remote sensing

MDPI

Technical Notes

Comparing ML Methods for Downscaling Near-Surface Air Temperature over the Eastern Mediterranean

Amit Blizer [1], Oren Glickman [2,*] and Itamar M. Lensky [1]

[1] Department of Geography and Environment, Bar-Ilan University; {amit.blizer, itamar.lensky}@biu.ac.il
[2] Department of Computer Science, Bar-Ilan University
* Correspondence: oren.glickman@biu.ac.il

Abstract: Near-surface air temperature (T_a) is a key variable in global climate studies. Global climate models such as ERA5 and CMIP6 predict various parameters at coarse spatial resolution (>9 km). As a result, local weather phenomena such as convection, thunderstorms, and urban heat islands are not reflected in the model's outputs. In this study, we address this limitation by downscaling the resolution of ERA5 (9 km) and CMIP6 (27 km) T_a to 1 km employing two different machine learning algorithms (XGBoost and Deep Learning). Our models leverage a diverse set of features, including data from satellites (land surface temperature and normalized difference vegetation index), from ERA5 and CMIP6 climate models (e.g., solar and thermal radiation, wind) and from digital elevation models to develop accurate machine learning prediction models. These models were rigorously validated against observations from 98 meteorological stations in the East Mediterranean (Israel) using cross-validation techniques, including leaving one group out on the station ID to avoid overfitting and dependence on geographic location. Our results demonstrate impressive accuracy, with the deep-learning based models obtaining Root Mean Squared Error (RMSE) values of 0.79°C (ERA5) and 1.32°C (CMIP6) for daily T_a, and 1.38°C (ERA5) for hourly T_a. Additionally, we explore the impact of the various input features and offer an extended application for future climate predictions. Finally, we propose an enhanced evaluation framework which addresses the problem of model overfitting. This work provides practical tools and insights for building and evaluating T_a downscaling models. The code and data are publicly shared online.

Keywords: Air temperature; Climate change; Climate models; Deep learning; Digital elevation models; Land surface temperature; Machine learning; Regression analysis; Remote sensing; Vegetation mapping.

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

In recent decades, global climate changes have had profound and often devastating effects, particularly at the local scale. Increased frequency and severity of events such as floods, droughts, and heatwaves have direct and indirect impacts across numerous fields. Near-surface air temperature at 2 m height (T_a) is a key variable of critical importance influencing these impacts. T_a plays a significant role in various fields, including agriculture ([1,2,3]), epidemiology ([4]), health ([5]), urban planning, and more. Thus, predicting climate trends at a high resolution is essential for developing climate change adaptation decision-support tools (climate services). However, obtaining accurate high-resolution T_a data can be challenging.

Meteorological stations do provide continuous measurements of T_a, however, their spatial distribution is irregular and coarse. To address this limitation, reanalysis data, which is a process that standardizes weather models using ground-based observation data ([6]), is used. The European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis provides hourly T_a data at a spatial resolution of 9 km. Addi-

tionally, ensemble forecasting uses a set (or ensemble) of forecasts to indicate the range of possible future states of the atmosphere. It runs multiple climate models with different initial conditions and selects the scenario with the highest probability. The Coupled Model Intercomparison Project Phase 6 (CMIP6) [7] is an ensemble of multiple climate models and re-analyses, offering climate change scenarios at a spatial resolution of 27 km. Despite these advances, the spatial resolution of ERA5 T_a data is too coarse and remains inadequate for many climate services. For example, local terrain effects are often missed in regions with rugged topography, and in agricultural areas characterized by small field sizes, the coarse resolution hinders precise climate analysis. Consequently, there is a growing need to enhance the spatial resolution of ERA5 (and CMIP6). Sharpening climate models T_a data to 1 km would contribute to the development of improved climate services.

Several parameters influence near-surface air temperature (T_a). Vegetation cover impacts T_a through evapotranspiration and the Normalized Difference Vegetation Index (NDVI) is often used to account for the impact of vegetation ([4]). Land surface temperature (LST) impacts T_a through sensible and latent heat fluxes ([8]), Solar radiation through surface heating, and winds through vertical mixing. Wind direction can influence T_a ([9]) through temperature advection ([10]). For example, in the Eastern Mediterranean winds with a northern component often advect cold air, while southeastern winds, which are common during transitional seasons, advect warm air. Additionally, topographic height significantly affects T_a ([11]) through air vertical motion, where the dry adiabatic lapse rate is ~9.8°C / 1000m. Topography in the Eastern Mediterranean varies significantly within short distances, leading to drastic changes in local temperature. For example, Mount Naftali, at 436 m above sea level, experiences maximum summer temperatures around 30°C, while just about 2 km away in the Hula Valley, at 70 m above sea level, temperatures can reach 37°C.

Multiple studies have undertaken the task of downscaling T_a using various methodologies, including statistical and classical machine learning (ML) techniques ([4,11,12,13,14,15,16]). These studies exhibit diversity in terms of geographical coverage, choice of algorithms (ranging from statistical and classical ML to Deep Learning), and the features utilized in model training. While some investigations focused on daily downscaling resolutions, either in terms of mean or min/max values, others delved into hourly downscaling predictions. A prevalent trend emerged in recent research, highlighting the efficacy of ML algorithms such as XGboost (XGB) and Random Forest (RF) in the T_a downscaling task. Furthermore, there is a growing body of work introducing Deep Learning (DL) models, including Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), demonstrating their potential and versatility in this context. These investigations span various regions worldwide, each presenting unique challenges and opportunities. To assess model performance, these studies typically employed cross-validation techniques, involving random data splitting for training and testing. The evaluation metric commonly employed to gauge the quality of downscaling predictions is the Root Mean Square Error (RMSE).

The summary of these works can be seen in Table 1. As can be seen from the comparison table, the reported RMSE values for downscaling daily T_a range between 0.85 to 2.14 and 1.58 to 1.77 for hourly predictions.

