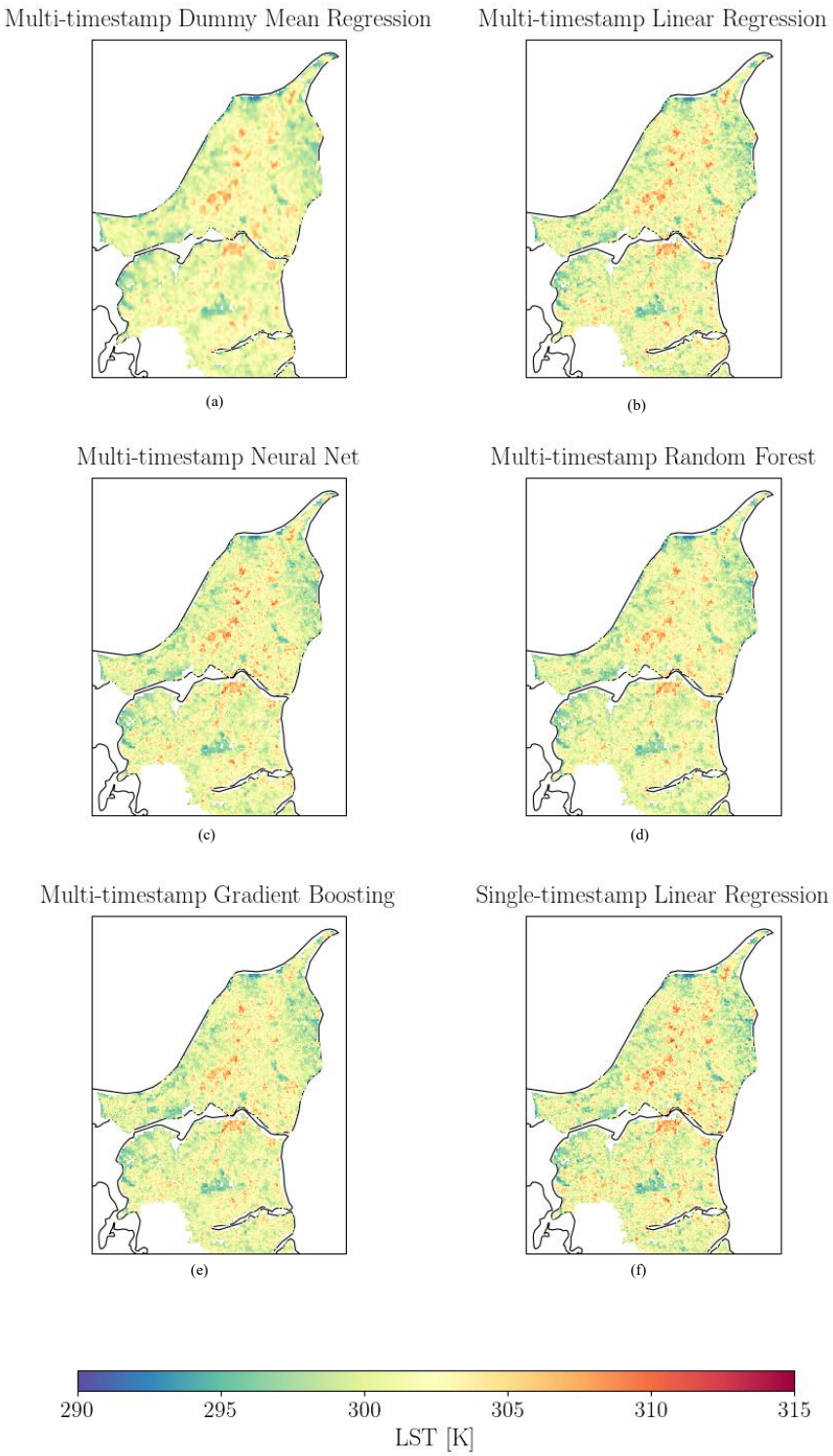


Figure A2. - Predicted fine LST (with residual correction) for Aalborg on 2023-06-08: multi-timestamp Dummy Mean Regression (a), multi-timestamp Linear Regression (b), multi-timestamp Neural Net (c), multi-timestamp Random Forest (d), multi-timestamp Gradient Boosting (e) and single-timestamp Linear Regression (f).

 doi:10.20944/preprints202604.0495.v1

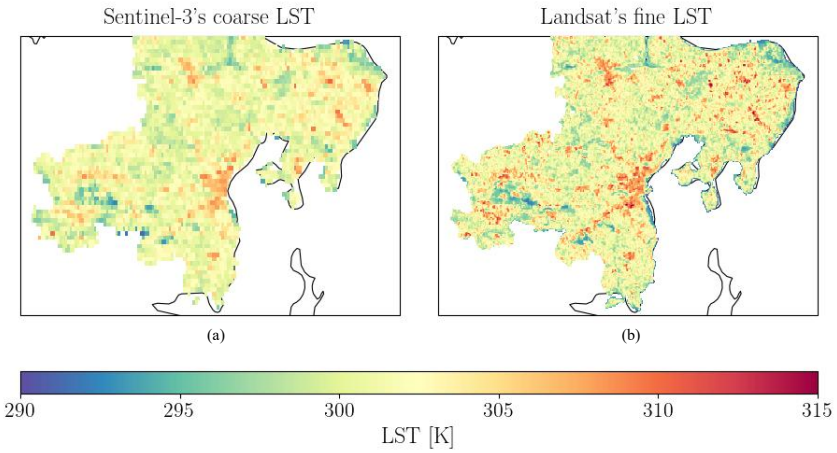


Figure A3. - Actual coarse (a) and fine (b) LST (with residual correction) for Aarhus on 2023-06-08.