Though previous research efforts have made valuable contributions to the field of downscaling near-surface air temperature (T_a), there are notable gaps in the existing body of work.

Firstly, there is a need for a comprehensive study that systematically compares the performance of different downscaling methods. This comparative analysis should encompass both daily and hourly resolutions to provide a holistic view of their strengths and weaknesses. Additionally, the consideration of additional features for model train-

ing, such as satellite data and topographic information, holds the potential to further improve accuracy in downscaling predictions.

Furthermore, while much attention has been given to downscaling historical temperature data, there is a growing demand for downscaling future climate predictions. This extension of the task presents unique challenges and opportunities, including addressing the relevant features available in the different climate change scenarios. Finally, there is room for improvement in the evaluation methodology, with a focus on refining existing techniques and exploring novel approaches that align with the intricacies of downscaling.

This study aims to address these challenges comprehensively. We propose a thorough investigation that compares various downscaling methods, explores the impact of additional features, extends the application to future climate predictions, and enhances the evaluation framework. Through this research, we aspire to contribute to the advancement of downscaling techniques, ultimately providing more accurate and reliable climate data for a wide range of applications.

The following sections of this paper detail our methodology, experiments, results, and discussions, presenting a significant step forward in the field of downscaling air temperature using machine and deep learning algorithms.

Table 1. Comparison of recent studies on Ta downscaling.

Reference	Area	Methods *	RMSE	
			Hourly	Daily
Sebbar2023 [12]	Morocco (High Atlas)	MLR, SVR, XGB	1.61	
Afshari2023 [13]	Germany (Urban)	CNN	1.77	
Zhou2020 [4]	Israel	RF	1.58	
Mouatadid2017 [14]	Canada	ANN (DL)		0.85
Arumugam2022 [16]	Inidia	Linear Regression		1.09
Zhang2022 [11]	Global	SVCM-SP		1.75
Karaman2023 [15]	Turkey	RF		2.14

* Methods: MLR - Multiple Linear Regression; CNN - Convolutional Neural Network; SVR - Support Vector Regression; XGB - Extreme Gradient Boosting; SVCM-SP - Spatially Varying Coefficient Models with Sign Preservation.

2. Data and Methods

In this section we present the data and methods employed in our study, which aims to compare machine-learning based methods for downscaling T_a. We compare the performance of 6 different ML implementations consisting of two different ML algorithms (XGBoost and Deep Learning), two different time resolutions (hourly and daily) and based on input data obtained from two different climate models (ERA5 and CMIP6). The "ground truth" data target labels for training and testing the supervised ML models consists of T_a observations from meteorological stations. Input features for the ML models consist of climate (ERA5 and CMIP6), satellite, and topography data obtained from publicly available resources. We compared the performance of the 6 different ML implementations using two different evaluation methodologies. To extend the application of the various downscaling models to future climate predictions of CMIP6 we used historical CMIP6 data and adjusted the features appropriately. The overall flow chart describing the various implementations' flow and architecture can be seen in Figure 1.

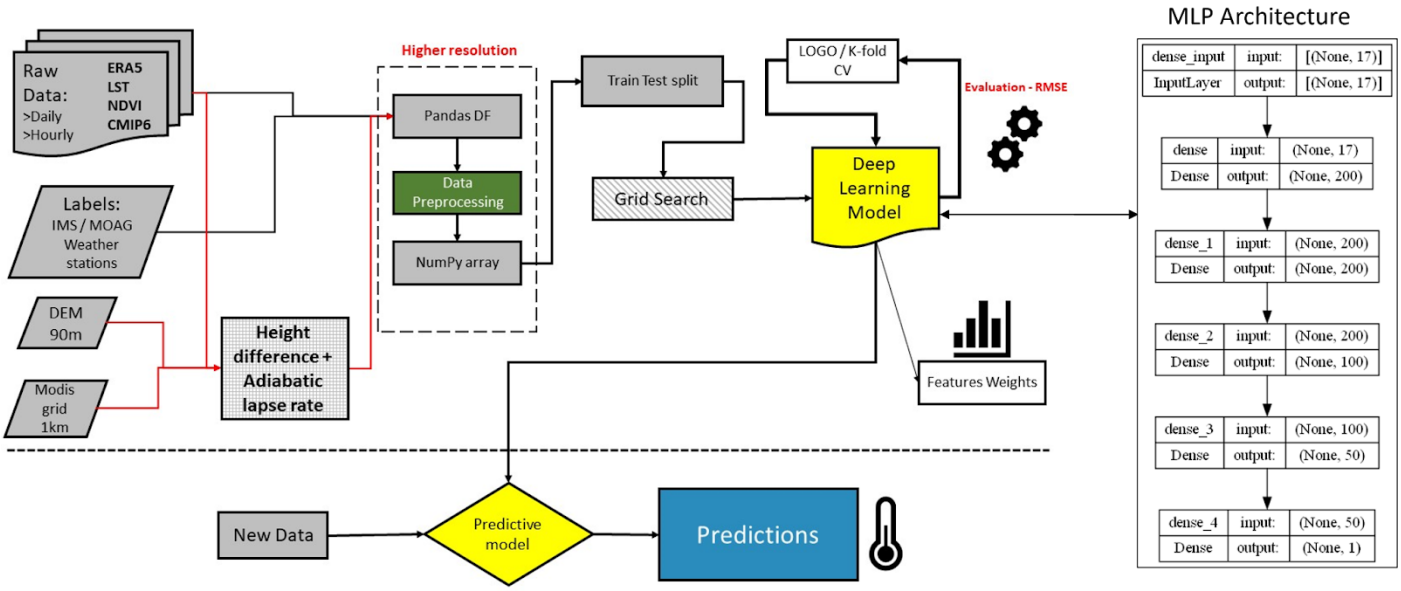

Figure 1. System Flow Chart and Multi Perceptron Neural Network (MLP) architecture.