 doi:10.20944/preprints202604.0495.v1

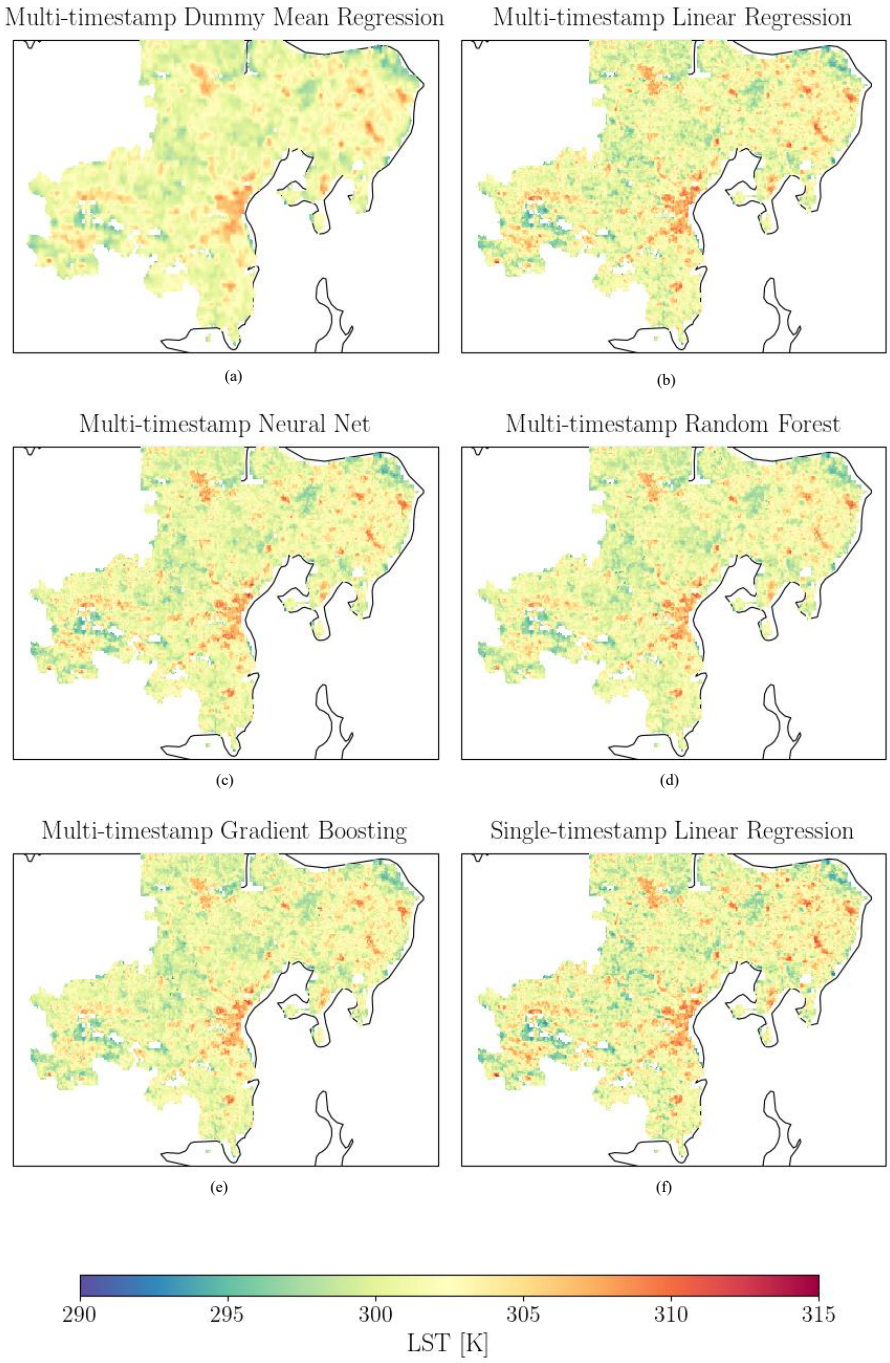


Figure A4. - Predicted fine LST (with residual correction) for Aarhus on 2023-06-08: multi-timestamp Dummy Mean Regression (a), multi-timestamp Linear Regression (b), multi-timestamp Neural Net (c), multi-timestamp Random Forest (d), multi-timestamp Gradient Boosting (e) and single-timestamp Linear Regression (f).

 doi:10.20944/preprints202604.0495.v1

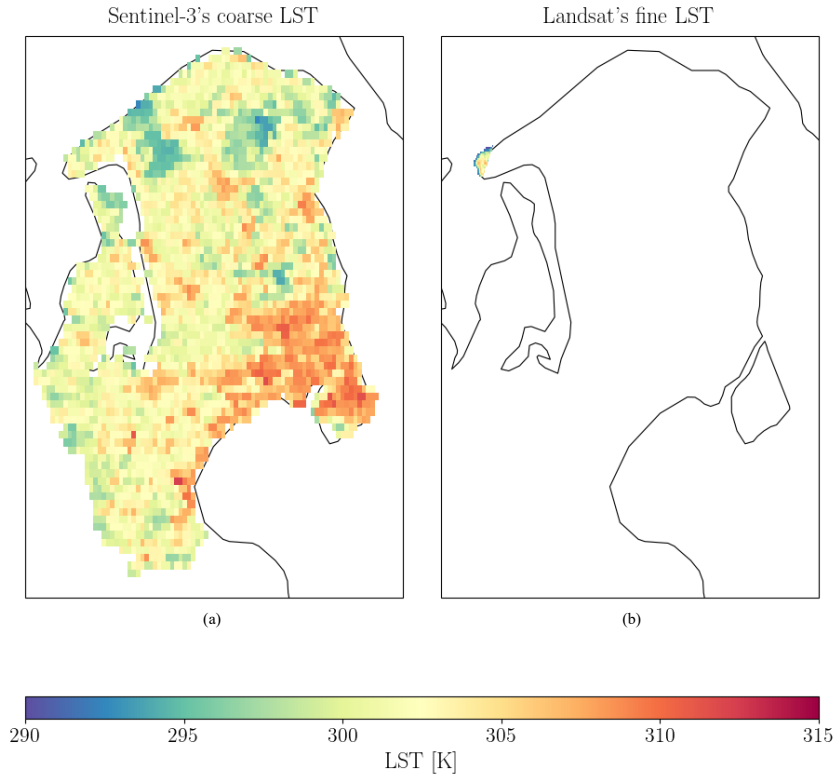


Figure A5. - Actual coarse (a) and fine (b) LST (with residual correction) for Copenhagen on 2023-06-08. Note that the Landsat swath does not fully cover the entire AOI in a single overpass (with the data for the Copenhagen region missing for this date).