2.1. Study Area

The study area, located in the Eastern Mediterranean (Figure 2a), is characterized by high spatial variability of climatic conditions ([17]) due to the region's complex orography (elevation: −430 up to 2814 m) and the spatial heterogeneity of land covers.

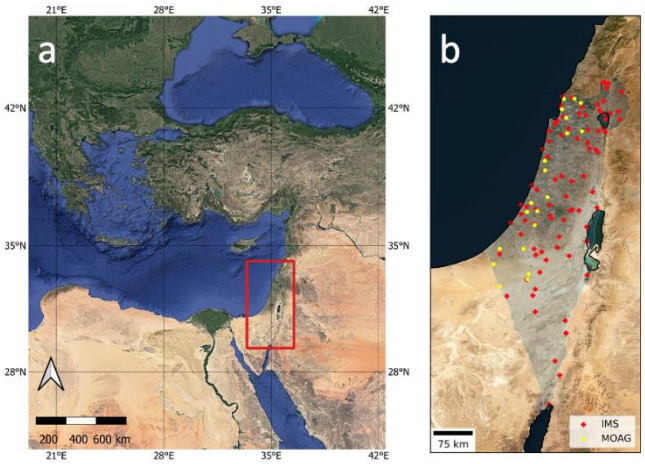

Figure 2. (a) The study area (red rectangle) in the Eastern Mediterranean. **(b)** Locations of the IMS (red) and MOAG (yellow) meteorological stations.

2.2. Data

The data for the supervised ML models consists of T_a observations as target values ("ground truth") and climate (ERA5 and CMIP6), satellite, and topography data used as model input features. The grids were projected on the same coordinate system (CRS) EPSG:4326.

2.2.1. Ground-Based Observations

We used T_a observations from the Israel Meteorological Service (IMS) and Ministry of Agriculture (MOAG) meteorological stations. The IMS data was obtained via an API

[18] in 10-minute temporal resolution and was resampled to hourly and daily temporal resolutions. The MOAG hourly data was provided by the Ministry of Agriculture as a CSV file. Figure 2b shows the geographical location of the 98 meteorological stations used in this study. The station data includes 12 years (2008-2019) of hourly T_a measurements with a total of 10.3 million hourly samples from all stations. 3.42 million hourly samples were used for the historical (ERA5) model and 6.87 million for the future (CMIP6) model. We used a different number of stations for each model with respect to the quality and availability of the data from the stations and the climate models. T_a measurement errors (i.e., values above 50°C or below -20°C) were excluded from the datasets. On top of that, we excluded stations with less than a minimum threshold (70%) of acceptable data samples.

2.2.2. ERA5 climate model data

ERA5 is the fifth generation ECMWF atmospheric reanalysis of the global climate covering the period from January 1940 to the present. ERA5 is produced by the Copernicus Climate Change Service (C3S) at ECMWF. ERA5 provides hourly estimates of a large number of atmospheric, land, and oceanic climate variables on a 9 km grid. It combines observational data from diverse sources, such as weather stations, satellites, and buoys, with a numerical weather prediction model to generate a comprehensive and consistent analysis of historical weather and climate conditions. With its high spatiotemporal resolution and global coverage, ERA5 provides a valuable resource for downscaling T_a and capturing detailed atmospheric information for machine and deep learning algorithms. ERA5 data containing T_a, U&V wind components, downward surface solar and thermal radiation, and net radiation were retrieved from the Climate Data Store (CDS) as netCDF files.

2.2.3. CMIP6 climate model data

Coupled Model Intercomparison Project (CMIP) is a project of the World Climate Research Programme (WCRP) that provides climate projections to understand past, present, and future climate changes ([19]). CMIP and its associated data infrastructure have become essential to the Intergovernmental Panel on Climate Change (IPCC) and other international and national climate assessments. CMIP6 historical data is available till the end of 2014, and future data till the end of the 21st century under different Representative Concentration Pathway greenhouse gas concentration (not emissions) trajectories adopted by the IPCC. The data was obtained utilizing the Google Earth Engine (GEE) platform.

2.2.4. MODIS satellite data

The spatial resolution and coverage of the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard the AQUA satellite, launched in 2002 as part of NASA's Earth Observing System, makes it a valuable resource for downscaling. We used the Shiff et al. ([20]) gap-filled MODIS LST dataset, which was found as the most suitable gap-filling method for MODIS LST data ([21]). We added MODIS daily 1 km LST and 250 m NDVI to the hourly dataset after up-sampling to hourly resolution using mean interpolation.

2.2.5. Digital Elevation Model data

As T_a is highly influenced by elevation, we obtained Digital Elevation Model (DEM) data for the study area. We downloaded the Shuttle Radar Topography Mission (SRTM) 90 m resolution DEM from the GEE platform ([22]). This data was used to calculate the mean elevation of all ERA5 (9 km) and CMIP6 (27 km) low-resolution (*LR*) grid cells in the study area. Similarly, we calculated the mean elevation of all 1 km high-resolution (*HR*) grid cells in the study area.

2.2.6. Corrected T_a based on height difference and adiabatic lapse rate

As the T_a values obtained from the climate models (ERA5 and CMIP6) are at a low resolution (9 and 27 km), we included a naïve correction of the T_a values based on the height difference between the mean height (h[LR]) of the low-resolution (LR) ERA5 or CMIP6 grid and the mean height (h[HR]) of the high-resolution (HR) 1 km grid. The corrected T_a [HR] value was calculated by using a vertical temperature gradient (dry adiabatic lapse rate) of 9.8°C/1000m as follows:

$$T_a\,[\mathrm{HR}] = T_a\,[\mathrm{LR}] - 9.8 * (\mathrm{h[HR]} - \mathrm{h[LR]})\,, \quad (1)$$

The corrected values were used as an additional input feature to the supervised machine learning models. Figure 3 illustrates the outcome of the ERA5 T_a correction for 4 PM 14-Jan-2015.

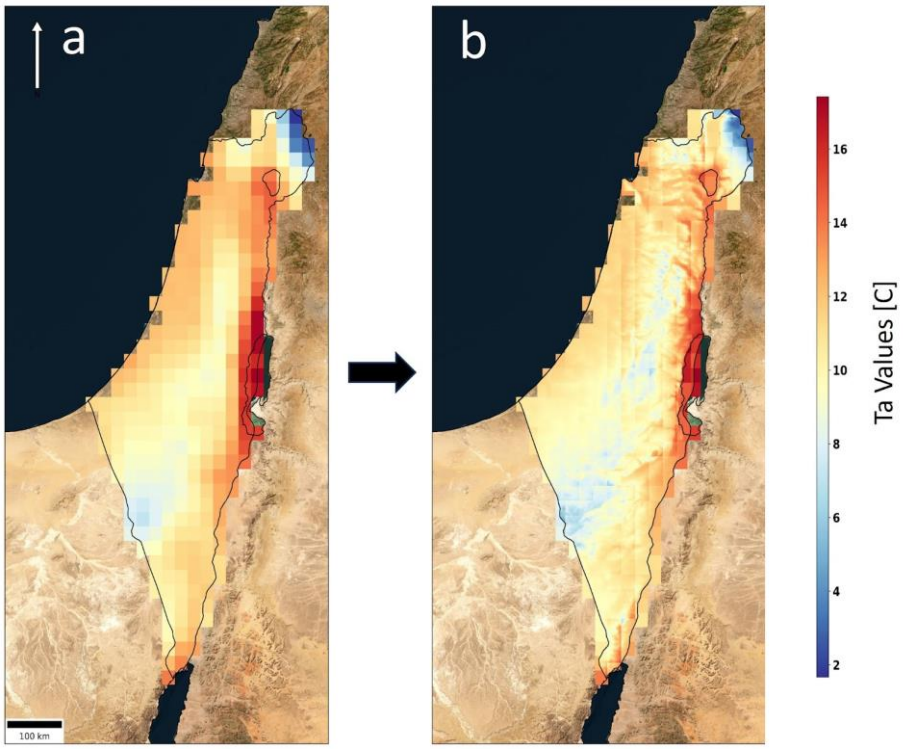

Figure 3. Downscaling ERA5 Ta based on dry adiabatic lapse as of 14-Jan-2015 at 4 PM.(a) Ta in 9km resolution. (b) Ta in 1km resolution

2.3. Methods

Following, we describe the two chosen ML algorithms, the implementation details for the past and future models based on the corresponding climate data as well as the two evaluation methodologies that were used.

2.3.1. XGBoost algorithm

One of the two algorithms employed in this study was XGBoost (Extreme Gradient Boosting), a powerful ensemble learning algorithm widely recognized for its robust predictive capabilities. For our XGBoost model, we configured specific hyperparameters to optimize its performance. We set the learning rate to 0.5 to control the step size at each iteration, ensuring a balance between model convergence and accuracy. A maximum depth of 6 branches was imposed on individual decision trees within the ensemble, regulating model complexity. A gamma rate of 0.1 was applied to specify the minimum reduction in loss required to make a new split, contributing to the overall model's ability to generalize. We employed 500 estimators, which represent the number of boosting rounds, allowing the model to progressively learn and adapt. This customized configu-

ration of XGBoost was based on trial-and-error experimentation on our training data enabling us to construct a predictive model tailored to our specific research data and task.

2.3.2. Deep Neural Network

In our study, we employed a Deep Neural Network (DNN) as one of the two modeling approaches. This deep-learning (DL) algorithm was implemented using Keras and TensorFlow, known for their effectiveness in developing intricate neural architectures. The Neural Network (NN) structured architecture consists of four fully connected layers, each with a varying number of neurons, configured in the following order: 200, 200, 100, and 50 neurons (Figure 1). To introduce non-linearity and enhance the model's expressive power, the Leaky-ReLu activation function was applied to each of these hidden layers. For the final dense layer, we used a single neuron, which aligns with our regression problem setting.

To initialize the weights in the NN layers, we employed a seeded Keras Glorot initializer, ensuring stability during training. Model performance was assessed using the Root Mean Square Error (RMSE) metric, which is suitable for regression tasks. To optimize the model's weights, we selected the Adam Optimizer, a popular choice for its ability to efficiently handle large datasets and complex architectures.

During training, a batch size of 2000 was employed to balance computational efficiency and convergence. To prevent overfitting and streamline model training, we incorporated early stopping callbacks, which halted training when performance on a validation set ceased to improve, thus preventing unnecessary iterations. This customized NN configuration, with careful consideration of architecture, activation functions, and optimization techniques, was integral to our research methodology, enabling us to effectively tackle the research questions at hand.

2.3.3. Models for downscaling historical data

We used the ERA5 dataset as historical climate model data. The training data of the historical model refers to a period of five years (2015-2019). For each of the historic ground observations, we collected ERA5, terrain, remote sensing, and temporal data to be used as input features for the supervised regression model. LST and NDVI daily values were up sampled to hourly resolution using mean interpolation. Overall, the raw historical data used to train the regression models include 3.2 million samples each with 15 different features (Table 2). Temporal features include the Day of Year (DOY, 0:364) and the Time of Day (TOD, 0:23) both of which were subjected to a *sin* and *cos* based trigonometric transformation [22] mapping each of them into two features in the range of [0, 1]. This transformation aids in accounting for the cyclic nature of temporal variables. As a means for the learning process to generalize to unseen stations we excluded features that contain station-specific information such as geographical coordinates or station ID.

RMSE was chosen to estimate the models' performance. We applied both the XGBoost and the DNN algorithms for both daily and hourly temporal resolutions – a total of 4 models. The hourly dataset included 3.3M samples and the daily dataset included 140K samples. The various ML algorithms were trained separately on hourly and daily data. Grid search was executed for each model to minimize loss rates.