 doi:10.20944/preprints202604.0495.v1

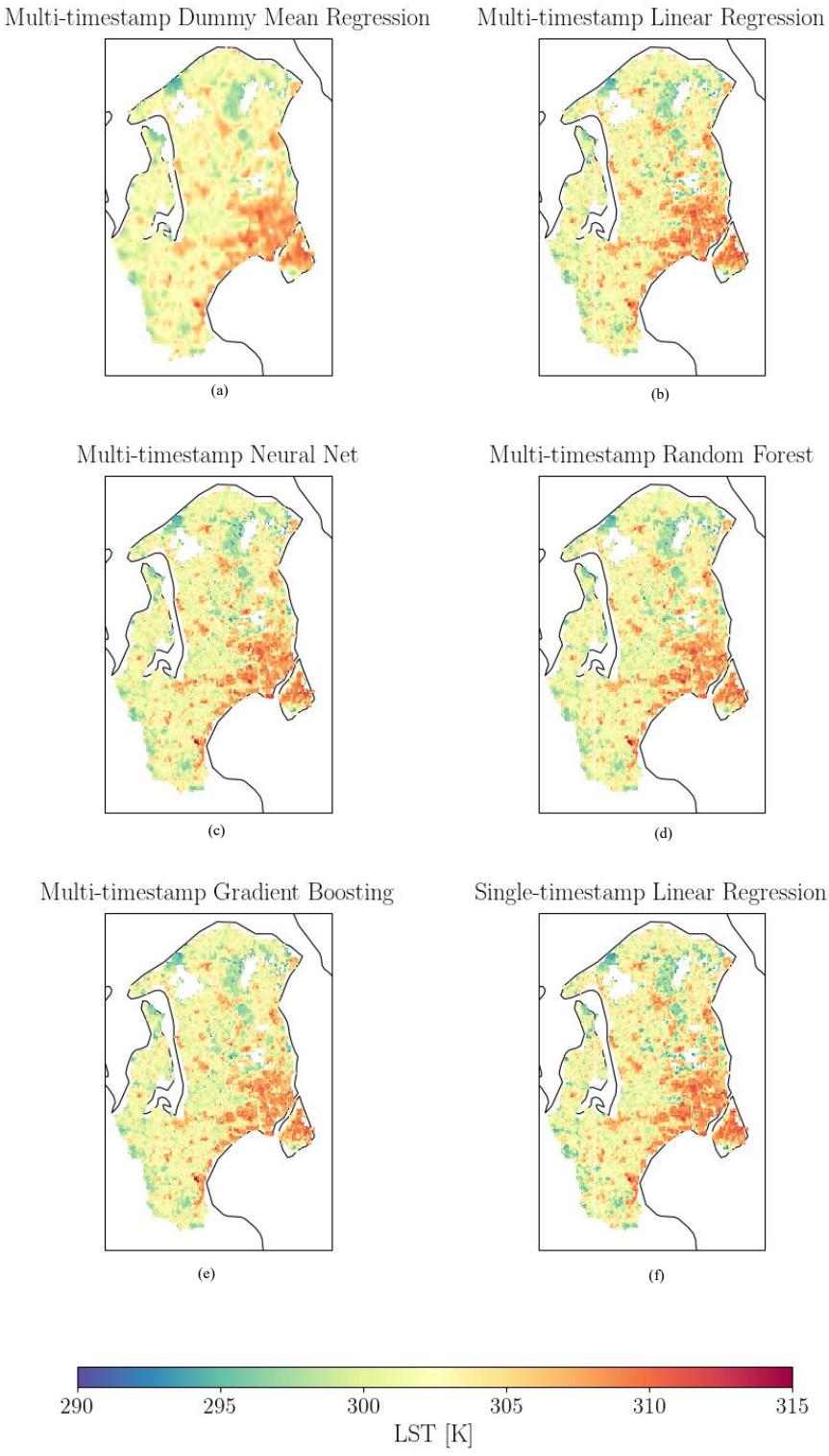


Figure A6. - Predicted fine LST (with residual correction) for Copenhagen on 2023-06-08: multi-timestamp Dummy Mean Regression (a), multi-timestamp Linear Regression (b), multi-timestamp Neural Net (c), multi-timestamp Random Forest (d), multi-timestamp Gradient Boosting (e) and single-timestamp Linear Regression (f).

References

1. Belward, A.; Bourassa, M.; Dowell, M.; Briggs, S.; Dolman, H. (A. J.); Holmlund, K.; Husband, R.; Quegan, S.; Simmons, A.; Sloyan, B.; et al. *The Global Observing System for Climate: Implementation Needs*; WMO: Geneva, Switzerland, 2016;