2.3.4. Models for downscaling future data

To develop models suitable for the downscaling of CMIP6 climate scenario predictions, we followed a methodology akin to the historical model training process outlined above (2.3.3). However, this necessitated specific adaptations to accommodate the characteristics of future data predictions. Consequently, we tailored the input features for the model, ensuring their alignment with the variables available in the CMIP6 climate scenario data, while omitting features, such as NDVI observations, that do not apply in this

context. In this setting we only run daily prediction models as the future climate predictions of CMIP6 are not available at an hourly resolution.

Table 2. The features used to train the various models. The first 13 features were used in training the models for downscaling historical data (ERA5), DOY and HOD were used in all models and the 10 remaining features were used to train the models for downscaling future data (CMIP6).

Feature name	Short name	Resolutions	Source	Historic/Future
Air temperature (2 m)	T2M	10km / Hourly	ERA5	Historic
Surface solar radiation downwards	SSRD			
Surface thermal radiation downwards	STRD			
Surface thermal radiation (net)	STRN			
Surface solar radiation (net)	SSRN			
U component	U_comp			
V component	V_comp			
Wind Speed	WS			
Wind Direction	WD			
Diff (ERA5 – MODIS) mean pixel height	HeightD	1km / -		
NDVI	NDVI	250m / Daily	MODIS AQUA	
LST	LST	1km / Daily		
Day Of Year	DOY			Both
Hour Of Day	HOD	- /Hourly		
Corrected Ta (lapse rate)	T2M_cor	1km / Hourly		
Daily mean near-surface wind speed	DMNSWS	27km / Daily	CMIP6	Future
Surface downwelling shortwave radiation	SDSR			
Surface downwelling longwave radiation	SDLW			
Near-surface specific humidity	NSSH			
Near-surface relative humidity	NSRH			
Daily Ta	DNSTA			
Daily min Ta	DMINSTA			
Daily max Ta	DMANSTA			
Mean height in CMIP6 pixel	MHCP	1km / -	90m DEM	
Mean height in MODIS pixel	MHMP			

The CMIP6 historical climate data, characterized by a spatial resolution of 27 km and a daily temporal resolution, served as the foundation for our model training. We harnessed a seven-year dataset spanning from 2008 to 2014, which encompassed 13 essential features. The feature set comprised daily mean near-surface wind speed, surface downwelling shortwave radiation, surface downwelling longwave radiation, near-surface specific humidity, near-surface relative humidity as well as daily T_a and daily minimum and maximum T_a.

In a similar manner to our approach for the ERA5-based models used in historical data downscaling, we employed both XGBoost and DNN Algorithms to assess the downscaling of CMIP6 data.

2.3.5. Evaluation and Validation Methodology

As stated above, RMSE was chosen to estimate the models' performance. To test our models, we applied two different validation techniques. In alignment with previous studies (see Table 1), we randomly split the data to a training set (80%) and a validation/test set (20%). The training set was used to train the machine learning model, while the validation/test set was used to evaluate its performance. It is important to note that in this setting observations from the same station (though from a different day and/or time) appear in both the train and test set and although we excluded features that con-

tain station-specific information such as geographical coordinates or station ID the learning models may be partially station-specific, and the learning process may not generalize properly to unseen stations. Thus, we also implemented a more rigorous evaluation technique consisting of K-fold cross-validation while leaving one group out (LOGO) on the station ID. These measures were instrumental in meticulously validating the model's proficiency in downscaling T_a data.

3. Results

Table 3. LOGO and No-LOGO RMSE values for XGboost and Neural Network, for daily and hourly ERA5 (9 km, 2015-2019) datasets and daily CMIP6 (27 km, 2008-2014) datasets.

Model/ Source	Algorithm	RMSE			
		NO-Logo		LOGO	
		Daily	Hourly	Daily	Hourly
Historic (ERA5)	XGB	0.82	1.29	1.44	2.65
	NN	0.79	1.37	0.90	2.10
Future (CMIP6)	XGB	1.32	-	2.05	-
	NN	1.27	-	1.56	-

- Note that the models for downscaling future (CMIP6) data were run only at a daily resolution.

3.1. Historical (ERA5) and Future (CMIP6) data models

Table 3 summarizes the RMSE results obtained by the various models. For the historical data models hourly dataset, the best-performing models obtained on the test set RMSE values of 1.37 for the NN and 1.39 for the XGB model. The corresponding results when evaluating using the LOGO procedure were higher (i.e., lower prediction accuracy) – 2.10 for the NN and 2.65 for the XGB model.

As with the historic hourly dataset, the daily RMSE values were, as expected, lower (better) than the hourly RMSE values. The best-performing models obtained RMSE values of 0.79 for the NN and 0.82 for the XGB model. The corresponding results when evaluating using the LOGO procedure were higher – 0.90 for the NN and 1.44 for the XGB model.

RMSE Results for the future daily models (CMIP6) were slightly higher - 1.27 and 1.32 for the NN and XGB models and 1.56 and 2.05 when applying LOGO.

To gain deeper insights into the model's performance, we computed the average RMSE across different diurnal (Time of Day, TOD) and seasonal (Day of Year, DOY) cycles for the observations in the test dataset. In Figure 4a, the RMSE values for the hourly historical (ERA5) NN model with LOGO are displayed over the course of the day, revealing a fluctuating pattern with peaks occurring between 4 and 9 a.m. and from 3 to 6 p.m. In Figure 4b, the corresponding RMSE values for this model are displayed over the course of year, revealing a peak in the spring in which RMSE values are comparatively high. Figure 5a shows RMSE values over the DOY for the daily historic model. As can be seen from the figure, higher RMSE values are obtained during spring and especially in autumn. Figure 5b shows RMSE values over the DOY for the daily future CMIP6 model showing high values at the end of the winter.

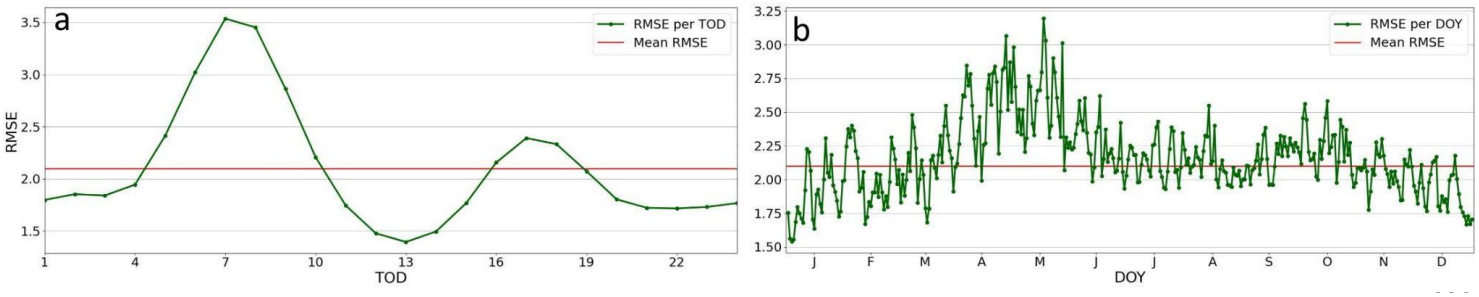

Figure 4. RMSE values for Leave One Group Out (on Station ID) using Neural Network, resulting from the hourly dataset at the (**a**) diurnal (TOD), and (**b**) seasonal (DOY) cycles.

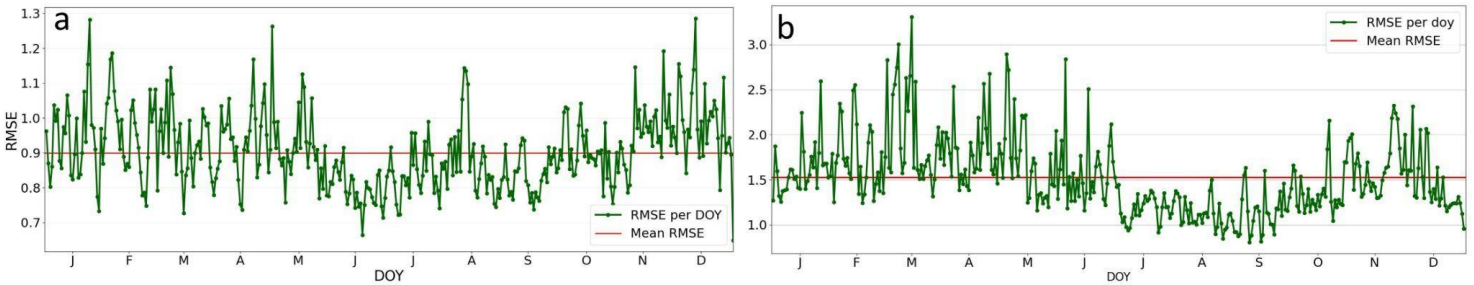

Figure 5. RMSE values per Day of Year (DOY) for Leave One Group Out (on Station ID) using Neural Network on the (a) ERA5 and (b) CMIP6 daily dataset.

3.2. Feature Importance

Investigating the feature importance across the diverse models employed in our study reveals valuable insights into their predictive capabilities and their contribution to the model performance. Figure 6 shows the feature weights for the XGboost daily CMIP6 model. The feature with the highest importance was Near-surface relative humidity followed by Surface downwelling shortwave radiation. For the historical hourly XGB model, the mean height difference had the highest importance followed by NDVI and LST (Figure 7b). While T_a from reanalysis (T2M) and wind U component (U_comp) were the prominent features for the historical daily XGB model (Figure 7a).

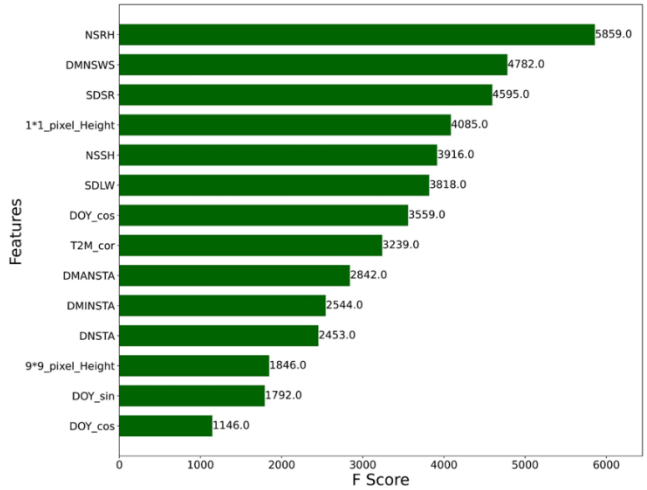

Figure 6. Feature weights for XGboost daily CMIP6 dataset

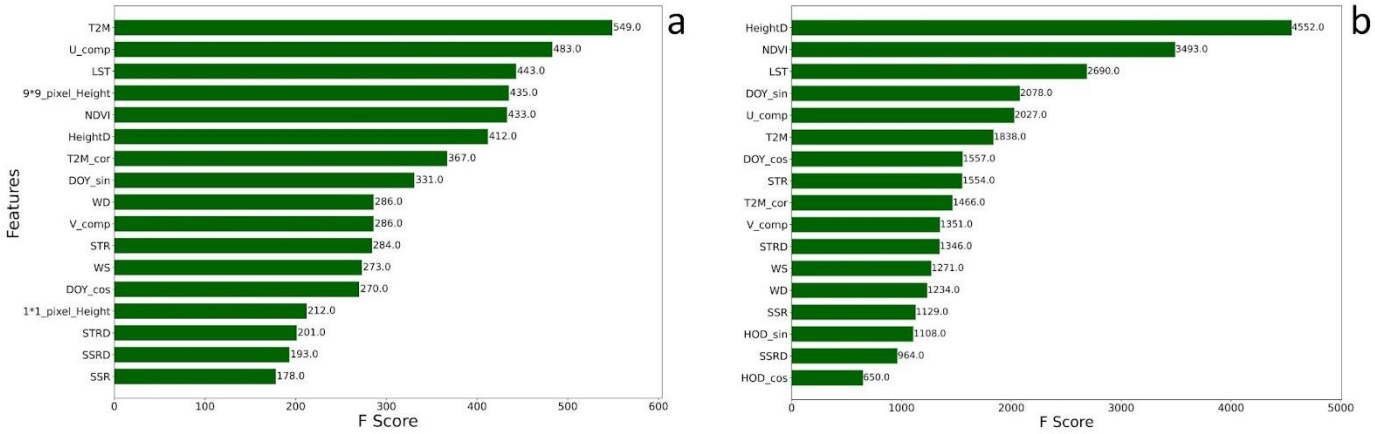

Figure 7. XGBoost features weights for historic ERA5 (a) daily, and (b) hourly model.