 doi:10.20944/preprints202604.0495.v1

2. Zhan, W.; Huang, F.; Quan, J.; Zhu, X.; Gao, L.; Zhou, J.; Ju, W. Disaggregation of Remotely Sensed Land Surface Temperature: A New Dynamic Methodology. *Journal of Geophysical Research: Atmospheres* **2016**, *121*, 10,538-10,554, doi:10.1002/2016JD024891.
3. Bechtel, B.; Demuzere, M.; Mills, G.; Zhan, W.; Sismanidis, P.; Small, C.; Voogt, J. SUHI Analysis Using Local Climate Zones—A Comparison of 50 Cities. *Urban Climate* **2019**, *28*, 100451, doi:10.1016/j.uclim.2019.01.005.
4. Cai, Y.; Chen, G.; Wang, Y.; Yang, L. Impacts of Land Cover and Seasonal Variation on Maximum Air Temperature Estimation Using MODIS Imagery. *Remote Sensing* **2017**, *9*, 233, doi:10.3390/rs9030233.
5. Liu, L.; Zhang, Y. Urban Heat Island Analysis Using the Landsat TM Data and ASTER Data: A Case Study in Hong Kong. *Remote Sensing* **2011**, *3*, 1535–1552, doi:10.3390/rs3071535.
6. Lopes, A.; Alves, E.; Alcoforado, M.J.; Machete, R. Lisbon Urban Heat Island Updated: New Highlights about the Relationships between Thermal Patterns and Wind Regimes. *Advances in Meteorology* **2013**, *2013*, 487695, doi:10.1155/2013/487695.
7. Wicki, A.; Parlow, E.; Feigenwinter, C. Evaluation and Modeling of Urban Heat Island Intensity in Basel, Switzerland. *Climate* **2018**, *6*, 55, doi:10.3390/cli6030055.
8. Wicki, A.; Parlow, E. Multiple Regression Analysis for Unmixing of Surface Temperature Data in an Urban Environment. *Remote Sensing* **2017**, *9*, doi:10.3390/rs9070684.
9. Oke, T.R.; Mills, G.; Christen, A.; Voogt, J.A. *Urban Climates*; Cambridge University Press, 2017; ISBN 978-0-521-84950-0.
10. Parlow, E. The Urban Heat Budget Derived from Satellite Data. *Geographica Helvetica* **2003**, *58*, 99–111, doi:10.5194/gh-58-99-2003.
11. Parlow, E.; Vogt, R.; Feigenwinter, C. The Urban Heat Island of Basel – Seen from Different Perspectives. *DIE ERDE – Journal of the Geographical Society of Berlin* **2014**, *145*, 96–110, doi:10.12854/erde-145-8.
12. Rigo, G.; Parlow, E.; Oesch, D. Validation of Satellite Observed Thermal Emission with In-Situ Measurements over an Urban Surface. *Remote Sensing of Environment* **2006**, *104*, 201–210, doi:10.1016/j.rse.2006.04.018.
13. Anderson, V.; Leung, A.C.W.; Mehdipoor, H.; Jänicke, B.; Milošević, D.; Oliveira, A.; Manavvi, S.; Kabano, P.; Dzyuban, Y.; Aguilar, R.; et al. Technological Opportunities for Sensing of the Health Effects of Weather and Climate Change: A State-of-the-Art-Review. *Int J Biometeorol* **2021**, *65*, 779–803, doi:10.1007/s00484-020-02063-z.
14. Parlow, E. Regarding Some Pitfalls in Urban Heat Island Studies Using Remote Sensing Technology. *Remote Sensing* **2021**, *13*, 3598, doi:10.3390/rs13183598.
15. Earth Resources Observation and Science (EROS) Center Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-1, Collection 2 2013.
16. Oliveira, A.; Lopes, A.; Correia, E.; Niza, S.; Soares, A. Heatwaves and Summer Urban Heat Islands: A Daily Cycle Approach to Unveil the Urban Thermal Signal Changes in Lisbon, Portugal. *Atmosphere* **2021**, *12*, 292, doi:10.3390/atmos12030292.
17. Pu, R.; Bonafoni, S. Thermal Infrared Remote Sensing Data Downscaling Investigations: An Overview on Current Status and Perspectives. *Remote Sensing Applications: Society and Environment* **2023**, *29*, 100921, doi:10.1016/j.rsase.2023.100921.
18. Hu, Y.; Tang, R.; Jiang, X.; Li, Z.-L.; Jiang, Y.; Liu, M.; Gao, C.; Zhou, X. A Physical Method for Downscaling Land Surface Temperatures Using Surface Energy Balance Theory. *Remote Sensing of Environment* **2023**, *286*, 113421, doi:10.1016/j.rse.2022.113421.
19. Li, Z.-L.; Tang, B.-H.; Wu, H.; Ren, H.; Yan, G.; Wan, Z.; Trigo, I.F.; Sobrino, J.A. Satellite-Derived Land Surface Temperature: Current Status and Perspectives. *Remote Sensing of Environment* **2013**, *131*, 14–37, doi:10.1016/j.rse.2012.12.008.
20. Voogt, J.A.; Oke, T.R. Thermal Remote Sensing of Urban Climates. *Remote Sensing of Environment* **2003**, *86*, 370–384, doi:10.1016/S0034-4257(03)00079-8.
21. Wen, J.; He, Y.; Yang, L.; Wan, P.; Gu, Z.; Wang, Y. A Two-Step Downscaling Model for MODIS Land Surface Temperature Based on Random Forests. *Atmosphere* **2025**, *16*, doi:10.3390/atmos16040424.