4. Discussion

In this section, we discuss the findings presented in the results section. Overall, the results indicate that both the XGB and the NN-based models downscaled T_a to 1km at hourly and daily temporal resolution with high accuracy improving over RMSE results reported in similar T_a-downscaling studies. This implies the strength of our models and approach. It is important to note that given that the various studies reported in Table 1 applied their models to different geographical areas and time points, one cannot make an exact 1-to-1 comparison of the results. To facilitate such future work, we share our models and data and are making them available to the academic community.

When comparing the two machine learning models – NN and XGB, both models obtained comparable results though the deep learning models consistently outperformed the XGBoost-based ones. This is possibly due to the large amount of training data and the strong capacity of NN to learn complex non-linear patterns and memorize information from the training data.

Results when applying the LOGO evaluation methodology, reveal higher RMASE (lower model precision) compared to the evaluation based on the conventional train/test approach as frequently employed in prior studies. It seems as if, despite the exclusion of station-specific features, our models demonstrate challenges in achieving robust generalization to unseen data points. This underscores the importance of adopting the LOGO methodology for reporting downscaling evaluation results. It also underscores the need for further research aimed at enhancing the adaptability of machine learning models to improve their generalization capabilities in such settings.

The future (CMIP6) model's performance was inferior to the ERA5 as expected due to fewer input features and coarser resolution. However, this model still obtains good results and can be used to be applied on future data.

Atmospheric processes' spatial and temporal scales are related. This is manifested in the feature importance that differs for the daily vs. hourly models. ERA5 coarse temporal resolution (Daily) T_a values are influenced by 9 km resolution ERA5 features, i.e., temperature and wind U (east-west) component (Figure 7a), while at the fine temporal resolution (hourly, Figure 7b) T_a values are influenced by fine spatial resolution features that capture the local conditions, i.e., topography and NDVI. LST is the third feature of importance in both models.

The diurnal variations of RMSE over TOD (Figure 4a) show a peak at sunrise (6 a.m.) when the maximal effect of nocturnal radiative cooling occurs. It is related to the spatial distribution of the meteorological stations where stations in rural areas are strongly affected while most other stations (in proximity to urban areas) are less prone

to this effect due to the urban heat island effect. On top of that, stations in low altitudes, especially in valleys will experience in clear sky nights much lower temperatures than others in higher latitudes or alternatively proximity to the seashore. The seasonal variations of RMSE over DOY (Figures 4b,5) show large variations in the spring and autumn when the synoptic circulation patterns change from steady summer patterns [24] to more variable winter patterns.

5. Conclusions

In this study, we successfully addressed the limitations of coarse spatial resolution global climate models, such as ERA5 and CMIP6, through the implementation of machine learning-driven downscaling techniques to enhance the resolution of near-surface air temperature (T_a) data. Leveraging a diverse array of features, including satellite data, climate model outputs, and digital elevation models, we developed accurate downscaling models, with the deep learning model proving most effective.

We strongly advocate for the adoption of the 'leaving one group out' (LOGO) evaluation methodology when testing and reporting model results, as it provides a robust validation approach enhancing the reliability of findings in climate-related studies. Additionally, we are committed to sharing our models and data with the academic community, facilitating further research and a deeper understanding of local climate dynamics. These resources offer invaluable insights for climate adaptation strategies and informed decision-making processes. The corresponding code and data of this work are publicly shared online.

Supplementary Materials: TBD

Author Contributions: Conceptualization, A.B. and I.M.L.; methodology, A.B., O.G. and I.M.L.; software, A.B.; validation, A.B., O.G. and I.M.L.; formal analysis, A.B., O.G. and I.M.L.; investigation, A.B., O.G. and I.M.L.; resources, A.B. and I.M.L.; data curation, A.B.; writing—original draft preparation, A.B., O.G. and I.M.L.; writing—review and editing, A.B., O.G. and I.M.L.; visualization, A.B.; supervision, O.G. and I.M.L; project administration, I.M.L.; funding acquisition, O.G. and I.M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Chief Scientist of the Israeli Ministry of Agriculture, grant numbers 20-01-0058 and 18-17-0012

Data Availability Statement: The code and data for this work is available at: https://github.com/estipollak/Downscaling-Near-Surface-Air-Temperature.

Acknowledgments: The authors thank Esti Polak for her assistance in organizing the code and making it suitable to be publicly available on GitHub.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. M. Blum, I. M. Lensky, and D. Nestel, "Estimation of olive grove canopy temperature from MODIS thermal imagery is more accurate than interpolation from meteorological stations," *Agric For Meteorol*, vol. 176, pp. 90–93, Jul. 2013, doi: 10.1016/J.AGRFORMET.2013.03.007.
2. M. Blum, D. Nestel, Y. Cohen, E. Goldshtein, D. Helman, and I. M. Lensky, "Predicting Heliothis (Helicoverpa armigera) pest population dynamics with an age-structured insect population model driven by satellite data," *Ecol Modell*, vol. 369, pp. 1–12, Feb. 2018, doi: 10.1016/J.ECOLMODEL.2017.12.019.
3. S. Shiff, I. M. Lensky, and D. J. Bonfil, "Using Satellite Data to Optimize Wheat Yield and Quality under Climate Change," *Remote Sensing 2021, Vol. 13, Page 2049*, vol. 13, no. 11, p. 2049, May 2021, doi: 10.3390/RS13112049.
4. B. Zhou *et al.*, "Estimating near-surface air temperature across Israel using a machine learning based hybrid approach," *Int J Climatol*, vol. 40, p. 14, 2020, doi: 10.1002/joc.6570.