 doi:10.20944/preprints202604.0495.v1

22. Hutengs, C.; Vohland, M. Downscaling Land Surface Temperatures at Regional Scales with Random Forest Regression. *Remote Sensing of Environment* **2016**, *178*, 127–141, doi:10.1016/j.rse.2016.03.006.
23. Mechri, R.; Ottlé, C.; Pannekoucke, O.; Kallel, A. Genetic Particle Filter Application to Land Surface Temperature Downscaling. *Journal of Geophysical Research: Atmospheres* **2014**, *119*, 2131–2146, doi:10.1002/2013JD020354.
24. Zhang, L.; Yan, H.; Qiu, L.; Cao, S.; He, Y.; Pang, G. Spatial and Temporal Analyses of Vegetation Changes at Multiple Time Scales in the Qilian Mountains. *Remote Sensing* **2021**, *13*, 5046, doi:10.3390/rs13245046.
25. Bisquert, M.; Sánchez, J.M.; Caselles, V. Evaluation of Disaggregation Methods for Downscaling MODIS Land Surface Temperature to Landsat Spatial Resolution in Barrax Test Site. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2016**, *9*, 1430–1438, doi:10.1109/JSTARS.2016.2519099.
26. Lacerda, L.N.; Cohen, Y.; Snider, J.; Huryna, H.; Liakos, V.; Vellidis, G. Field Scale Assessment of the TsHARP Technique for Thermal Sharpening of MODIS Satellite Images Using VENμS and Sentinel-2-Derived NDVI. *Remote Sensing* **2021**, *13*, 1155, doi:10.3390/rs13061155.
27. SLSTR Processing Available online: https://sentiwiki.copernicus.eu/web/slstr-processing (accessed on 26 January 2026).
28. United States Geological Survey EarthExplorer Available online: https://earthexplorer.usgs.gov/ (accessed on 11 February 2026).
29. Ecosystem, C.D.S. Sentinel-3 Level 2 SYN Data Collections Available | Copernicus Data Space Ecosystem Available online: https://dataspace.copernicus.eu/news/2024-12-17-sentinel-3-level-2-syn-data-collections-available (accessed on 26 January 2026).
30. Agam, N.; Kustas, W.P.; Anderson, M.C.; Li, F.; Neale, C.M.U. A Vegetation Index Based Technique for Spatial Sharpening of Thermal Imagery. *Remote Sensing of Environment* **2007**, *107*, 545–558, doi:10.1016/j.rse.2006.10.006.
31. Choudhury, B.J.; Ahmed, N.U.; Idso, S.B.; Reginato, R.J.; Daughtry, C.S.T. Relations between Evaporation Coefficients and Vegetation Indices Studied by Model Simulations. *Remote Sensing of Environment* **1994**, *50*, 1–17, doi:10.1016/0034-4257(94)90090-6.
32. Copernicus Data Space Ecosystem OData – Documentation Available online: https://documentation.dataspace.copernicus.eu/APIs/OData.html (accessed on 11 February 2026).
33. Chapman, L. Assessing Topographic Exposure. *Meteorological Applications* **2000**, *7*, 335–340, doi:10.1017/S1350482700001729.
34. Ecosystem, C.D.S. Copernicus DEM - Global and European Digital Elevation Model | Copernicus Data Space Ecosystem Available online: https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM (accessed on 26 January 2026).
35. Spatial without Compromise · QGIS Web Site Available online: https://qgis.org/ (accessed on 26 January 2026).
36. Tree Cover Density 2018 (Raster 10 m, 100 m), Europe, Yearly Available online: https://land.copernicus.eu/en/products/high-resolution-layer-forests-and-tree-cover/tree-cover-density-2018-raster-10-m-100-m-europe-yearly (accessed on 26 January 2026).
37. Oliveira, A.; Lopes, A.; Niza, S. Local Climate Zones Classification Method from Copernicus Land Monitoring Service Datasets: An ArcGIS-Based Toolbox. *MethodsX* **2020**, *7*, 101150, doi:10.1016/j.mex.2020.101150.
38. Grimmond, C.S.B.; Oke, T.R. Heat Storage in Urban Areas: Local-Scale Observations and Evaluation of a Simple Model. *Journal of Applied Meteorology and Climatology* **1999**, *38*, 922–940, doi:10.1175/1520-0450(1999)038<0922:HSIUAL>2.0.CO;2.
39. Ruel, J.-C.; Pin, D.; Spacek, L.; Cooper, K.; Benoit, R. The Estimation of Wind Exposure for Windthrow Hazard Rating: Comparison between Strongblow, MC2, Topex and a Wind Tunnel Study. *Forestry (Lond)* **1997**, *70*, 253–266, doi:10.1093/forestry/70.3.253.
40. Marine Regions Available online: https://www.marineregions.org/sources.php (accessed on 26 January 2026).

 doi:10.20944/preprints202604.0495.v1

41. Imperviousness Density 2018 (Raster 10 m and 100 m), Europe, 3-Yearly Available online: https://land.copernicus.eu/en/products/high-resolution-layer-imperviousness/imperviousness-density-2018 (accessed on 26 January 2026).
42. Ait-Bachir, R.; Granero-Belinchon, C.; Michel, A.; Michel, J.; Briottet, X.; Drumetz, L. Land Surface Temperature Super-Resolution With a Scale-Invariance-Free Neural Approach: Application to MODIS. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2025**, *18*, 14480–14494, doi:10.1109/JSTARS.2025.3573610.
43. Zhou, J.; Liu, S.; Li, M.; Zhan, W.; Xu, Z.; Xu, T. Quantification of the Scale Effect in Downscaling Remotely Sensed Land Surface Temperature. *Remote Sensing* **2016**, *8*, 975, doi:10.3390/rs8120975.
44. Yang, Y.; Li, X.; Pan, X.; Zhang, Y.; Cao, C. Downscaling Land Surface Temperature in Complex Regions by Using Multiple Scale Factors with Adaptive Thresholds. *Sensors* **2017**, *17*, 744, doi:10.3390/s17040744.
45. Sánchez, J.M.; Galve, J.M.; González-Piqueras, J.; López-Urrea, R.; Niclòs, R.; Calera, A. Monitoring 10-m LST from the Combination MODIS/Sentinel-2, Validation in a High Contrast Semi-Arid Agroecosystem. *Remote Sensing* **2020**, *12*, 1453, doi:10.3390/rs12091453.
46. Sattari, F.; Hashim, M.; Sookhak, M.; Banihashemi, S.; Pour, A.B. Assessment of the TsHARP Method for Spatial Downscaling of Land Surface Temperature over Urban Regions. *Urban Climate* **2022**, *45*, 101265, doi:10.1016/j.uclim.2022.101265.
47. Wang, Z.; Sui, L.; Zhang, S. Generating Daily Land Surface Temperature Downscaling Data Based on Sentinel-3 Images. *Remote Sensing* **2022**, *14*, 5752, doi:10.3390/rs14225752.
48. Pedregosa, F.; Pedregosa, F.; Varoquaux, G.; Varoquaux, G.; Org, N.; Gramfort, A.; Gramfort, A.; Michel, V.; Michel, V.; Fr, L.; et al. Scikit-Learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.
49. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: San Francisco California USA, August 13 2016; pp. 785–794.
50. Wilkinson, G.N.; Rogers, C.E. Symbolic Description of Factorial Models for Analysis of Variance. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **1973**, *22*, 392–399, doi:10.2307/2346786.
51. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Association for Computing Machinery: New York, NY, USA, July 25 2019; pp. 2623–2631.
52. Ding, L.; Zhou, J.; Ma, J.; Zhu, X.; Wang, W.; Li, M. A Spatial Downscaling Approach for Land Surface Temperature by Considering Descriptor Weight. *IEEE Geoscience and Remote Sensing Letters* **2023**, *20*, 1–5, doi:10.1109/LGRS.2023.3255785.
53. Hernanz, A.; García-Valero, J.A.; Domínguez, M.; Rodríguez-Camino, E. A Critical View on the Suitability of Machine Learning Techniques to Downscale Climate Change Projections: Illustration for Temperature with a Toy Experiment. *Atmospheric Science Letters* **2022**, *23*, e1087, doi:10.1002/asl.1087.
54. Gao, F.; Kustas, W.P.; Anderson, M.C. A Data Mining Approach for Sharpening Thermal Satellite Imagery over Land. *Remote Sensing* **2012**, *4*, 3287–3319, doi:10.3390/rs4113287.
55. Dominguez, A.; Kleissl, J.; Luvall, J.C.; Rickman, D.L. High-Resolution Urban Thermal Sharpener (HUTS). *Remote Sensing of Environment* **2011**, *115*, 1772–1780, doi:10.1016/j.rse.2011.03.008.
56. Inamdar, A.K.; French, A.; Hook, S.; Vaughan, G.; Luckett, W. Land Surface Temperature Retrieval at High Spatial and Temporal Resolutions over the Southwestern United States. *Journal of Geophysical Research: Atmospheres* **2008**, *113*, doi:10.1029/2007JD009048.
57. Inamdar, A.K.; French, A. Disaggregation of GOES Land Surface Temperatures Using Surface Emissivity. *Geophysical Research Letters* **2009**, *36*, doi:10.1029/2008GL036544.
58. Merlin, O.; Duchemin, B.; Hagolle, O.; Jacob, F.; Coudert, B.; Chehbouni, G.; Dedieu, G.; Garatuza, J.; Kerr, Y. Disaggregation of MODIS Surface Temperature over an Agricultural Area Using a Time Series of Formosat-2 Images. *Remote Sensing of Environment* **2010**, *114*, 2500–2512, doi:10.1016/j.rse.2010.05.025.
59. Jeganathan, C.; Hamm, N.A.S.; Mukherjee, S.; Atkinson, P.M.; Raju, P.L.N.; Dadhwal, V.K. Evaluating a Thermal Image Sharpening Model over a Mixed Agricultural Landscape in India. *International Journal of Applied Earth Observation and Geoinformation* **2011**, *13*, 178–191, doi:10.1016/j.jag.2010.11.001.