5. X. Basagaña *et al.*, "Low and High Ambient Temperatures during Pregnancy and Birth Weight among 624,940 Singleton Term Births in Israel (2010–2014): An Investigation of Potential Windows of Susceptibility," *Environ Health Perspect*, vol. 129, no. 10, Oct. 2021, doi: 10.1289/EHP8117.
6. C. Di Napoli, C. Barnard, C. Prudhomme, H. L. Cloke, and F. Pappenberger, "ERA5-HEAT: A global gridded historical dataset of human thermal comfort indices from climate reanalysis," *Geosci Data J*, vol. 8, no. 1, pp. 2–10, Jun. 2021, doi: 10.1002/GDJ3.102.
7. B. Thrasher, W. Wang, A. Michaelis, F. Melton, T. Lee, and R. Nemani, "NASA Global Daily Downscaled Projections, CMIP6," *Scientific Data 2022 9:1*, vol. 9, no. 1, pp. 1–6, Jun. 2022, doi: 10.1038/s41597-022-01393-4.
8. I. M. Lensky, U. Dayan, and D. Helman. "Synoptic circulation impact on the near-surface temperature difference outweighs that of the seasonal signal in the Eastern Mediterranean." *Journal of Geophysical Research: Atmospheres*, 123, 11,333–11,347. https://doi.org/10.1029/2017JD027973
9. X. Min *et al.*, "Spatially Downscaling IMERG at Daily Scale Using Machine Learning Approaches Over Zhejiang, Southeastern China," *Front Earth Sci (Lausanne)*, vol. 8, p. 146, Jun. 2020, doi: 10.3389/FEART.2020.00146/BIBTEX.
10. I. M. Lensky and U. Dayan, "Satellite observations of land surface temperature patterns induced by synoptic circulation," *International Journal of Climatology*, vol. 35, no. 2, pp. 189–195, Feb. 2015, doi: 10.1002/JOC.3971.
11. T. Zhang *et al.*, "A global dataset of daily maximum and minimum near-surface air temperature at 1 km resolution over land (2003–2020)," *Earth Syst Sci Data*, vol. 14, no. 12, pp. 5637–5649, Dec. 2022, doi: 10.5194/ESSD-14-5637-2022.
12. B. E. Sebbar *et al.*, "Machine-Learning-Based Downscaling of Hourly ERA5-Land Air Temperature over Mountainous Regions," *Atmosphere 2023, Vol. 14, Page 610*, vol. 14, no. 4, p. 610, Mar. 2023, doi: 10.3390/ATMOS14040610.
13. A. Afshari, J. Vogel, and G. Chockalingam, "Statistical Downscaling of SEVIRI Land Surface Temperature to WRF Near-Surface Air Temperature Using a Deep Learning Model," *Remote Sensing 2023, Vol. 15, Page 4447*, vol. 15, no. 18, p. 4447, Sep. 2023, doi: 10.3390/RS15184447.
14. S. Mouatadid, S. Easterbrook, and A. R. Erler, "A Machine Learning Approach to Non-uniform Spatial Downscaling of Climate Variables," *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2017-November, pp. 332–341, Dec. 2017, doi: 10.1109/ICDMW.2017.49.
15. Ç. Hasan Karaman and Z. Akyürek, "Evaluation of near-surface air temperature reanalysis datasets and downscaling with machine learning based Random Forest method for complex terrain of Turkey," *Advances in Space Research*, 2023, doi: 10.1016/J.ASR.2023.02.006.
16. P. Arumugam, · N R Patel, and · V Kumar, "Estimation of air temperature using the temperature/vegetation index approach over Andhra Pradesh and Karnataka," 2022, doi: 10.1007/s12665-022-10180-8.
17. I. M. Lensky and U. Dayan, "Detection of Finescale Climatic Features from Satellites and Implications for Agricultural Planning." *Bull. Amer. Meteor. Soc.*, 92, 1131–1136. Sep 2011. doi: 10.1175/2011BAMS3160.1.
18. 10 & 1 -minutes data (API). Israel Meteorological Service. Available online: https://ims.gov.il/en/ObservationDataAPI
19. V. Eyring *et al.*, "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," *Geosci Model Dev*, vol. 9, no. 5, pp. 1937–1958, May 2016, doi: 10.5194/GMD-9-1937-2016.
20. S. Shiff, D. Helman, and I. M. Lensky, "Worldwide continuous gap-filled MODIS land surface temperature dataset," *Scientific Data 2021 8:1*, vol. 8, no. 1, pp. 1–10, Mar. 2021, doi: 10.1038/s41597-021-00861-7.
21. Y. Mo, Y. Xu, Y. Liu, Y. Xin, and S. Zhu, "Comparison of gap-filling methods for producing all-weather daily remotely sensed near-surface air temperature," *Remote Sens Environ*, vol. 296, p. 113732, Oct. 2023, doi: 10.1016/J.RSE.2023.113732.
22. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens Environ*, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/J.RSE.2017.06.031.
23. Bescond, P. L. "Cyclical features encoding, it's about time." *URL: https://towardsdatascience. com/cyclical-features-encoding-its-about-time-ce23581845ca (viitattu 15. 12. 2020)* (2020).
24. A. Bitan and H. Sa'Aroni, "The horizontal and vertical extension of the Persian Gulf pressure trough," International Journal of Climatology, vol. 12, no. 7, pp. 733–747, 1992, doi: 10.1002/JOC.3370120706.