 doi:10.20944/preprints202604.0495.v1

60. Zakšek, K.; Oštir, K. Downscaling Land Surface Temperature for Urban Heat Island Diurnal Cycle Analysis. *Remote Sensing of Environment* **2012**, *117*, 114–124, doi:10.1016/j.rse.2011.05.027.
61. Li, W.; Ni, L.; Li, Z.-L.; Wu, H. Downscaling Land Surface Temperature by Using Random Forest Regression Algorithm. In Proceedings of the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium; July 2018; pp. 2527–2530.
62. Wu, H.; Li, W. Downscaling Land Surface Temperatures Using a Random Forest Regression Model With Multitype Predictor Variables. *IEEE Access* **2019**, *7*, 21904–21916, doi:10.1109/ACCESS.2019.2896241.
63. Kustas, W.P.; Norman, J.M.; Anderson, M.C.; French, A.N. Estimating Subpixel Surface Temperatures and Energy Fluxes from the Vegetation Index–Radiometric Temperature Relationship. *Remote Sensing of Environment* **2003**, *85*, 429–440, doi:10.1016/S0034-4257(03)00036-1.
64. Mukherjee, S.; Joshi, P.K.; Garg, R.D. Evaluation of LST Downscaling Algorithms on Seasonal Thermal Data in Humid Subtropical Regions of India. *International Journal of Remote Sensing* **2015**, *36*, 2503–2523, doi:10.1080/01431161.2015.1041175.
65. Wang, J.; Tang, B.-H.; Zhu, X.; Fan, D.; Li, M.; Chen, J. A Comparative Analysis of Five Land Surface Temperature Downscaling Methods in Plateau Mountainous Areas. *Front. Earth Sci.* **2025**, *12*, doi:10.3389/feart.2024.1488711.
66. Oliveira, A.; Lopes, A.; Niza, S.; Soares, A. An Urban Energy Balance-Guided Machine Learning Approach for Synthetic Nocturnal Surface Urban Heat Island Prediction: A Heatwave Event in Naples. *Science of The Total Environment* **2022**, *805*, 150130, doi:10.1016/j.scitotenv.2021.150130.